Ouija: Incorporating prior knowledge in single-cell trajectory learning using Bayesian nonlinear factor analysis

Kieran Campbell[1,2] and Christopher Yau[2,3]

[1]Department of Physiology, Anatomy and Genetics, University of Oxford, UK
[2]Wellcome Trust Centre for Human Genetics, University of Oxford, UK
[3]Department of Statistics, University of Oxford, UK

June 23, 2016

Abstract

Pseudotime estimation algorithms from single cell molecular profiling allows the recovery of temporal information from otherwise static profiles of individual cells. This pseudotemporal information can be used to characterise transient events in temporally evolving biological systems. Conventional algorithms typically employ an unsupervised approach that do not model any explicit gene behaviours making them hard to apply in the context of strong prior knowledge. Our approach Ouija takes an alternate approach to pseudotime by explicitly focusing on switch-like marker genes that are ordinarily used to confirm the accuracy of unsupervised pseudotime algorithms. Instead of using the marker genes for confirmation, we derive pseudotimes directly from the marker genes. We show that in many instances a small panel of marker genes can recover pseudotimes that are consistent with those obtained using the whole transcriptome. Ouija therefore provides a powerful complimentary approach to existing approaches to pseudotime estimation.

1 Introduction

The advent of high-throughput single-cell technologies has revolutionised single-cell biology by allowing dense molecular profiling for studies involving 100-10,000s of cells [1–6] . The increased availability of single cell data has driven the development of novel analytical methods specifically tailored to single cell properties [7, 8]. Due to the difficulties of conducting genuine time series experiments, one important important development has been the emergence of computational techniques, known as 'pseudotime ordering' algorithms, to extract temporal information from snapshot molecular profiles of individual cells. These algorithms exploit studies in which the cells captured are behaving asynchronously and therefore each is at a different stage of some underlying temporal biological process, such as cell differentiation. In sufficient numbers, it is possible to infer an ordering of the cellular profiles that correlates with actual temporal dynamics and these approaches have promoted insights into a number of time-evolving biological systems [9–18].

A predominant feature of current pseudotime algorithms is that they emphasise an "unsupervised" approach where pseudotimes are learned using no specific prior knowledge of gene behaviour. Typically, the high-dimensional molecular profiles for each cell are projected on to a reduced dimensional space by using a (non)linear transformation of all the underlying molecular properties. In this reduced dimensional space, it is hoped that any temporal variation is sufficiently strong to cause the cells to align against a trajectory along which pseudotime can be measured. This approach is therefore subject to a number of analysis choices, e.g. the choice of dimensionality reduction technique, the trajectory fitting algorithm, etc., that could lead to considerable variation in the pseudotime estimates obtained. In order to verify that any specific set of pseudotime estimates are biologically plausible, it is typical for investigators to retrospectively examine specific marker genes or proteins to confirm that the predicted (pseudo)temporal behaviour reflects *a priori* beliefs. An iterative "semi-supervised" process maybe therefore be required to concentrate pseudotime algorithms on behaviours that are both consistent with the measured data and compliant with a limited amount of known gene behaviour.

In this paper we present an orthogonal approach implemented in a latent variable model statistical framework called Ouija that can integrate prior expectations of gene behaviour along trajectories using Bayesian nonlinear factor analysis. Our approach therefore using known or putative marker genes directly for pseudotime estimation rather than as a device for retrospectively validating pseudotime estimates. In particular, our model focuses on switch-like expression behaviour and assumes that the marker gene expression follows a noisy switch pattern that corresponds to up- or down-regulation over time. Crucially, we explicitly model when a gene turns on or off as well as how quickly this behaviour occurs. We can then place Bayesian priors on this behaviour that allows us to learn temporal gene behaviours that are consistent with existing biological knowledge.

Our hypothesis in developing Ouija was that using only small panels of marker gene expression profiles, with a vague knowledge of their temporal behaviour, could learn pseudotimes equivalent to method using whole transcriptome methods. To test this we applied Ouija to three previously published single-cell RNA-seq datasets of differentiating cells. In the following, we demonstrate how Ouija can recover pseudotemporal orderings of cells using as low as five marker genes, providing estimates consistent with whole-transcriptome algorithms such as Monocle [12] and Waterfall [16]. We go on to show how Ouija outperforms various pseudotemporal ordering algorithms for carefully constructed synthetic data for small panels of genes. Finally, we provide a use-case for Ouija when trying to recover the trajectory of a known pathway of genes in the presence of a secondary, unknown confounding pathway. Ouija is available as a `R` package at `www.github.com/kieranrcampbell/ouija`.

2 Methods

2.1 Overview

We give a high-level overview of our pseudotime inference framework here and provide more technical details in the following sub-sections. The aim of pseudotime ordering is to associate a p-dimensional expression measurement (the data) to a latent unobserved pseudotime. Mathematically we can express this as the following:

$$\underbrace{y_c}_{\text{Expression}} = \underbrace{f}_{\text{Mapping}}\Big(\underbrace{t_c}_{\text{Pseudotime}}\Big) + \underbrace{e_c}_{\text{Noise}} \tag{1}$$

where the function f maps the one-dimensional pseudotime to the p-dimensional observation space in which the data lies. The challenge lies in the fact that *both* the mapping function f and the pseudotimes are *unknown*.

Our objective here is to use parametric forms for the mapping function f that will enable relatively fast computations whilst characterising certain gene expression temporal behaviours. Our premise is that prior knowledge might exist for a set of marker genes whose temporal behaviour is known to be approximately switch-like and therefore can be used to infer pseudotime orderings. We use sigmoid mapping functions that can capture "switch-like" behaviour over time as shown in Figure 1B and are parameterised by key quantities for which estimates (and associated uncertainty) might be useful: the activation strength and the activation time. These correspond to measures of how rapidly the gene expression level changes and when the (in)activation of the gene occurs.

The approach we adopt is therefore a form of latent variable model implemented ass *non-linear parametric factor analysis* where the factors correspond to the pseudo-times and the factor loadings correspond to the parameters of the sigmoidal function which provides the non-linearity. In addition, we model dropouts and a strict empirically motivated mean-variance relationship (see Supplementary Information) which is required to provide constraints on the latent variable model since nothing on the right hand side of Equation 1 is actually measured or observed.

2.2 Statistical Model

2.2.1 Model Specification

We index C cells by $c \in 1, \ldots, C$ and G genes by $g \in 1, \ldots, G$. Let $y_{cg} = [\mathbf{Y}]_{cg}$ denote the (log-transformed) observed cell-by-gene expression matrix. Let $\mathbf{t}$ be an N-length pseudotime vector (one for each cell).

