




METHOD ARTICLE

# Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers

[version 1; referees: 2 approved]

Kieran R Campbell<sup>1,2</sup>, Christopher Yau <sup>2,3</sup>

<sup>1</sup>Department of Physiology, Anatomy and Genetics, University of Oxford, Oxford, OX1 3QX, UK

<sup>2</sup>Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, OX3 7BN, UK

<sup>3</sup>Centre for Computational Biology, Institute of Cancer and Genomic Sciences, University of Birmingham, Birmingham, B15 2TT, UK



**v1** First published: 15 Mar 2017, 2:19 (doi: [10.12688/wellcomeopenres.11087.1](https://doi.org/10.12688/wellcomeopenres.11087.1))  
Latest published: 15 Mar 2017, 2:19 (doi: [10.12688/wellcomeopenres.11087.1](https://doi.org/10.12688/wellcomeopenres.11087.1))

## Abstract

Modeling bifurcations in single-cell transcriptomics data has become an increasingly popular field of research. Several methods have been proposed to infer bifurcation structure from such data, but all rely on heuristic non-probabilistic inference. Here we propose the first generative, fully probabilistic model for such inference based on a Bayesian hierarchical mixture of factor analyzers. Our model exhibits competitive performance on large datasets despite implementing full Markov-Chain Monte Carlo sampling, and its unique hierarchical prior structure enables automatic determination of genes driving the bifurcation process. We additionally propose an Empirical-Bayes like extension that deals with the high levels of zero-inflation in single-cell RNA-seq data and quantify when such models are useful. We apply our model to both real and simulated single-cell gene expression data and compare the results to existing pseudotime methods. Finally, we discuss both the merits and weaknesses of such a unified, probabilistic approach in the context practical bioinformatics analyses.

## Open Peer Review

Referee Status:  

	Invited Referees	
	1	2
<b>version 1</b> published 15 Mar 2017	 report	 report

1 **Anthony Gitter** , University of Wisconsin–Madison USA

2 **Luca Pinello**, Massachusetts General Hospital and Harvard Medical School USA

## Discuss this article

Comments (0)

**Corresponding author:** Christopher Yau ([cyou@well.ox.ac.uk](mailto:cyou@well.ox.ac.uk))

**How to cite this article:** Campbell KR and Yau C. **Probabilistic modeling of bifurcations in single-cell gene expression data using a Bayesian mixture of factor analyzers [version 1; referees: 2 approved]** Wellcome Open Research 2017, 2:19 (doi: [10.12688/wellcomeopenres.11087.1](https://doi.org/10.12688/wellcomeopenres.11087.1))

**Copyright:** © 2017 Campbell KR and Yau C. This is an open access article distributed under the terms of the [Creative Commons Attribution Licence](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

**Grant information:** This work was supported by the Wellcome Trust Core Award [090532]; a UK Medical Research Council funded doctoral studentship to KC; a UK Medical Research Council New Investigator Research Grant to CY [MR/L001411/1]; the John Fell Oxford University Press Research Fund to CY; the Li Ka Shing Foundation via a Oxford-Stanford Big Data in Human Health Seed Grant to CY. *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

**Competing interests:** No competing interests were disclosed.

**First published:** 15 Mar 2017, 2:19 (doi: [10.12688/wellcomeopenres.11087.1](https://doi.org/10.12688/wellcomeopenres.11087.1))

## Introduction

Trajectory analysis of single-cell RNA-seq (scRNA-seq) data has become a popular method that attempts to infer lost temporal information, such as a cell's differentiation state<sup>1,2</sup>. Such analyses reconstruct a measure of a cell's progression through some biological process, known as a *pseudotime*. Recently, attention has turned to modeling bifurcations where, part-way along such trajectories, cells undergo some fate decision and branch into two or more distinct cell types.

Several methods have been proposed to infer bifurcation structure from single-cell data. Wishbone<sup>3</sup> constructs a  $k$ -nearest neighbor graph and uses shortest paths from a *root* cell to define pseudotimes, using inconsistencies over multiple paths to detect bifurcations. Diffusion Pseudotime (DPT)<sup>4</sup> similarly constructs a transition matrix where each entry may be interpreted as a diffusion distance between two cells. Bifurcations are inferred by identifying the anti-correlation structure of random walks from both a root cell and its maximally distant cell. While DPT arguably has a probabilistic interpretation, neither method specifies a fully generative model that incorporates measurement noise, while both infer bifurcations retrospectively after constructing pseudotimes. A further algorithm Monocle<sup>5</sup> learns pseudotimes based on dimensionality reduction using the DDRTree algorithm<sup>6</sup> and provides post-hoc inference of genes involved in the bifurcation process using generalized linear models.

Here we propose a Bayesian hierarchical mixture of factor analyzers for inferring bifurcations from single-cell data. Factor analysis and its close relative principal component analysis (PCA) are frequently used in the context of single-cell gene expression modeling, both for visualization and trajectory inference (see e.g. 7,8). Since developmental bifurcations involve two related processes, it is therefore natural to extend such models to involve a mixture of two factor analyzers in a Bayesian hierarchical setting that relates expression patterns between branches.

The model we propose is unique compared to existing bifurcation inference methods in the following: (1) by specifying a fully generative probabilistic model we incorporate measurement noise into inference and provide full uncertainty estimates for all parameters; (2) we simultaneously infer cell "pseudotimes" and branching structure as opposed to post-hoc branching inference as is typically performed; and (3) our hierarchical shrinkage prior structure automatically detects features involved in the bifurcation, providing statistical support for detecting which genes drive fate decisions.

In the following, we introduce our model and apply it to both synthetic datasets and demonstrate its consistency with existing algorithms on real single-cell data. We further propose a zero-inflated variant that takes into account zero-inflation, and quantify the levels of dropout at which such models are beneficial. We highlight the multiple natural solutions to bifurcation inference when using gene expression data alone and finally discuss both the merits and drawbacks of using such a unified probabilistic model.

## Methods

### Statistical model

We begin with an  $N \times G$  matrix of suitably normalized gene expression measurements for  $N$  cells and  $G$  genes, where  $\mathbf{y}_i$  denotes the  $i^{\text{th}}$

row vector corresponding to the expression measurement of cell  $i$ . We assign a pseudotime  $t_i$  to each cell, along with a binary variable  $\gamma_i$  indicating to which of  $B$  branches cell  $i$  belongs:

$$\gamma_i = b \text{ if cell } i \text{ on branch } b \quad (1)$$

with  $b \in 1, \dots, B$ .

The pseudotime  $t_i$  is a surrogate measure of a cell's progression along a trajectory while it is the behavior of the genes - given by the factor loading matrix - that changes between the branches. We therefore introduce  $B$  factor loading matrices  $\Lambda_b = [\mathbf{c}_b \ \mathbf{k}_b]$ ,  $b \in 1, \dots, B$  for each branch modeled.

The likelihood of a given cell's gene expression measurement conditional on all the parameters is then given by

$$\mathbf{y}_i | \gamma_i, \Lambda_{\gamma_i}, t_i, \boldsymbol{\tau} \sim \text{Normal}(\mathbf{c}_{\gamma_i} + \mathbf{k}_{\gamma_i} t_i, \boldsymbol{\tau}^{-1} \mathbb{1}_G) \quad (2)$$

where  $\mathbb{1}_G$  is the  $G \times G$  identity matrix.

We motivate the prior structure as follows: if the bifurcation processes share some common elements then the behavior of a non-negligible subset of the genes will be (near) identical across branches. It is therefore reasonable that the factor loading gradients  $\mathbf{k}_\gamma$  should be similar to each other unless the data suggests otherwise. We therefore place a prior of the form

$$\mathbf{k}_{\gamma_i} \sim \text{Normal}(\boldsymbol{\theta}, \boldsymbol{\chi}^{-1} \mathbb{1}_G) \quad (3)$$

where  $\boldsymbol{\theta}$  denotes a common factor gradient across branches. This has similar elements to Automatic Relevance Determination (ARD) models with the difference that rather than shrinking regression coefficients to zero to induce sparsity, we shrink factor loading gradients towards a common value to induce similar behavior between mixture components. We can then inspect the posterior precision to identify genes involved in the bifurcation: if  $\chi_g$  is very large then the model is sure that  $k_{0g} \approx k_{1g}$  and gene  $g$  is not involved in the bifurcation; however, if  $\chi_g$  is relatively small then  $|k_{0g} - k_{1g}| \gg 0$  and the model indicates that  $g$  is involved in the bifurcation.