Our statistical model can be specified as a Bayesian hierarchical model where the likelihood model is given by a bimodal distribution formed from a mixture of zero-component (dropout) and an non-zero expressing cell population:

$$y_{cg} \sim \begin{cases} \theta_{cg} + (1-\theta_{cg})\text{Student}(\mu_{cg}, \sigma^2_{cg}) & \text{if } y_{cg} = 0 \\ (1-\theta_{cg})\text{Student}(\mu_{cg}, \sigma^2_{cg}) & \text{if } y_{cg} > 0 \end{cases}, \tag{2}$$

$$\theta_{cg} \sim \text{Bernoulli}(\text{logit}^{-1}(\beta_0 + \beta_1 \mu_{cg})), \tag{3}$$

$$\boldsymbol{\beta} \sim \text{Normal}(0, 0.1). \tag{4}$$

The relationship between dropout rate and expression level is expressed as a logistic regression model [19]. Furthermore, we impose a mean-variance relationship of the form:

$$\mu_{cg} = \mu^{(0)}_g f(t_c, k_g, t^{(0)}), \tag{5}$$

$$t_c \sim \text{Normal}(0.5, 1), \tag{6}$$

$$\sigma^2_{cg} = (1+\phi)\mu_{cg} + \epsilon, \tag{7}$$

$$\phi \sim \text{Gamma}(\alpha_\phi, \beta_\phi) \tag{8}$$

where ϕ is a dispersion parameter. This model is motivated by empirical observations of marker gene behaviour (see Supplementary Information).

We define the sigmoid function as

$$f(t_c; k_g, t^{(0)}_g) = \frac{2}{1 + \exp\left(-k_g(t_c - t^{(0)}_g)\right)}, \tag{9}$$

where k_g and $t^{(0)}_g$ denote the activation strength and activation time parameters for each gene

and $\mu_g^{(0)}$ the average peak expression with default priors

$$\mu_g^{(0)} \sim \text{Gamma}(\delta/2, 1/2), \tag{10}$$
$$k_g \sim \text{Normal}(\mu_g^{(k)}, 1/\tau_g^{(k)}), \tag{11}$$
$$t_g^{(0)} \sim \text{Normal}(\mu_g^{(t)}, 1/\tau_g^{(t)}), \tag{12}$$

If available, user-supplied prior beliefs can be encoded in these priors by specifying the parameters $\mu_g^{(k)}, \tau_g^{(k)}, \mu_g^{(t)}, \tau_g^{(t)}$. Otherwise, inference can be performed using uninformative hyperpriors on these parameters. Specifying $\mu_g^{(k)}$ encodes a prior belief in the strength and direction of the activation of gene g along the trajectory with $\tau_g^{(k)}$ (inversely-) representing the strength of this belief. Similarly, specifying $\mu_g^{(t)}$ encodes a prior belief of where in the trajectory gene g exhibits behaviour (either turning on or off) with $\tau_g^{(t)}$ encoding the strength of this belief.

2.2.2 Inference

We perform posterior inference using Markov Chain Monte Carlo (MCMC) stochastic simulation algorithms, specifically the No U-Turn Hamiltonian Monte Carlo approach [20] implemented in the STAN probabilistic programming language [21] which we use to implement our model. The parameter $\epsilon = 0.01$ is used to avoid numerical issues in MCMC computation.

2.3 Principal Component Analysis assumes linear gene activations

If the mapping functions f are restricted to a linear form then the model becomes:

$$\begin{aligned} k_g &\sim \text{Normal}(\mu_g^{(k)}, 1/\tau^{(k)}) \\ y_{cg} &\sim \text{Normal}(k_g t_c, 1/\tau_g) \end{aligned} \tag{13}$$

Note that the data can always be centred making it redundant to model an intercept. If we set $\mu_g^{(k)} = 0$ and $\tau^{(k)} = 1$ then this model reduces to Factor Analysis. In other words, performing factor analysis on single-cell RNA-seq data is entirely equivalent to finding a trajectory where gene expression is linear over time with no prior expectations on how the genes behave. If we model a common precision across all genes so $\tau_g = \tau$ then then model reduces further to (probabilistic) principal components analysis [22], providing an explicit interpretation for the results of principal component analysis on single-cell data.

3 Results

3.1 Ouija recovers whole-transcriptome pseudotimes using small gene panels

Our hypothesis in developing Ouija was that using only small panels of marker genes with a vague knowledge of the behaviour of those genes we can learn pseudotimes equivalent to whole-transcriptome methods. To test this we applied Ouija to four previously published single-cell RNA-seq datasets of differentiating cells.

The first dataset on which we tested Ouija was differentiating human myoblasts from [12]. Here cells begin as proliferating cells before a serum switch induces differentiation as myoblasts, with contaminating interstitial mesenchymal cells removed prior to the analysis (full details in supplementary methods). The original publication identified *CDK1*, *ID1*, *MYOG*, *MEF2C* and *MYH3* as key regulators of muscle differentiation. *CDK1* and *ID1* are expected to decrease

in expression as the cells leave their proliferating phase while *MYOG*, *MEF2C* and *MYH3* are expected to switch on as myogenesis begins.

To encode such information in the Ouija framework we set $\boldsymbol{\mu}^{(k)} = (-10, -10, 10, 10, 10)$ and left priors on $t^{(0)}$ centred around the midpoint of the trajectory. The results of the trajectory fitting can be see in Figure 2A. Using only these five marker genes Ouija recovers the published trajectory with a Pearson correlation of 0.79 (note that the published trajectory does not necessarily correspond to the *true* trajectory). The behaviour of the marker genes at the MAP pseudotime estimate is naturally consistent with our prior expectations, with clear evidence for strong pseudotemporal regulation of *CDK1*, *MYOG* and *MYH3*. Our method also provides evidence that the "switch-like" inactivation of *ID1* mentioned in the original paper occurs at around the half way point in the pseudotemporal trajectory, which is not evident from examining the gene expression plot alone.

The second dataset on which we tested Ouija was on mouse quiescent neural stem cells (qNSCs) [16]. qNSCs continuously regenerate new neurons and in this study the authors examined the cells at the very beginning of this process as they progress from qNSCs to activated neural stem cells (aNSCs) and on to early intermediate progenitor cells (eIPCs). We used six pseudotemporal marker genes identified in the paper (*Sox11*, *Eomes*, *Stmn1*, *Apoe*, *Aldoc* and *Gfap*) to fit the pseudotemporal trajectory. *Sox11* and *Eomes* were known markers of qNSCs while *Apoe* and *Gfap* were known markers of eIPCs. The genes *Stmn1* and *Aldoc* were newly discovered markers of qNSCs and eIPCs respectively.

We encoded the expression patterns as prior beliefs similarly to above using Ouija and fitted the trajectories, the results of which can be seen in Figure 2B. The inferred pseudotimes have a Pearson correlation of 0.72 with the published pseudotimes, which due to the nature of the Waterfall algorithm is very similar to principal component 1 on a large gene set. The behaviour of the genes at the Ouija MAP estimate is consistent with our prior knowledge, but with clear evidence that the marker gene *Gfap* is not ubiquitously expressed among qNSCs.

The third dataset to which we applied Ouija was a single cell expression experiment on differentiating planarian neural stem cells [23]. The authors proposed three stem-cell marker genes (*piwi-1*, *vasa-1*, *hdac-1*) expected to be down regulated along the trajectory with three neural marker genes (*pc-2*, *chat*, *ascl-1*) expected to be up regulated. We encoded this prior expectation in Ouija and fitted the trajectory, as shown in Figure 2C. Our results uncovered clear evidence for down regulation of the three stem-cell marker genes but only show consistent evidence for early up regulation of *chat* and not of *pc-2* or *ascl-1*.

3.1.1 Clustering cell types based on pseudotime continuity

The fourth dataset to what we applied Ouija tracked the differentiation of embryonic precursor cells into haematopoietic stem cells (HSCs) [24]. The cells begin as haemogenic endothelial cells (ECs) before travelling down a differentiation trajectory that successively transforms them into pre-HSC and finally HSC cells. The authors identified six marker genes that would be down-regulated along the differentiation trajectory, with early down-regulation of *Nrp2* and *Nr2f2* as the cells transform from ECs into pre-HSCs, and late down-regulation of *Nrp1*, *Hey1*, *Efnb2* and *Ephb4* as the cells emergy from pre-HSCs to become HSCs.

We encoded this prior expectation in the subset of genes and performed pseudotime inference using Ouija, the results of which can be seen in Figure 2D. As expected *Nrp2* and *Nr2f2* exhibit early down-regulation with *Nrp1*, *Hey1*, *Efnb2* and *Ephb4* all exhibiting late down-regulation.

The cells in this experiment correspond to distinct types as they progress along the differentiation trajectory, namely EC cells, T1 cells ($CDK45^{-}$ pre-HSCs), T2 cells ($CDK45^{+}$ pre-HSCs) and HSC cells at the E12 and E14 developmental stages. We were interested to discover if we could recapitulate these cell types using the six marker genes mentioned alone.

The pseudotime orderings inferred by Ouija are probabilistic so we can naturally quantify the number of times a given cell appears before another in the trajectory. Consequently, there will be a large uncertainty in the orderings of cells within each cell type but there will be a consistent ordering between cell types, *if distinct cell types exist in the dataset*. Otherwise, we expect the uncertainty in the ordering to be consistent along the trajectory. As a result we have a method to distinguish between discrete and continuous cell types based on expression, and if the discrete cell types exist we can uncover them by examining the ordering consistency.

Specifically, for C cells we form a $C \times C$ matrix M, where M_{ij} = proportion of times cell i is ordered before cell j in the posterior MCMC traces[1]. To aid visualisation the rows and columns of M are reordered to reflect the MAP pseudotime ordering of cells. To cluster the cells in this setting we perform univariate Gaussian mixture model clustering on $\mathbf{x}$, the first principal component[2] of M. As a result, each discovered cell type i has an associated 'centre' in pseudotime μ_i and variance σ_i^2, meaning we can classify a cell at hypothetical pseudotime t_c by $\text{argmax}_i \text{ Normal}(t_c|\mu_i, \sigma_i^2)$. For visualisation we can also form a decision boundary d_{ij} between adjacent cell types i and j as the solution to $\text{Normal}(d_{ij}|\mu_i, \sigma_i^2) = \text{Normal}(d_{ij}|\mu_j, \sigma_j^2)$.

We applied this methodology to the HSC trajectories after probabilistic pseudotime inference using Ouija. A heatmap of the consistency matrix M can be seen in Figure 3A. The mixture modelling was performed using the `R` package `mclust` [25], which automatically chose four clusters, overlain on Figure 3A. The regions of pseudotime for which each of the clusters is the maximum likelihood cluster are plotted as background colouring along with gene expression in Figure 3B. The originally measured cell type is plotted as a function of pseudotime in Figure 3C, and the confusing matrix between originally measured cell type and Ouija cell type is shown in Figure 3D.

3.2 Ouija outperforms alternative pseudotime algorithms on synthetic data

To compare the performance of Ouija to other pseudotemporal ordering algorithms on small gene sets we created synthetic gene expression data with known pseudotimes then re-inferred the pseudotimes using the algorithms. Briefly, pseudotimes were generated on a uniform distribution on $[0, 1)$ and k_g, $\mu_g^{(0)}$ and $t_g^{(0)}$ generated for each gene. Gene expression representing $\log_2(\text{TPM} + 1)$ was then generated using a Gaussian likelihood censored from below at zero. Care was taken to ensure the mean-variance and dropout relations were comparable to real single-cell RNA-seq data. Full details can be found in supplementary methods.

To benchmark the algorithms we generated data for $C = 100$ cells for $G \in \{6, 9, 12, 15\}$ genes, with forty replicates of each dataset for varying values of k, t_0 and μ_0. We benchmarked Ouija against principal component 1 (PC1), Diffusion Pseudotime (DPT) [17] and Monocle 2 [12]. Ouija allows us to specify priors on k and t_0 via $k_g \sim \text{Normal}(\mu_g^{(k)}, 1/\tau_g^{(k)})$ and $t_g^{(0)} \sim \text{Normal}(\mu_g^{(t)}, 1/\tau_g^{(t)})$. We benchmarked Ouija in three situations. Firstly, with *random priors* on k where $\mu_g^{(k)}$ is set to arbitrary values and $\mu_g^{(t)}$ is set to 0.5 (essentially non-informative) for all g. Secondly, *directional priors* where $\mu_g^{(k)}$ is set to a constant multiplied by the sign of the true k_g and $\mu_g^{(t)} = 0.5$. This situation is most likely to correspond to the real-life use case where the researcher knows the direction of change of the genes along the trajectory along with a guess of the magnitude of that change. Finally, we considered priors with *true values* where $\mu_g^{(k)}$ and $\mu_g^{(t)}$ are set to the true values used in the synthetic data generation. In general it is highly unrealistic for researchers to know the true parameter values but we include this case for comparison to the more realistic scenario of directional priors. In all situations the precision

[1] Note that M_{ii} is undefined and $M_{ij} = 1 - M_{ji}$

[2] $\mathbf{x}$ is a C-length vector where each entry can be thought of as a *consistency score* of that cell to its neighbours

parameters are set to 1.