With these considerations the overall model is given by the following hierarchical (M)ixtures of (F)actor (A)nalyzers (MFA) specification:

$$\begin{aligned} \boldsymbol{\omega} &\sim \text{Dirichlet}(1/B, \dots, 1/B) \\ \gamma_i &\sim \text{Categorical}(\boldsymbol{\omega}) \\ \eta &\sim \text{Normal}(\bar{\eta}, \tau_\eta^{-1}) \\ \boldsymbol{\theta}_g &\sim \text{Normal}(\bar{\boldsymbol{\theta}}, \tau_\theta^{-1}) \\ \chi_g &\sim \text{Gamma}(\alpha_x, \beta_x) \\ \mathbf{c}_{\gamma_i} &\sim \text{Normal}(\boldsymbol{\eta}, \tau_c^{-1}) \\ \mathbf{k}_{\gamma_i} &\sim \text{Normal}(\boldsymbol{\theta}, \boldsymbol{\chi}^{-1} \mathbb{1}_G) \\ t_i &\sim \text{Normal}(0, 1) \\ \boldsymbol{\tau} &\sim \text{Gamma}(\alpha, \beta) \\ \mathbf{y}_i &\sim \text{Normal}(\mathbf{c}_{\gamma_i} + \mathbf{k}_{\gamma_i} t_i, \boldsymbol{\tau}^{-1} \mathbb{1}_G) \end{aligned} \quad (4)$$

where  $\tilde{\eta}$ ,  $\tilde{\theta}$ ,  $\tau_\eta$ ,  $\tau_\theta$ ,  $\tau_c$ ,  $\alpha_\chi$ ,  $\beta_\chi$ ,  $\alpha$  and  $\beta$  are hyperparameters fixed by the user. By default we set the non-informative prior  $\alpha_\chi = \beta_\chi = 10^{-2}$  to maximize how informative the posterior of  $\chi$  is in identifying genes that show differential expression across the branches.

As the model exhibits complete conditional conjugacy, inference was performed using Gibbs sampling (Supplementary File 1). Details of computer software (MFA) implementing these methods is given in Software availability<sup>9</sup>.

### Modeling zero-inflation

Single-cell data is known to exhibit *dropout* where the failure to reverse-transcribe lowly expressed mRNA results in zero counts in the expression matrix. The issue has been extensively studied in the context of scRNA-seq, resulting in algorithms that take into account the resulting zero inflation, such as ZIFA<sup>7</sup> or SCDE<sup>10</sup>.

We can incorporate tractable zero-inflation into our model by considering a per-gene dropout probability given by

$$p(\text{dropout in gene } g) = \exp\left(-\frac{\lambda}{N} \sum_{i=1}^N x_{ig}\right) \quad (5)$$

where  $x_{ig}$  is the unobserved true expression of gene  $g$  in cell  $i$  and  $\lambda$  is a global dropout parameter estimated in an Empirical-Bayes manner. This exponential model empirically fits multiple scRNA-seq datasets well (Supplementary File 1). Incorporating this zero-inflated likelihood modifies the model in 4 to

$$\begin{aligned} x_i &\sim \text{Normal}(\mathbf{c}_{y_i} + \mathbf{k}_{y_i} t_i, \boldsymbol{\tau}^{-1} \mathbb{1}_G) \\ h_{ig} &\sim \text{Bernoulli}\left(\exp\left(-\frac{\lambda}{N} \sum_i x_{ig}\right)\right) \\ y_{ig} &= \begin{cases} x_{ig} & \text{if } h_{ig} = 0 \\ 0 & \text{if } h_{ig} = 1 \end{cases} \end{aligned} \quad (6)$$

While incorporating zero-inflation in the likelihood leads to a less-misspecified model, we must perform inference on an additional  $N_0$  parameters, where  $N_0$  is the number of zero measurements in the expression matrix. For single cell RNA-seq data this can be as high as 90% of all measurements, leading to a significant additional computational burden.

Furthermore, such a dropout model assumes a per-gene dropout probability dependent on the mean latent expression, though in reality the dropout probability would depend on the latent expression itself. This compromise allows us to estimate the parameter  $\lambda$  by fitting for each gene the proportion of cells expressed versus the mean expression.

### Multiple solutions to bifurcation inference

It is common in bifurcation inference methods to specify additional information aside to gene expression data alone. For example, Wishbone requires the specification of a *root* cell that signifies the beginning of pseudotime. DPT also allows for the specification of a root

cell or picks the furthest from a random cell if unspecified. Monocle equivalently allows re-fitting of the pseudotimes with the constraint that one of the inferred ‘states’ is the initial or root state.

We argue that such requirements are necessary due to a fundamental invariance in the gene expression of bifurcating cells. Figure 1 shows a conceptual model of three end-states (1–3) and a gene that is expressed in one end state (2), but not the others. We can envisage three possible bifurcation routes here: state 1 is the initial state that bifurcates to 2 & 3 (1 → 2, 3), or equivalently 3 → 1, 2 or 2 → 1, 3. If 1 or 3 is the initial state then the gene exhibits differential expression across the branches, while if we start at 2 the gene exhibits concordant expression across the branches. Note that for a bifurcation we require some genes that show differential expression between the branches and some that show concordant expression - lacking the former would give a non-branching trajectory and lacking the latter would give separate cell types.

The above reasons that in a single-gene case the initial state is indistinguishable from the gene expression alone. We can easily generalize this to the multiple-gene case, due to the fact that the labels in Figure 1 are statistically non-identifiable. The equivalent geometric argument is that you can ‘spin’ Figure 1 about  $\mathbf{B}$  for each gene (and optionally invert the expression to give two states of non-zero expression).

While in algorithms, such as Wishbone and DPT, this non-identifiability is solved by setting an initial cell or state, the equivalent in our model is the correct initialization of the pseudotimes. PCA is applied to the data before inference and the principal component that best corresponds to the trajectory based on the expression of known genes is used to initialize the pseudotimes. Such trajectories correspond to local modes in the posterior space that are sufficiently narrow the probability of the Gibbs sampler moving to another local mode is negligible. A future extension that would solve this non-identifiability would involve placing priors on the behavior of certain genes across the branches, which combined with more efficient inference would pick out the ‘true’ trajectory.

Please note that an earlier version of this article can be found on bioRxiv (doi: [10.1101/076547](https://doi.org/10.1101/076547)).