The results can be seen in Figure 4. Performance was assessed using Pearson correlation to the true pseudotimes as well as Kendall's tau, a measure of concordance of the ordering of two sets. For both metrics three trends are obvious. Firstly, as the number of genes in the marker panel increases, the accuracy of the pseudotemporal inference also increases. Secondly, PCA, DPT, Monocle 2 and Ouija with random priors are all largely comparable in their results, with Ouija with both directional priors and true values performing significantly better. Finally, the performance of Ouija with directional priors and the true values are largely comparable, implying exact knowledge of the biological parameters isn't necessary for the inclusion of prior information to be useful.

3.3 Ouija discovers trajectories in the presence of confounding pathways

So far all analyses (on both real and synthetic data) have involved genes that are all known to be within the same pathway and so they share a common pseudotime. A more realistic scenario is when we are sure of the behaviour of a small set of genes but have no prior knowledge of some others that may or may not belong to the same pathway. This is akin to the situation using whole-transcriptome methods where the genes used include those involved in the process of interest such as differentiation as well as those involved in nuisance processes such as cell cycle or apoptosis. However, it is far easier to know *a priori* which genes are involved in the process as opposed to those which are not.

To test this in an extreme case we generated synthetic data where half the genes belong to one pseudotime t_1 and half belong to a second t_2 (supplementary methods). This represents a central process of interest (t_1) governed by genes $\mathcal{G}^*$ that we wish to infer from the gene expression matrix along with a confounding or nuisance process (t_2) governed by genes $\tilde{\mathcal{G}}$ that we wish to ignore. To test Ouija we integrated our prior knowledge for $\mathcal{G}^*$ and left noninformative priors on $\tilde{\mathcal{G}}$, and compared inference to PCA, Monocle 2 and Diffusion Pseudotime (DPT), where no such prior knowledge can be encoded.

The results can be seen in 5 for 40 replications of the synthetic dataset (supplementary methods). It can be seen that PCA, DPT and Monocle 2 perform comparably with median absolute Pearson correlations of 0.31, 0.33 and 0.39 respectively. In comparison, Ouija achieves a median absolute Pearson correlation of 0.86 which is similar to the case in which no confounding factors are present at all.

While such a use-case is artificial in the extreme it serves as proof of concept that by integrating relevant prior information Ouija can recover pseudotemporal orderings in the presence of secondary, confounding trajectories. Furthermore, this is not meant to serve as criticism of either DPT or Monocle 2 which are not designed for small gene panels with confounding factors but instead transcriptome-wide orderings where the dominant source of variation is the process of interest.

4 Conclusions

We have developed a novel approach for pseudotime estimation based on modelling switching expression behaviour over time for marker genes. Our strategy provides an orthogonal and complimentary approach to unsupervised, whole transcriptome methods that do not explicitly model any gene-specific behaviours and do not readily permit the inclusion of prior knowledge.

We demonstrate that the selection of a few marker genes allows comparable pseudotime estimates to whole transcriptome methods on real single cell data sets. Furthermore, using a

parametric gene behaviour model and full Bayesian inference we are able to recover posterior uncertainty information about key parameters, such as the gene activation time, that allows us to explicitly determine a potential ordering of gene (de)activation events over (pseudo)time. The posterior ordering uncertainty can also be used to identify homogeneous phases of transcriptional activity that might correspond to transient, but discrete, cell states.

Although our focus was on switching expression behaviour, alternative parametric functions could be used to capture other gene behaviours. However, it is critical to recognise that in a latent variable modelling framework such as this, prior information has a strong influence over the final outcome. Therefore any constraints should match *a priori* knowledge of the marker genes under investigation. Otherwise the pseudotime estimation could forcibly impose a pre-specified temporal form on the data. Furthermore, whilst we do not explicitly address branching processes in this work, our framework provides a natural and simple extension to allow for multiple lineages and cell fates using a sparse mixture of factor analyzers in which each lineage is denoted by a separate mixture component and the factors loadings are shrunk to common values to denote shared branches. This will be developed in future work.

In summary, Ouija provides a novel contribution to the increasing plethora of pseudotime estimation methods available for single cell gene expression data.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

K.C. and C.Y. conceived the study. K.C. developed software and performed computer simulations. K.C. and C.Y. wrote the manuscript.

Acknowledgements

K.C. is supported by a UK Medical Research Council funded doctoral studentship. C.Y. is supported by a UK Medical Research Council New Investigator Research Grant (Ref. No. MR/L001411/1), the Wellcome Trust Core Award Grant Number 090532/Z/09/Z, the John Fell Oxford University Press (OUP) Research Fund and the Li Ka Shing Foundation via a Oxford-Stanford Big Data in Human Health Seed Grant.

References

1. Kalisky, T. & Quake, S. R. Single-cell genomics. *Nature methods* **8,** 311–314 (2011).
2. Shapiro, E., Biezuner, T. & Linnarsson, S. Single-cell sequencing-based technologies will revolutionize whole-organism science. *Nature reviews. Genetics* **14,** 618–30. ISSN: 1471-0064 (Sept. 2013).
3. Macaulay, I. C. & Voet, T. Single cell genomics: advances and future perspectives. *PLoS genetics* **10,** e1004126. ISSN: 1553-7404 (Jan. 2014).
4. Wills, Q. F. & Mead, A. J. Application of Single Cell Genomics in Cancer: Promise and Challenges. *Human molecular genetics,* ddv235 (2015).
5. Linnarsson, S. & Teichmann, S. A. Single-cell genomics: coming of age. *Genome biology* **17,** 1 (2016).

6. Liu, S. & Trapnell, C. Single-cell transcriptome sequencing: recent advances and remaining challenges. *F1000Research* **5** (2016).

7. Stegle, O., Teichmann, S. a. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics* **16,** 133–145. ISSN: 1471-0056 (Jan. 2015).