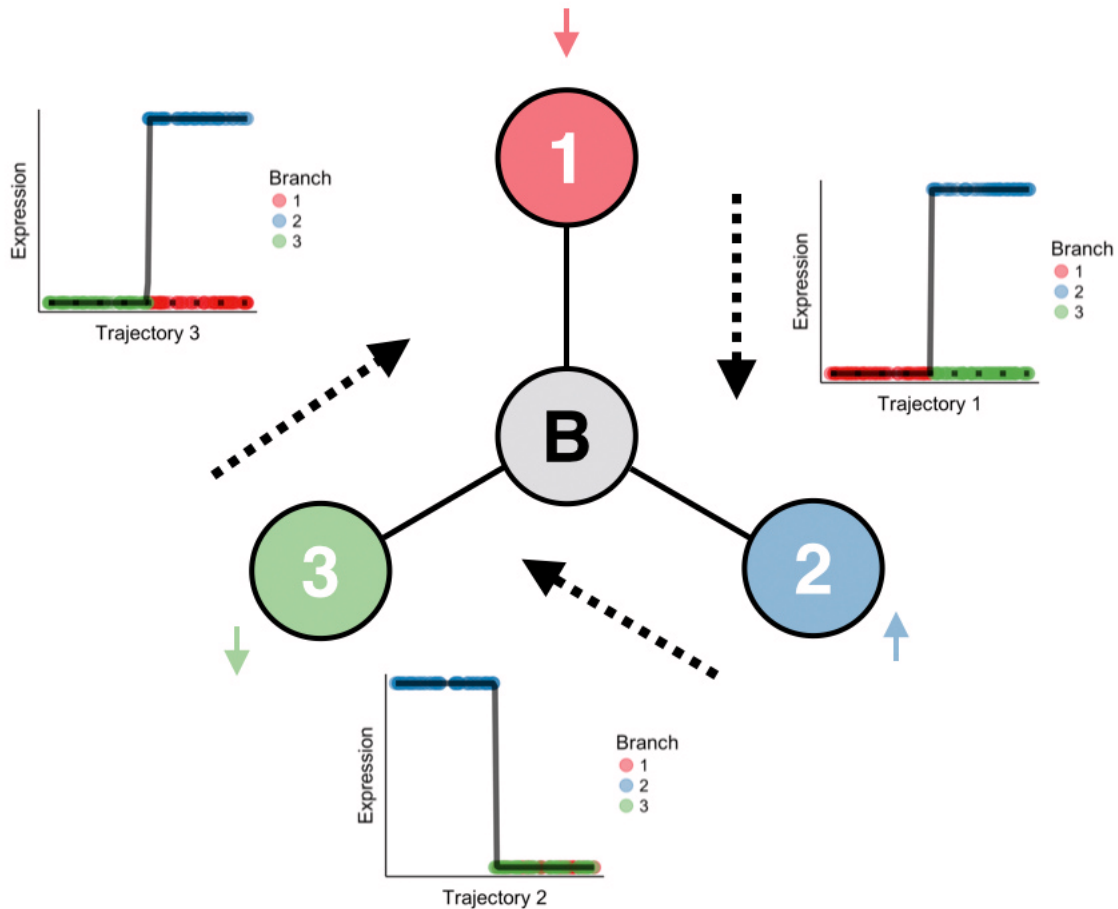
## Results

### Synthetic datasets

We first demonstrate our method on a synthetic ‘toy’ dataset of 300 bifurcating cells and 60 genes, half of which exhibit differential behavior across the bifurcation and half of which show similar behavior.

Our synthetically generated data is mildly mis-specified with respect to our model to demonstrate robustness when using real genomic data. For example, the generated gene behavior across pseudotime is sigmoidal, which we have previously successfully used to model real single-cell datasets<sup>11,12</sup>.

Pseudotimes were inferred using Gibbs sampling (Supplementary File 1) for  $10^5$  iterations. PCA representations of the synthetic data can be seen in Figures 2A and B, showing the characteristic  $Y$

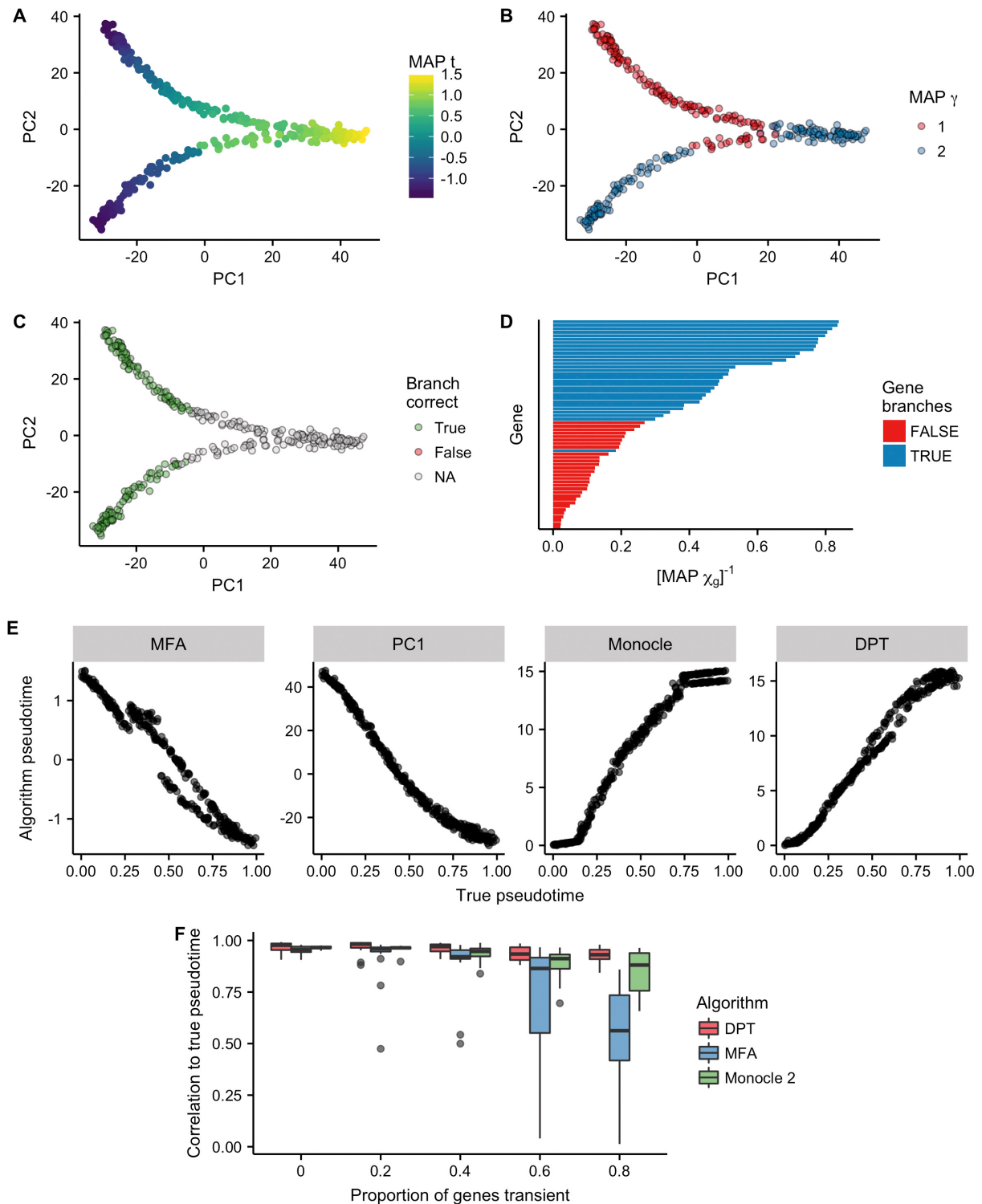


**Figure 1. Multiple solutions to bifurcation inference.** Starting with three cell states, we would like to infer a bifurcation process from one to the other two. If a single gene is up-regulated in one of the states, yet down-regulated in the other two, then clearly any state may act as the beginning of the trajectory. For example, if we start in state 1 then the gene is up-regulated along state 2 and stays constant in state 3; if we start in state 2 then the gene is down-regulated in states 1 & 3; if we start in state 3 then the gene is up-regulated in state 2 and remains down-regulated in state 1. However, due to the non-identifiability this is true if we add additional genes that are up-regulated in one or two of the cell states. The equivalent geometric argument is that we can build the transcriptomic profiles across all genes by spinning the figure about B (with possible inversion) and “adding” that gene. No matter how many additional genes we add, any one of the three states can act as the root state or beginning of pseudotime. Therefore, in the absence of any additional information there are always three equally valid solutions to bifurcation inference from gene expression data alone.

shape associated with bifurcating data, colored by both maximum *a posteriori* (MAP) pseudotime and branch assignment estimates, respectively. We compared the Pearson correlation of the estimated pseudotimes to the true pseudotimes (Figure 2C) for both MFA, PC1 (the first principal component of the data), Monocle and Diffusion Pseudotime, giving values of 0.98, 0.98, 0.98 and 0.99 (to 2 s.f.), respectively. Broad benchmarking of pseudotime algorithms to “ground-truth” data is difficult, due to the inherent assumptions that are necessary about how genes expression evolves along trajectories. However, such toy examples demonstrates the consistency of multiple algorithms on our toy dataset.

One weakness of our model is that it assumes gene expression changes as a linear function of time. This allows us to perform fast conjugate Gibbs sampling, but is highly unrealistic for real data. The synthetic data generated is based on sigmoidal changes

across pseudotime, which being nonlinear is already mildly mis-specified with respect to our model. However, genes may also exhibit transient behavior, in which they are briefly down- or up-regulated before returning to their initial state. We sought to quantify the robustness of MFA to transient gene expression by performing extensive simulations. Specifically, we generated synthetic datasets with 0%, 20%,..., 80% of genes exhibiting transient expression, and inferred the pseudotimes using DPT, MFA and Monocle 2. This was repeated 20 times for each percentage of transient genes. The results can be seen in Figure 2F. The performance of MFA remains competitive up to around 40% of genes exhibiting transient expression, after which DPT and Monocle 2 perform significantly better. However, MFA is highly consistent with DPT and Monocle 2 on the two real datasets examined (Figure 4 and Figure 5), implying the occurrence of transient expression is limited enough in practice for the linearity assumption to be feasible.