8. Trapnell, C. Defining cell types and states with single-cell genomics. *Genome Res* **25,** 1491–8 (2015).

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

10. Bendall, S. C. *et al.* Single-cell trajectory detection uncovers progression and regulatory coordination in human B cell development. *Cell* **157,** 714–25. ISSN: 1097-4172 (Apr. 2014).

11. Marco, E. *et al.* Bifurcation analysis of single-cell gene expression data reveals epigenetic landscape. *Proceedings of the National Academy of Sciences of the United States of America* **111,** E5643–50. ISSN: 1091-6490 (Dec. 2014).

12. Trapnell, C. *et al.* The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology* **32,** 381–6. ISSN: 1546-1696 (Apr. 2014).

13. Moignard, V. *et al.* Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology* **33.** ISSN: 1087-0156. doi:10.1038/nbt.3154. <http://www.nature.com/doifinder/10.1038/nbt.3154> (Feb. 2015).

14. Reid, J. E. & Wernisch, L. Pseudotime estimation: deconfounding single cell time series. *bioRxiv,* 019588 (2015).

15. Hanchate, N. K. *et al.* Single-cell transcriptomics reveals receptor transformations during olfactory neurogenesis. *Science,* science.aad2456–. ISSN: 0036-8075 (Nov. 2015).

16. Shin, J. *et al.* Single-Cell RNA-Seq with Waterfall Reveals Molecular Cascades underlying Adult Neurogenesis. English. *Cell Stem Cell* **17,** 360–372. ISSN: 19345909 (Aug. 2015).

17. Haghverdi, L., Buettner, M., Wolf, F. A., Buettner, F. & Theis, F. J. Diffusion pseudotime robustly reconstructs lineage branching. *bioRxiv,* 041384 (2016).

18. Ji, Z. & Ji, H. TSCAN: Pseudo-time reconstruction and evaluation in single-cell RNA-seq analysis. *Nucleic acids research,* gkw430 (2016).

19. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nature methods* **11,** 740–2. ISSN: 1548-7105 (July 2014).

20. Homan, M. D. & Gelman, A. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *The Journal of Machine Learning Research* **15,** 1593–1623 (2014).

21. Carpenter, B. *et al.* Stan: a probabilistic programming language. *Journal of Statistical Software* (2015).

22. Tipping, M. E. & Bishop, C. M. Probabilistic principal component analysis. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **61,** 611–622 (1999).

23. Molinaro, A. M. & Pearson, B. J. In silico lineage tracing through single cell transcriptomics identifies a neural stem cell population in planarians. *Genome biology* **17,** 1 (2016).

24. Zhou, F. *et al.* Tracing haematopoietic stem cell formation at single-cell resolution. *Nature* **533,** 487–492 (2016).

25. Fraley, C. & Raftery, A. E. MCLUST: Software for model-based cluster analysis. *Journal of Classification* **16,** 297–306 (1999).