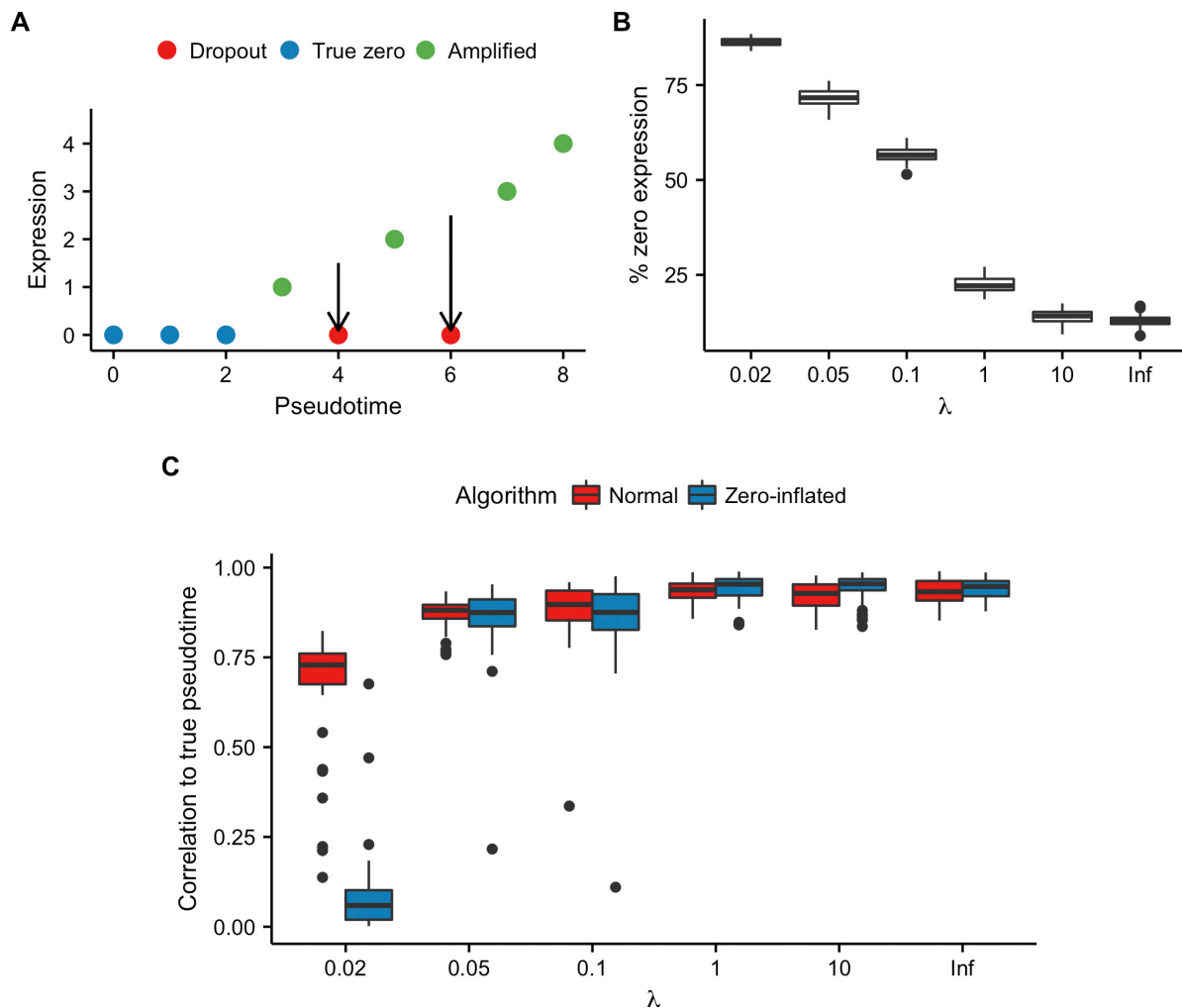
**Figure 2. Probabilistic inference of bifurcations in synthetic data.** **A** Principal component analysis representation of a toy dataset for 300 cells and 60 genes, colored by the maximum *a posteriori* (MAP) pseudotime estimates. **B** Equivalent representation as **(A)** color by the MAP branch estimate. **C** Equivalent representation showing whether each branch was assigned correctly. Due to the non-identifiability of mixture components, we map component indices from true to inferred such that the agreement is maximized. **D** The inverse MAP estimates of  $\chi$  largely identify which genes in the dataset exhibit different behavior across the two branches. **E** Comparison of different pseudotime inference algorithms to the ground truth pseudotime on this particular dataset. The algorithms MFA, PC1 (principal component 1), Monocle and DPT had correlations of 0.98, 0.98, 0.98, 0.99 (to 2 s.f.), respectively. **F** The correlation of inferred pseudotimes to ground truth depending on the proportion of genes in the dataset exhibiting transient behavior. MFA shows competitive performance up to around 40% of genes begin transient despite an inherent linear assumption.

One notable difference between MFA and existing bifurcation inference algorithms is in the pre-bifurcation branch assignment. Algorithms, such as Wishbone and DPT, will assign a separate branch to cells preceding the bifurcation. However, MFA will typically assign pre-bifurcation cells to one of the two branches modeled, with the other branch beginning at the bifurcation. A bifurcation process consists of two temporal processes that have a common origin but differing end points. Thus, due to nonidentifiability, cells pre-bifurcation can equally be said to be on one branch with the second beginning at the bifurcation point. Importantly, no matter how we assign the branches under this regime, the observed behavior of genes as a function of both pseudotime and branch assignment will be consistent, which is necessary for biological insight.

### Benefits of modeling zero-inflation

Single-cell RNA-seq data is known to exhibit *dropout*, where a failure to reverse transcribe lowly-expressed mRNA results in zero counts. We have created a variant of MFA that employs an Empirical-Bayes like approach to account for such dropout (see *Methods*). However, a zero count for a particular gene in a particular cell may also be a *true zero* where no mRNA in the cell is present.

We expect such true zeros to be useful for pseudotime inference. [Figure 3A](#) shows a conceptual model where a gene is up-regulated along pseudotime with two cells exhibiting dropout. The true zeros (in blue) help pseudotime inference as the low-expression implies they are at the beginning of pseudotime. However,



**Figure 3. Effects of modeling zero-inflation.** **A** Zero counts observed in single-cell RNA-seq data may be attributed to either *true zeros*, where no mRNA of a given gene is produced in a cell, or *dropout*, where there is a failure to reverse-transcribe the low levels of starting material. Alternatively, a count is registered and the gene is *amplified*. In theory not accounting for dropouts will reduce the accuracy of pseudotime inference the two red counts at pseudotimes of 4 and 6 would be ordered with the blue counts. However, in practice it is impossible to distinguish between *dropouts* and *true zeros*. **B** The percentage of counts with zero expression across 50 replicates for each value of  $\lambda$  used in dropout simulations. **C** The Pearson correlation to true pseudotime using both the non-zero-inflated and zero-inflated variants of MFA as a function of  $\lambda$  used to generate the dataset. Accounting for zero-inflation shows marginal benefits if only a small percentage counts are dropouts. However, for high dropout percentages (> 80%) the algorithm has to “impute” such a large percentage of the data that correlations to the true pseudotime reduce to near-zero.

the cells exhibiting dropout (in red) would potentially impede pseudotime inference as MFA would order them with the true zero cells at the beginning of the trajectory.

Accounting for such dropouts involves modifying the model so that zero counts are likely if the underlying latent expression is low. Therefore, the red dropout cells in [Figure 3A](#) would be effectively imputed (via Gibbs updates) upwards towards the mean expression line, increasing the accuracy of pseudotime inference. However, as there is no way to distinguish between true zeros and dropouts, we also “impute” the expression of the true zeros, which may itself decrease the accuracy of pseudotime inference.

We sought to quantify the benefits of modeling zero inflation against the drawbacks of losing the information contained in “true zeros”. We created multiple synthetic datasets ([Supplementary File 1](#)), while varying the dropout parameter  $\lambda \in \{0.02, 0.05, 0.1, 1, 10, \infty\}$ , where  $\lambda = 0.02$  has the largest levels of dropout, while  $\lambda = \infty$  has no dropout, only true zeros. This was repeated 50 times for each  $\lambda$ , and the proportion of zero counts in each dataset can be seen in [Figure 3B](#). We subsequently re-inferred the pseudotimes using MFA with both the zero-inflated and standard variants.

The resulting correlations with the true pseudotimes across the range of  $\lambda$  and MFA variants can be seen in [Figure 3C](#). At very high levels of dropout ( $\lambda = 0.02$ , where  $> 80\%$  of counts are zeros) the zero-inflated variant performs considerably worse than the non-zero-inflated variant, with virtually no correspondence to the true pseudotimes compared to  $\rho \approx 0.75$ . We suggest this is due to the inference procedure, effectively imputing such a large proportion of the data that there are too many degrees of freedom to effectively infer the trajectory. For the remaining values of  $\lambda$  the zero-inflated variant infers pseudotimes largely comparable to those of the non-zero inflated version, with marginal improvements in accuracy when there is moderate dropout ( $\lambda = 1, 10$ ). We conclude that incorporating zero-inflation into pseudotime inference is sensible, but the variable quality across the (unknown in practice) dropout range along with considerable additional computational cost render it unnecessary for most practical purposes.

#### Application to single-cell RNA-seq data

We next applied our method to previously published single-cell RNA-seq data of 4,423 hematopoietic progenitor/stem cells, differentiating into myeloid and erythroid precursors<sup>13</sup>.

To reduce the dataset to a computationally feasible size we used only genes expressed in at least 20% of cells with a variance in normalized expression greater than 5. We performed Gibbs sampling for  $4 \times 10^4$  iterations using default hyperparameter values, except for  $\tau_\theta = \tau_\eta = 1$ , and initialized the pseudotimes to the second principal component of the data. The results can be seen in [Figures 4\(A and B\)](#). The MAP pseudotime estimates clearly recapitulates the trajectory in the data, as shown using a tSNE representation from [3](#), while the MAP estimates of  $\gamma_i$  detects the branching structure in the data, consistent with previous methods.

We went on to analyze the genes suggested by the model to be involved in the bifurcation process. [Figure 4C](#) shows the inverse

posterior mean of  $\chi_g$ , with larger values indicating more evidence that gene  $g$  is involved in the bifurcation process. For illustration purposes, we plot the expression of *ELANE* and *CAR2*, which the model suggests will show differential behavior across the bifurcation, along with *RPL26*, which the model suggests will show common behavior ([Figure 4D](#)).

We next sought to compare the performance of MFA to existing bifurcation inference algorithms, in particular Wishbone, DPT and Monocle (v2), along with the second principal component of the data (PC2), which we noted from exploratory analyses was highly correlated with the existing Wishbone values. We sub-sampled down to 1,000 cells for Monocle comparisons for computational convenience and used the previously published results for Wishbone (from [3](#)). The root cell for DPT was selected as the cell with the minimum value for the second principal component and similarly the root state for Monocle was chosen such that it contained that cell. Otherwise, algorithms were run with default parameters.

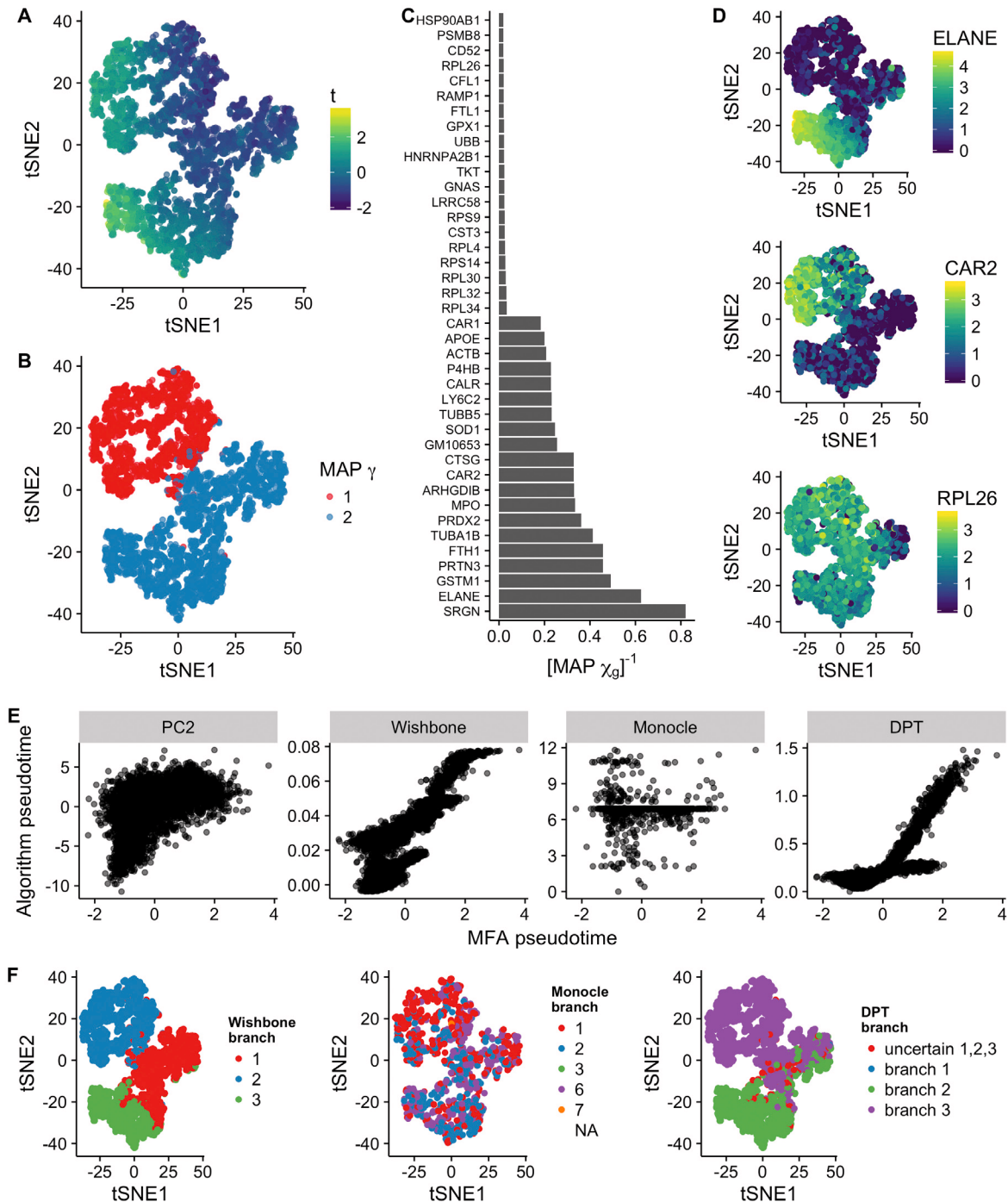
The comparison of the inferred pseudotimes with that MFA can be seen in [Figure 4E](#). There is high correlations with PC2 ( $\rho = 0.54$ ), Wishbone ( $\rho = 0.83$ ), and DPT ( $\rho = 0.78$ ). However, there is virtually no correlation with Monocle ( $\rho = 0.01$ ), though as this low correlation only occurs with Monocle we assume it is not an issue with MFA. We also sought to compare branch allocations across the algorithms, which is difficult due to the non-identifiability of the statistical models involved. [Figure 4F](#) shows a tSNE representation of the cells colored by branch allocation for each of Wishbone, Monocle and DPT. We see that MFA is largely consistent with Wishbone and DPT, detecting a bifurcation at the “pinch” in the tSNE plot, but as with the pseudotimes there is barely any correspondence in branch allocations with Monocle (which, as of version 2, does not allow pre-specification of the number of branches to model).

#### Application to single-cell mass-cytometry data

We next applied MFA to single-cell mass cytometry data, tracking the differentiation of 22,850 monocytes and erythrocytes from hematopoietic stem and progenitor cells across 12 markers as published in [14](#) and previously analyzed in [3](#). For computational convenience with all algorithms, we sub-sampled the data down to 2,000 randomly chosen cells, with the exception of Monocle, which we subsequently sub-sampled further down to 1,000 cells. We found that due to the small number of proteins measured there was too much freedom for the MFA model to infer mixtures using the default parameter settings. We therefore had to encourage large levels of similarity across the two branches by setting  $\alpha_\chi = 5 \times 10^3$  and  $\beta_\chi = 1$ .