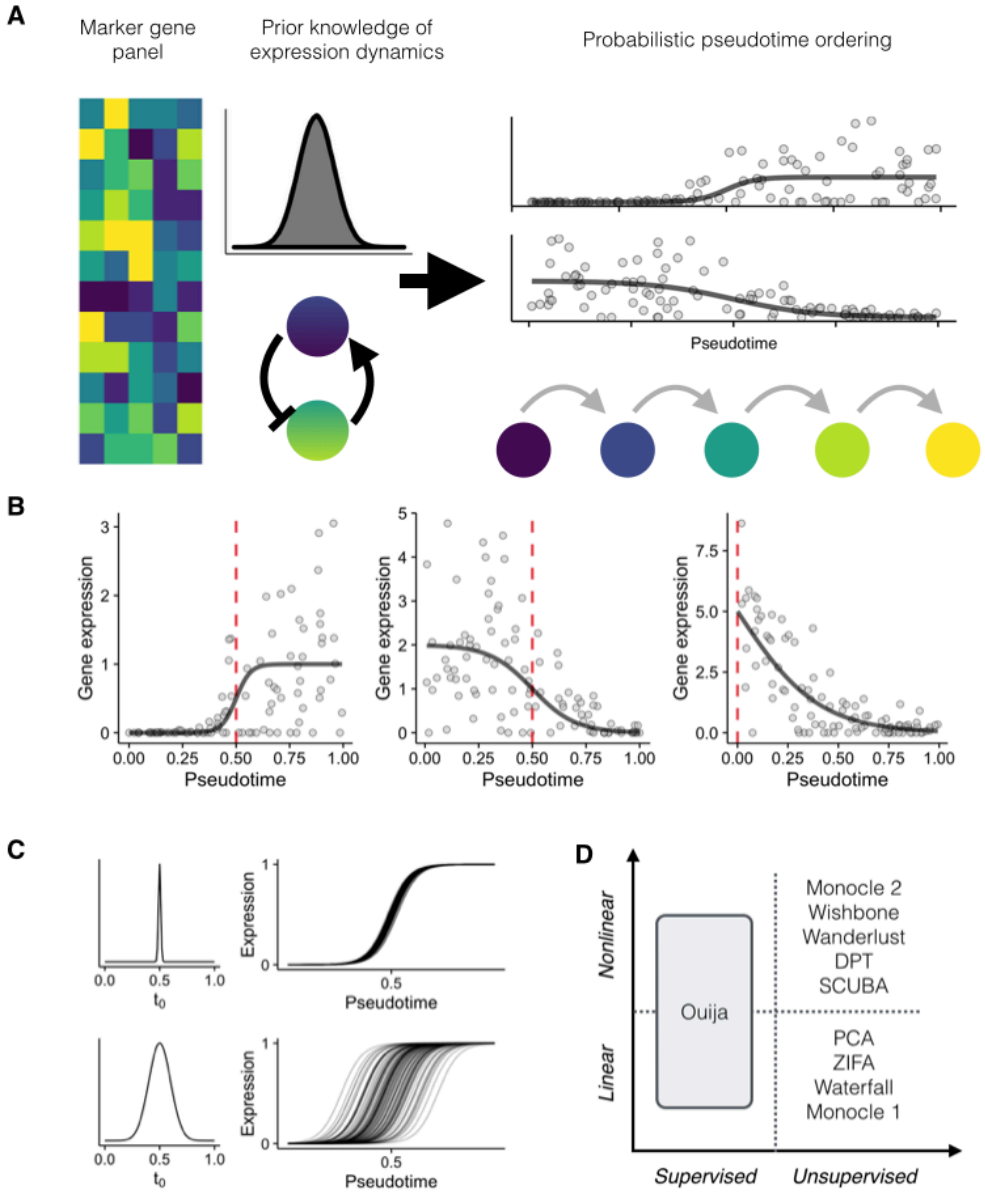


Figure 1: Overview of Ouija. A A small panel of genes is chosen with *a priori* knowledge about their expression dynamics, either as cell-type specific marker genes or from pathway interaction databases such as KEGG. The pseudotimes are then inferred using a Bayesian hierarchical non-linear factor analysis model. Differential expression and clustering of additional genes across pseudotime can then be inferred using standard methods such as those described in [12]. **B** Examples of sigmoidal gene expression for three different sets of parameters. Points represent simulated gene expression, with the solid black line denoting the sigmoid curve and the red dashed line denoting the activation time. The sigmoid function is parameterised by (i) μ_0 - the average peak expression level, (ii) k - the activation strength and (iii) t_0 - the activation time. *Left* Fast 'switch-on' behaviour with parameters $k = 30$, $\mu_0 = 1$, $t_0 = 0.5$, *Centre* Slower 'switch-off' behaviour with parameters $k = -10$, $\mu_0 = 2$, $t_0 = 0.5$ and *Right* Decaying behaviour with parameters $k = -5$, $\mu_0 = 10$, $t_0 = 0$. **C** The effect of prior expectations on the 'activation time' t_0 parameter. A highly peaked prior (top) corresponds to confident knowledge of where in the trajectory a gene turns on or off, while a more diffuse prior (bottom) indicates more uncertainty as to where in the trajectory the gene behaviour occurs. **D** Comparison of Ouija to alternative pseudotime inference algorithms. Ouija is the only algorithm that explicitly allows incorporation of prior knowledge of gene behaviour.

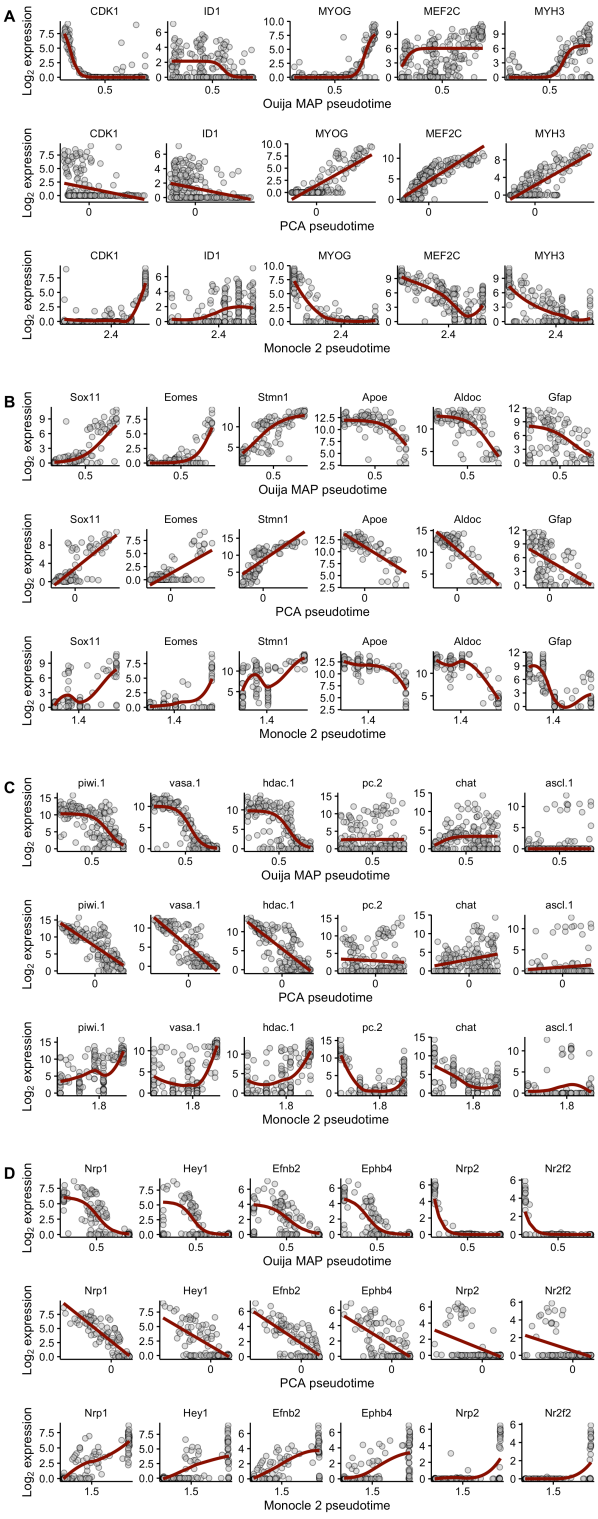
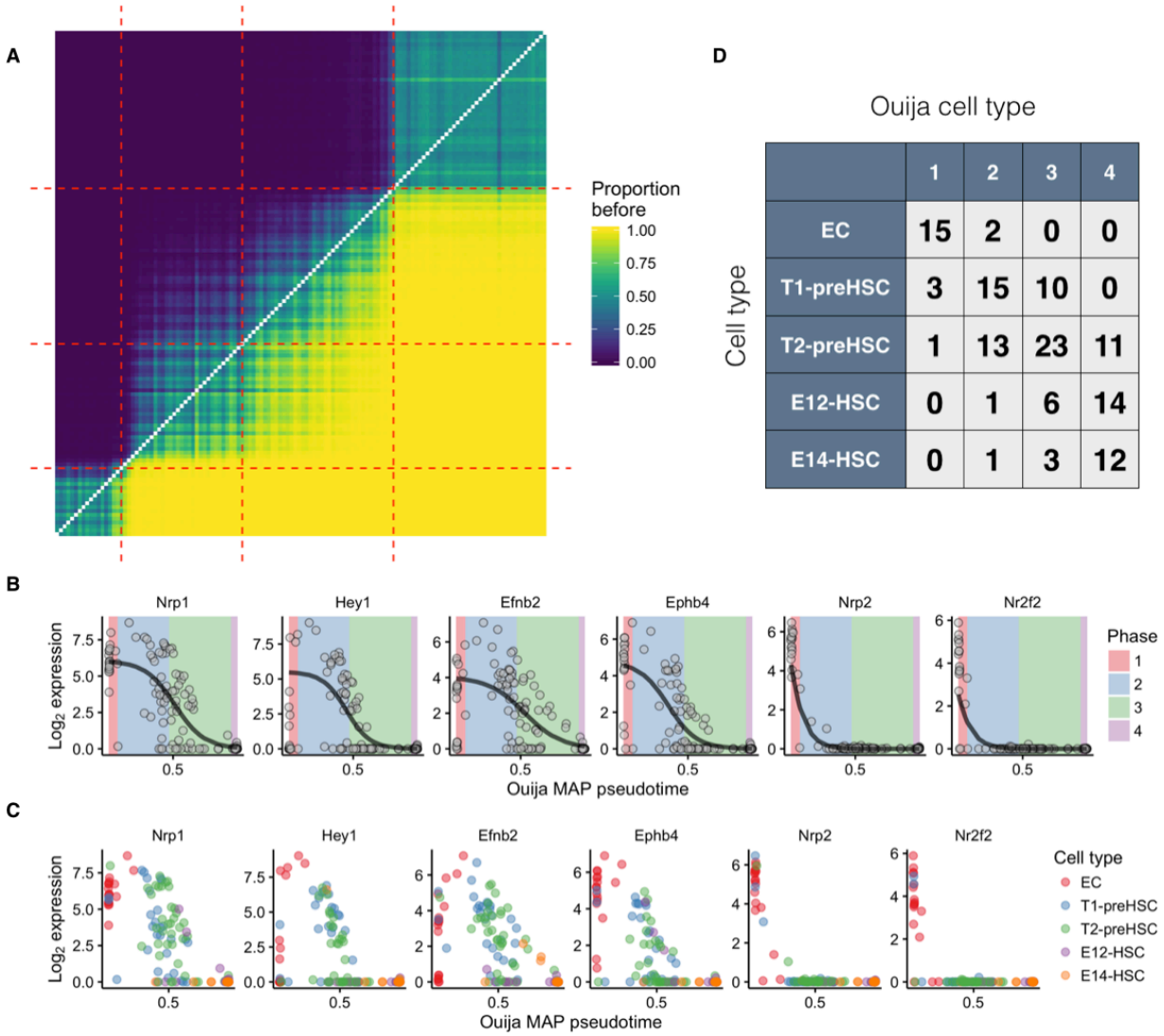


Figure 2: Comparison of pseudotimes inferred using Ouija to the published values. **A** For the dataset described in [12] of differentiating myoblasts. Using only five pseudotemporal marker genes described in the publication (*CDK1*, *ID1*, *MYOG*, *MEF2C*, *MYH3*) Ouija infers pseudotimes with a Pearson correlation of 0.79 compared to the transcriptome-wide method used in the original publication (`Monocle`). **B** For the dataset described in [16] of neurogenesis in mice. Using six marker genes described in the publication (*Sox11*, *Eomes*, *Stmn1*, *Apoe*, *Aldoc*, *Gfap*), Ouija infers pseudotimes with a Pearson correlation of 0.72 compared to the transcriptome-wide method used in the original publication (`Waterfall`).



	1	2	3	4
EC	15	2	0	0
T1-preHSC	3	15	10	0
T2-preHSC	1	13	23	11
E12-HSC	0	1	6	14
E14-HSC	0	1	3	12

Figure 3: Pseudotime ordering and cell type identification of haematopoeietic stem cell differentiation A Consistency matrix of pseudotime ordering. Entry in the i^{th} row and j^{th} column is the proportion of times cell i was ordered before cell j in the MCMC posterior traces. Gaussian mixture modelling on the first principal component of the matrix identified four clusters as shown by the red-dashed lines. **B** HSC gene expression as a function of pseudotime for six marker genes. Background colour denotes the maximum likelihood estimate for the Ouija inferred cell type in that region of pseudotime. **C** HSC gene expression as a function of pseudotime for six marker genes with cells coloured by known cell type. **D** Confusion matrix for cell types identified in original study (rows) and Ouija inferred (columns). Ouija inferred cluster 1 corresponds to EC cells, clusters 2 & 3 correspond to pre-HSC cells while cluster 4 corresponds to HSC cells.