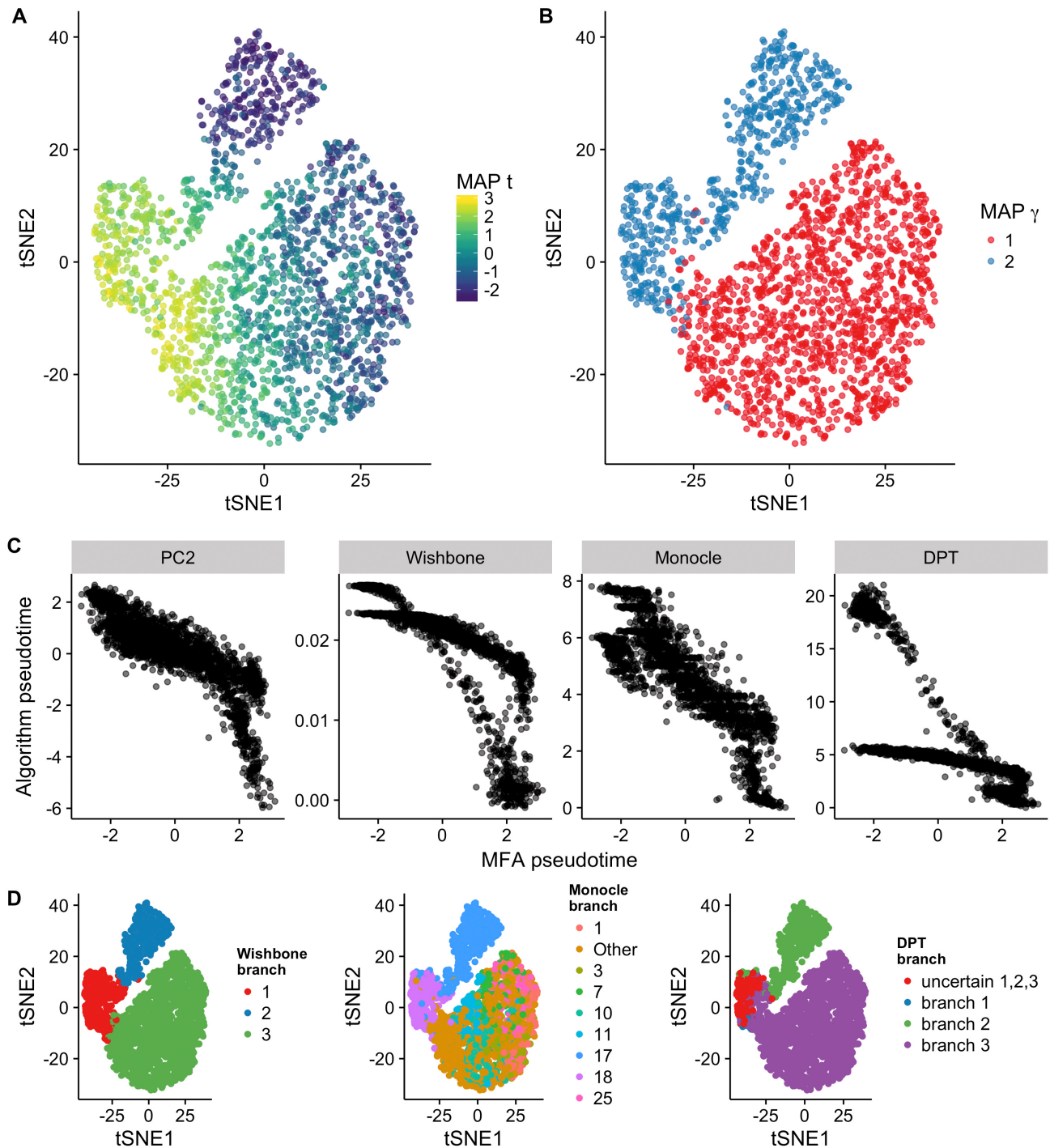
The results can be seen in [Figure 5](#). [Figure 5A](#) shows a tSNE representation (as published in [3](#)) showing the inferred MAP pseudotimes correctly following the left-right trajectory, while [Figure 5B](#) correctly shows the MAP  $\gamma$  values identifying a bifurcation at the “pinch” in the plot.

We subsequently compared the inferred pseudotimes and branching to those found using the alternative algorithms. We found good correspondence to all other methods ([Figure 5C](#)),



**Figure 4. Inference of bifurcations in scRNA-seq data of 4,423 hematopoietic progenitor/stem cells differentiating into myeloid and erythroid precursors<sup>3</sup>.** **A** tSNE representation colored by the maximum *a posteriori* (MAP) pseudotime. **B** Equivalent plot as **A** colored by MAP  $\gamma$  (branch assignment). **C** Inverse map  $\chi$  showing both the 20 largest and 20 smallest values indicating which genes do and do not show differential behavior across the bifurcation. **D** tSNE representation of the dataset colored by gene expression. Both *ELANE* and *CAR2* were predicted by the inverse  $\chi$  values to show differing expression across the branches, while *RPL26* was predicted to show similar expression. **E** Scatter plots of pseudotime values compared to those inferred by PC2, Wishbone, Monocle, and DPT. These had Pearson correlations of 0.54, 0.83, 0.01, and 0.78, respectively. **F** tSNE representations of the dataset colored by branch allocation of alternative algorithms shows good agreement with Wishbone and DPT.





**Figure 5. Inference of bifurcations in single-cell mass cytometry data of a subsample of 2,000 hematopoietic progenitor/stem cells differentiating into monocyte and erythrocyte progenitors.** **A** A tSNE representation colored by the maximum *a posteriori* (MAP) pseudotime. **B** Equivalent plot as **A** colored by MAP  $\gamma$  (branch assignment). **C** Scatter plots of MFA pseudotime compared to PC2, Wishbone, Monocle, and DPT, with Pearson correlations of 0.84, 0.86, 0.80 and 0.69 respectively. **D** tSNE representation colored by branch assignment of Wishbone, Monocle, and DPT. As of version 2, Monocle does not allow for the number of branches to be selected *a priori* and typically returns a large number. For the convenience of visualization we therefore only display the 30% most frequent states and group the remaining infrequent ones into “Other”. The figures suggest a good agreement of branch assignment of MFA with Wishbone and DPT, and moderate agreement with Monocle.

with Pearson correlations of 0.84, 0.86, 0.80 and 0.69 for PC2, Wishbone, Monocle, and DPT, respectively. We further compared the branch assignment of MFA to those of the alternative algorithms (Figure 5D). As of version 2, Monocle does not allow for the number of branches to be selected *a priori* and typically returns a large number. For the convenience of visualization we therefore only display the 30% most frequent states and group the remaining infrequent ones into “Other”. We find good agreement between MFA and Monocle and DPT, and similarities with the Monocle assignments (MFA branch 2 loosely corresponds to Monocle branch 17).

## Discussion

In this paper we have presented a Bayesian hierarchical mixture of factor analyzers for inference of bifurcating trajectories in single-cell data. Our model is unique compared to existing efforts in that it (a) is fully generative, incorporating measurement noise into inference, (b) jointly infers both the pseudotimes and branches compared to post-hoc inference of branch detection, and (c) jointly infers which genes are differentially regulated across the branches. We also proposed an extension that accounts for the high levels of zero-inflation present in single-cell RNA-seq data. We applied our model to a range of synthetic and real datasets and demonstrated it performs competitively with existing methods.

There is a natural trade-off in designing such models between flexibility and practicality. The implicit assumption of MFA that gene expression develops linearly across pseudotime allows for fast Markov-Chain Monte Carlo sampling and joint inference of branch structure. However, it is potentially highly mis-specified: the predicted expression can become negative leading to erroneous inference (see Supplementary File 1). A solution to this would be to not explicitly assume a strongly parametric form of gene expression and consider nonparametric methods. However, such methods are often overly flexible, requiring either additional capture information to correctly infer pseudotimes<sup>15</sup> or hard-setting the pseudotimes prior to inferring the branching structure<sup>16</sup>. As such there is a natural trade-off between the expressivity of such models and being able to perform valid statistical inference that fully incorporates parameter variation without additional constraints or “tweaking”.