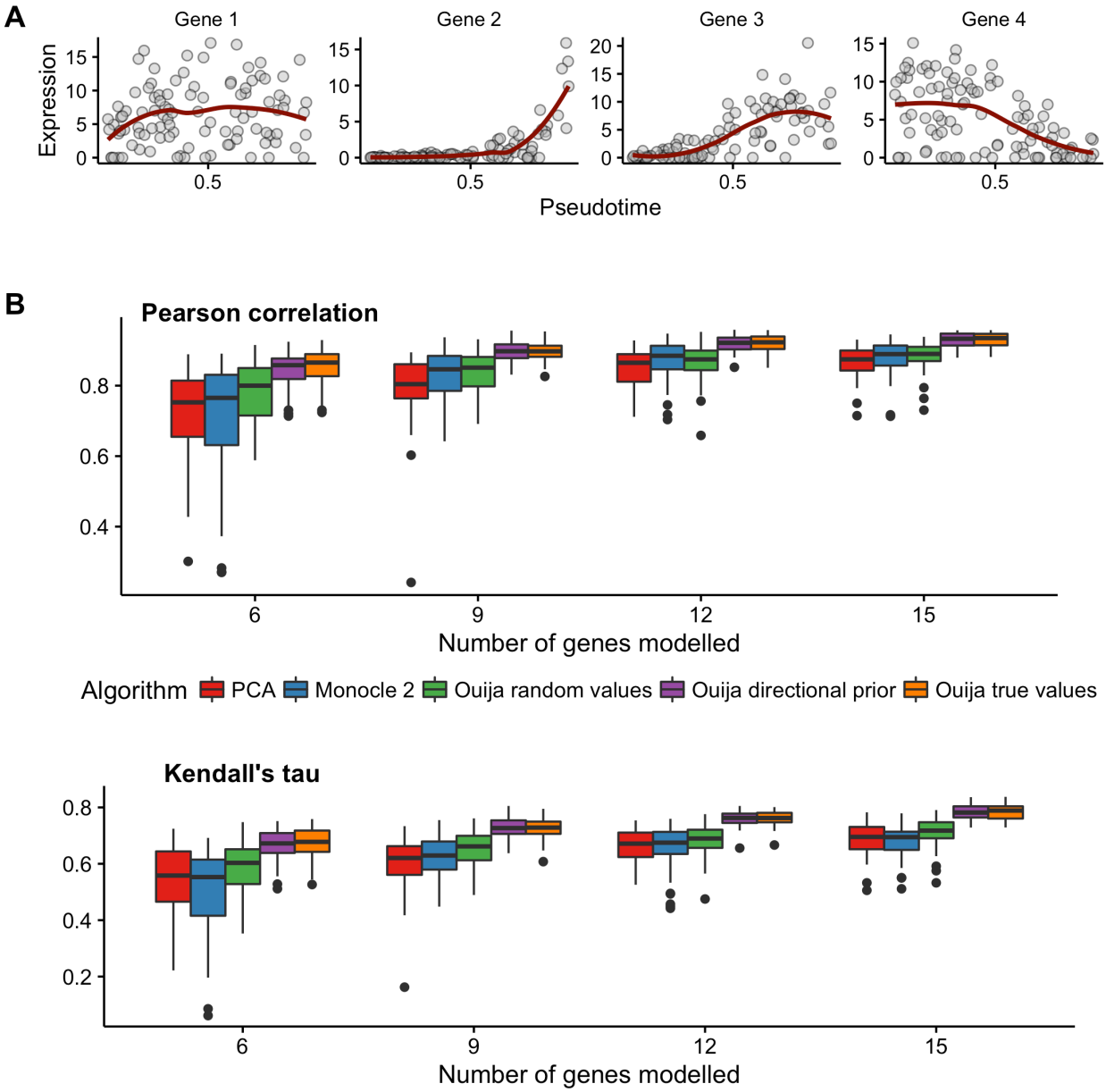


Figure 4: Comparison with alternative pseudotime algorithms using synthetic data. **A** Example plots of synthetic gene expression across pseudotime for four genes. The algorithm for generating synthetic pseudotemporally regulated gene expression data may be found in supplementary methods. **B** We compared Ouija against three alternative pseudotemporal methods (principal component 1, Monocle 2 and DPT). We generated synthetic data as described in supplementary methods for gene expression matrices of different sizes ($G = 6, 9, 12, 15$) with forty replicates in each and attempted to re-infer the pseudotimes. We used both Pearson correlation and Kendall's tau as measures of accuracy compared to the true values. Results show that PC1, Monocle 2, DPT and Ouija with non-informative priors perform comparably. Ouija with priors given by directional and true priors are significantly more accurate in all cases.

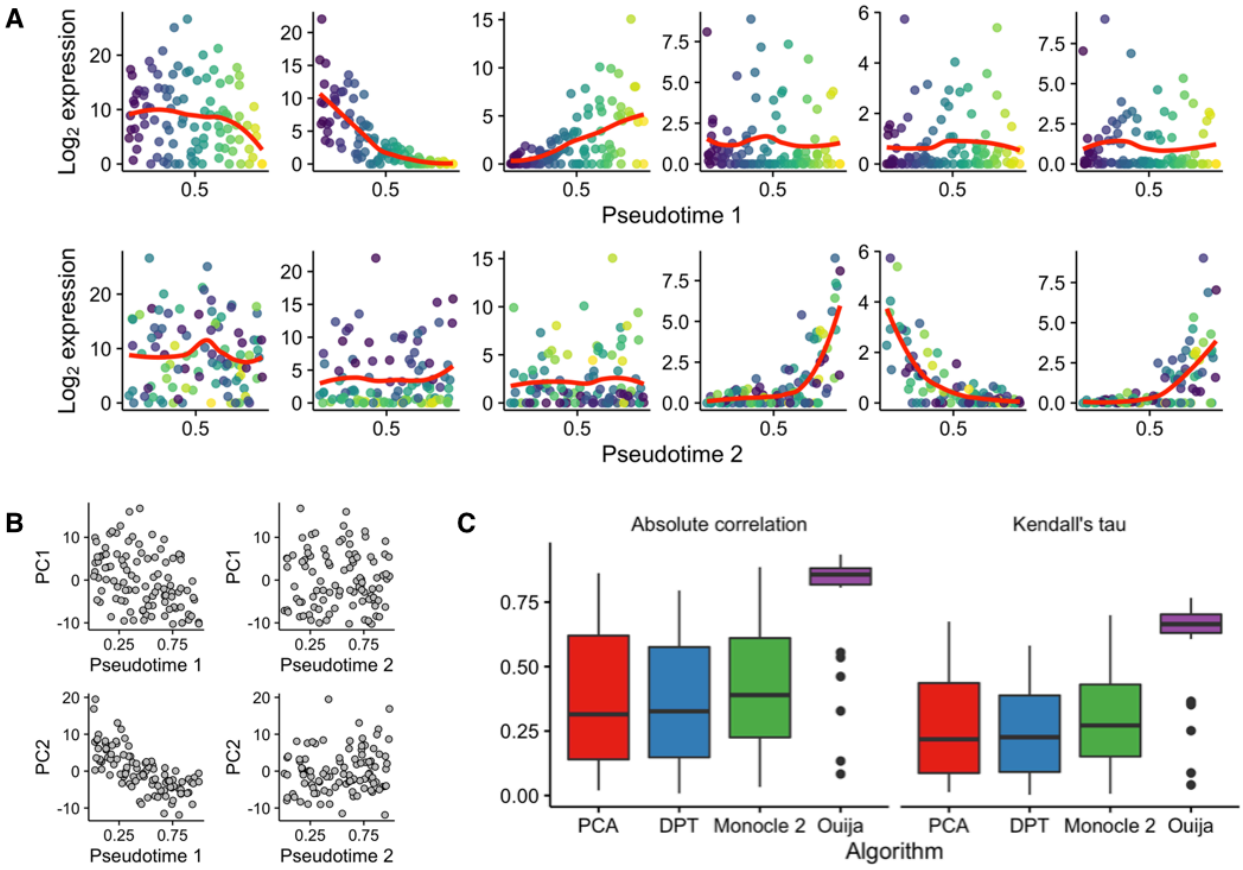


Figure 5: Performance of pseudotime algorithms in the presence of confounding pathways. 40 synthetic datasets were generated where half the genes correspond to the true pseudotime process of interest with the other half corresponding to a secondary nuisance trajectory. We benchmarked PCA, DPT and Monocle 2 against Ouija with informative priors. We encoded priors that expressed a preference for the inference of trajectory 1 with noninformative priors on genes involved in the second trajectory. The results show Ouija clearly outperforms the other algorithms with a median absolute Pearson correlation of 0.86 compared to 0.31 - 0.39.