## Supplementary material

**Supplementary File 1: Further methods and analysis.** This file contains (1) additional technical details of the statistical inference methods, (2) further description of the zero-inflation, (3) details of the simulation strategy and (4) a discussion of the limitations of the linear model.

[Click here to access the data.](#)

There are several extensions that can be applied to our model. While the model performs well on large single cell RNA-seq datasets, it could be scaled up further using Stochastic Variational Inference<sup>17</sup>, which due to the model’s conditionally conjugate structure could be implemented without resorting to approximations. As mentioned previously, one main weakness of the model is the unrealistic assumption of linear changes in expression over pseudotime, leading to severe model specification. One could therefore consider alternative nonlinear functions, such as sigmoids (previously used in 8), or nonparametric models such as Gaussian Process Latent Variable Models (previously used in 15,18), with appropriate structural constraints.

## Software availability

MFA software available from: <http://www.github.com/kieran-rcampbell/mfa>

Archived source code as at time of publication: doi, [10.5281/zenodo.3459819](https://doi.org/10.5281/zenodo.3459819)

License: GNU General Public License (GPL)

---

## Author contributions

KRC and CY conceived the study. KRC developed software and performed computer simulations. KRC and CY wrote the manuscript.

## Competing interests

No competing interests were disclosed.

## Grant information

This work was supported by the Wellcome Trust Core Award [090532]; a UK Medical Research Council funded doctoral studentship to KC; a UK Medical Research Council New Investigator Research Grant to CY [MR/L001411/1]; the John Fell Oxford University Press Research Fund to CY; the Li Ka Shing Foundation via a Oxford-Stanford Big Data in Human Health Seed Grant to CY.

*The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

## References

1. Wagner A, Regev A, Yosef N: **Revealing the vectors of cellular identity with single-cell genomics.** *Nat Biotechnol.* 2016; **34**(11): 1145–1160.  
[PubMed Abstract](#) | [Publisher Full Text](#)
2. Bacher R, Kendziorski C: **Design and computational analysis of single-cell RNA-sequencing experiments.** *Genome Biol.* 2016; **17**(1): 63.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
3. Setty M, Tadmor MD, Reich-Zeliger S, *et al.*: **Wishbone identifies bifurcating developmental trajectories from single-cell data.** *Nat Biotechnol.* 2016; **34**(6): 637–645.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
4. Haghverdi L, Büttner M, Wolf FA, *et al.*: **Diffusion pseudotime robustly reconstructs lineage branching.** *Nat Methods.* 2016; **13**(10): 845–8.  
[PubMed Abstract](#) | [Publisher Full Text](#)
5. Qiu X, Hill A, Packer J, *et al.*: **Single-cell mRNA quantification and differential analysis with census.** *Nat methods.* 2017; **14**(3): 309–315.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
6. Mao Q, Wang L, Tsang IW, *et al.*: **A novel regularized principal graph learning framework on explicit graph representation.** 2015; arXiv preprint arXiv: 1512.02752.  
[Reference Source](#)
7. Pierson E, Yau C: **ZIFA: Dimensionality reduction for zero-inflated single-cell gene expression analysis.** *Genome Biol.* 2015; **16**(1): 241.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
8. Campbell K, Yau C: **Ouija: Incorporating prior knowledge in single-cell trajectory learning using bayesian nonlinear factor analysis.** *bioRxiv.* 2016; 060442.  
[Publisher Full Text](#)
9. Campbell KR, Yau C: **kieranrcampbell/mfa: Bioconductor-ready version [Data set].** *Zenodo.* 2017.  
[Data Source](#)
10. Kharchenko PV, Silberstein L, Scadden DT: **Bayesian approach to single-cell differential expression analysis.** *Nat Methods.* 2014; **11**(7): 740–742.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
11. Campbell KR, Yau C: **Order Under Uncertainty: Robust Differential Expression Analysis Using Probabilistic Models for Pseudotime Inference.** *PLoS Comput Biol.* 2016; **12**(11): e1005212.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
12. Campbell KR, Yau C: **switchde: inference of switch-like differential expression along single-cell trajectories.** *Bioinformatics.* 2016; pii: btw798.  
[PubMed Abstract](#) | [Publisher Full Text](#)
13. Paul F, Arkin Y, Giladi A, *et al.*: **Transcriptional Heterogeneity and Lineage Commitment in Myeloid Progenitors.** *Cell.* 2015; **163**(7): 1663–1677.  
[PubMed Abstract](#) | [Publisher Full Text](#)
14. Bendall SC, Simonds EF, Qiu P, *et al.*: **Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum.** *Science.* 2011; **332**(6030): 687–696.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
15. Reid JE, Wernisch L: **Pseudotime estimation: deconfounding single cell time series.** *Bioinformatics.* 2016; **32**(19): 2973–80.  
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
16. Lönnberg T, Svensson V, James KR, *et al.*: **Single-cell RNA-seq and computational analysis using temporal mixture modeling resolves  $T_H1/T_{FH}$  fate bifurcation in malaria.** *Sci Immunol.* 2017; **2**(9): p.eaal2192.  
[Publisher Full Text](#)
17. Hoffman MD, Blei DM, Wang C, *et al.*: **Stochastic variational inference.** *Journal of Machine Learning Research.* 2013; **14**(1): 1303–1347.  
[Reference Source](#)
18. Campbell K, Yau C: **Bayesian gaussian process latent variable models for pseudotime inference in single-cell rna-seq data.** *bioRxiv.* 2015; 026872.  
[Publisher Full Text](#)

# Open Peer Review

Current Referee Status:  

---

## Version 1

Referee Report 09 May 2017

doi:[10.21956/wellcomeopenres.11959.r21991](https://doi.org/10.21956/wellcomeopenres.11959.r21991)



### Luca Pinello

Department of Pathology, Massachusetts General Hospital and Harvard Medical School, Boston, MA, USA

Campbell and Yau present a probabilistic method called MFA to infer bifurcating trajectories and differentially regulated genes across branches from single cell transcriptomic data.

The method is based on a Bayesian hierarchical mixture of factor analyzers. The authors also discuss an extension of this method to deal with dropout events commonly observed in scRNA-seq datasets. Although they claim that the model obtained with this procedure is less misspecified, the computational requirements associated with it may be a burden, especially for very sparse gene expression matrices and unnecessary for practical purposes.

MFA is evaluated on a synthetic dataset, scRNA-seq and mass cytometry data and compared with existing methods, although limited to Wishbone, Monocle2 and Diffusion Pseudotime.

Overall, the manuscript is well written, the results clearly described and the source code provided.

It may be helpful to consider/clarify the following points to improve the manuscript:

1. The computational requirements of MFA are not discussed in depth and running times are not reported. Please add a table comparing the execution times of the different methods tested using the full datasets presented and not the down-sampled versions. If a method fails to run in reasonable time, just report this fact (and if possible show the speed-up of MFA using down-sampled datasets).
2. For different datasets, different values of (hyper)parameters are used but no clear guidelines are provided to the user in how to set those values. Please explain the rationale and if possible clear metrics that the users can use for tuning those parameters.
3. *“PCA is applied to the data before inference and the principal component that best corresponds to the trajectory based on the expression of known genes is used to initialize the pseudotimes”*. Please explain how to initialize the pseudo-time in absence of known genes.
4. I agree with Dr. Gitter about adding a short intro to factor analysis before jumping to equation 2.
5. Please describe (or better show with a synthetic dataset) how MFA performs when more than one branching point is present. For example, using the synthetic dataset presented in Rizvi et al 2017

Nat Biotech (see Figure 2).

6. The authors claim that explicitly modeling the dropout events doesn't always justify the computational cost. I think it may be worth to test this idea on real datasets, especially droplet based (drop-seq, in-drop or 10x genomics) in which this problem is more pronounced. Good candidate datasets to show how the method performs in those settings are presented in van Dijk et al.<sup>1</sup> where their imputation strategy clearly improve pseudotime estimations or Zheng et al. 2016 Nat Comm.
7. Application to scRNA-seq:
  - 1) What is the running time without down-sampling? How comparable are the results with or without down-sampling?
  - 2) Two genes (*ELANE* and *CAR2*) are presented to illustrate the bifurcation process, what other genes are significant by this analysis? It may be worth to show in a sup table the ranking obtained for each branch using MFA (are *ELANE* and *CAR2* on top?).
  - 3) "The comparison of the inferred pseudotimes with that MFA can be seen in Figure 4E. There is high correlations with PC2 ( $\rho = 0.54$ ), Wishbone ( $\rho = 0.83$ ), and DPT ( $\rho = 0.78$ ).". Please describe more explicitly how the correlation is calculated taking into account the fact that different approaches may have different number of branches (and that some genes may be relevant only in a sub-branch).
8. Application to mass-cytometry data:
  - 1) Please report the results and running time using the whole dataset. If a method fails to run in a reasonable time exclude it from the comparison.
  - 2) Custom parameters used (see point 2)
9. In the plot where multiple cells are displayed as circles, it may be worth to remove the black border to improve the perception of the density (for example Figure 2e or Figure 4e).

## References

1. van Dijk D, Nainys J, Sharma R, Kathail P, Carr AJ, Moon KR, Mazutis L, Wolf G, Krishnaswamy S, Pe'er D: MAGIC: A diffusion-based imputation method reveals gene-gene interactions in single-cell RNA-sequencing data. *BioRxiv*. 2017.

## Is the rationale for developing the new method (or application) clearly explained?

Yes

## Is the description of the method technically sound?

Yes

## Are sufficient details provided to allow replication of the method development and its use by others?

Yes

**If any results are presented, are all the source data underlying the results available to ensure full reproducibility?**

Yes

**Are the conclusions about the method and its performance adequately supported by the findings presented in the article?**

Partly

**Competing Interests:** No competing interests were disclosed.

**Referee Expertise:** Single cell (epi)genomics, genome editing (see more on [www.pinellolab.org](http://www.pinellolab.org))

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

Referee Report 06 April 2017

doi:[10.21956/wellcomeopenres.11959.r21016](https://doi.org/10.21956/wellcomeopenres.11959.r21016)



**Anthony Gitter** 

Department of Biostatistics and Medical Informatics, University of Wisconsin–Madison, Madison, WI, USA

This manuscript presents Mixtures of Factor Analysers (MFA), a hierarchical Bayesian model for studying the branching structure in single-cell RNA-seq and mass cytometry data. Single-cell RNA-seq can provide snapshots of cells progressing through dynamic biological processes, many of which exhibit branching structure in which the expression levels of one subset of cells diverges from the others. These types of dynamic behaviors are encountered not only in differentiation but also in stimulus response and other processes, which has sparked a need to computationally model the overall branching structure of the process, how cells progress through the process, and how gene expression levels of some genes differ along the branches. These and related inference problems are what MFA aims to solve.

MFA is a generative probabilistic model that uses factor analysis to model the expression properties of branches in a single-cell RNA-seq dataset. A prior is used to encourage similar factor loading gradients in each branch. An optional extension models the dropout phenomenon in which technical artifacts can cause non-zero mRNA abundances to be reported as zeros. MFA is compared with several popular existing algorithms that are not generative probabilistic models: Wishbone, Diffusion Pseudotime, and Monocle 2. These comparisons are conducted in a fair manner on simulated and real data. The assessment of the benchmarking is balanced, as is the overall conclusion that in most cases MFA is competitive with these existing approaches even if there is not evidence that it definitively outperforms them by some quantitative metric.

The balanced discussion is a strength of the manuscript overall. Figures 2E and 3C both show the scenarios in which MFA's performance degrades with respect to dropout levels or the fraction of transient genes. The authors conclude that in many practical analyses the extension for incorporating zero-inflation is not worth the added computational cost. They also present the limitations of their linear model and offer suggestions for improving the scalability of the inference and the linearity assumption.

The open source software is another asset and follows the best practices for scientific code. The code is

available in GitHub, and an archival version has been deposited in Zenodo. The Zenodo version's title states that it is the "Bioconductor-ready version", and providing the mfa R package through Bioconductor would indeed further enhance its utility.

Overall, the manuscript is easy to read, and the model is technically sound and well-motivated. I have only minor comments that may improve the accessibility to a broader audience and help clarify some points.

Minor comments:

\* The Methods section assumes that the reader is already familiar with factor analysis, as this technique is not explained. It would be helpful to introduce the approach and the meaning of  $c$  and  $k$  in this biological context.

\* There has been other related work on branching trajectories in single-cell datasets. A few examples include:

- SLICER (DOI:10.1186/s13059-016-0975-3) <sup>1</sup>
- TSCAN (DOI:10.1093/nar/gkw430) <sup>2</sup>
- Topslam (DOI:10.1101/057778) <sup>3</sup>
- Mpath (DOI:10.1038/ncomms11988) <sup>4</sup>

Very briefly discussing some of these methods and expanding the discussion of how GPfates (reference 16, DOI:10.1126/sciimmunol.aal2192)<sup>5</sup> relates to MFA would help readers understand MFA's advantages and disadvantages. I do not think it is necessary to benchmark against additional algorithms.

\* The parameter  $B$ , the number of branches, appears to be user-defined, but this is not explicitly stated in the text. It would help users to offer guidance on selecting this crucial parameter.

\* The sensitivity to the hyperparameter values is not assessed. It is not clear what model behavior was observed when modeling the mass cytometry dataset that led to the decision to use non-default values for  $\alpha_x$  and  $\beta_x$  and how users should make those decisions on new datasets.

\* I understand the mathematical invariance presented in Figure 1, but the biological argument is not intuitive. In the bifurcation 2  $\rightarrow$  1,3 states 1 and 3 have the same expression level, which would suggest that this single gene does not exhibit branching behavior. Rather, it switches from a high to low state in all cases.

\* The simulation with 300 bifurcating cells and 60 genes may have been too simple. Even the first principal component of the data recovers the true pseudotimes well. All of the methods perform extremely well, making it difficult to assess their relative performances.

\* I expected that the red cells along the lower curve in Figure 2B would be labeled as False in Figure 2C. Visually, the branch point appears to occur around 15 on the PC1 axis.

\* Running Monocle 2 on a smaller set of sub-sampled cells than the other methods could put it at a disadvantage. The scalability of MFA is not discussed or related to the runtimes of the other methods. Could MFA run on the full mass cytometry dataset in a reasonable amount of time or is sub-sampling required?

\* There are a few potential typos:

- Abstract: "apply or model" -> "apply our model"
- Abstract: "context practical" -> "context of practical"
- Supplement page 2: The identity matrix symbol in the line of Equation 1 for k is not correct
- Supplement page 4: Presumably the p(dropout in gene g) equation in the text should also have a 1/N term to match Equation 18
- Supplement page 7: "conduisive" -> "conductive"

## References

1. Welch JD, Hartemink AJ, Prins JF: SLICER: inferring branched, nonlinear cellular trajectories from single cell RNA-seq data. *Genome Biol.* 2016; **17** (1): 106 [PubMed Abstract](#) | [Publisher Full Text](#)
2. Ji Z, Ji H: TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic Acids Res.* 2016; **44** (13): e117 [PubMed Abstract](#) | [Publisher Full Text](#)
3. Zwiessele M, Lawrence ND: Topslam: Waddington Landscape Recovery for Single Cell Experiments. *bioRxiv.* 2017. [Publisher Full Text](#) | [Reference Source](#)
4. Chen J, Schlitzer A, Chakarov S, Ginhoux F, Poidinger M: Mpath maps multi-branching single-cell trajectories revealing progenitor cell progression during development. *Nat Commun.* 2016; **7**: 11988 [PubMed Abstract](#) | [Publisher Full Text](#)
5. Lönnberg T, Svensson V, James KR, Fernandez-Ruiz D, Sebina I, Montandon R, Soon MS, Fogg LG, Nair AS, Liligeto U, Stubbington MJ, Ly LH, Bagger FO, Zwiessele M, Lawrence ND, Souza-Fonseca-Guimaraes F, Bunn PT, Engwerda CR, Heath WR, Billker O, Stegle O, Haque A, Teichmann SA: Single-cell RNA-seq and computational analysis using temporal mixture modelling resolves Th1/Tfh fate bifurcation in malaria. *Sci Immunol.* 2017; **2** (9). [PubMed Abstract](#) | [Publisher Full Text](#)

**Competing Interests:** No competing interests were disclosed.

**I have read this submission. I believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.**

---