A descriptive marker gene approach to single-cell pseudotime inference

Kieran R Campbell1,2,5 and Christopher Yau2,3,4∗

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
4 Centre for Computational Biology, University of Birmingham, UK
5 Current address: Department of Statistics, University of British Columbia, Vancouver BC, Canada

∗To whom correspondence should be addressed.
Associate Editor: XXXXXXX
Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Pseudotime estimation from single-cell gene expression data allows the recovery of temporal information from otherwise static profiles of individual cells. Conventional pseudotime inference methods emphasise an unsupervised transcriptome-wide approach and use retrospective analysis to evaluate the behaviour of individual genes. However, the resulting trajectories can only be understood in terms of abstract geometric structures and not in terms of interpretable models of gene behaviour.

Results: Here we introduce an orthogonal Bayesian approach termed “Ouija” that learns pseudotimes from a small set of marker genes that might ordinarily be used to retrospectively confirm the accuracy of unsupervised pseudotime algorithms. Crucially, we model these genes in terms of switch-like or transient behaviour along the trajectory, allowing us to understand why the pseudotimes have been inferred and learn informative parameters about the behaviour of each gene. Since each gene is associated with a switch or peak time the genes are effectively ordered along with the cells, allowing each part of the trajectory to be understood in terms of the behaviour of certain genes. We demonstrate that this small panel of marker genes can recover pseudotimes that are consistent with those obtained using the entire transcriptome. Furthermore, we show that our method can detect differences in the regulation timings between two genes and identify “metastable” states - discrete cell types along the continuous trajectories - that recapitulate known cell types.


Contact: kieran.campbell@stat.ubc.ca, c.yau@bham.ac.uk

Supplementary information: Supplementary text, figures, and tables are available at Bioinformatics online.

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 (Kalisky and Quake, 2011; Shapiro et al., 2013; Macaulay and Voet, 2014; Wills and Mead, 2015). The increased availability of single-cell data has driven the development of novel analytical methods specifically tailored to single cell properties (Stegle et al., 2015; Trapnell, 2015). The difficulties in conducting genuine time-series experiments at the single-cell level has led to the development of computational techniques known as pseudotime ordering algorithms that extract temporal information from snapshot molecular profiles of individual cells. These algorithms exploit studies in which the captured cells behave 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 (Qiu
et al., 2011; Bendall et al., 2014; Trapnell et al., 2014; Reid and Wernisch,
2015; Hancock et al., 2015; Shim et al., 2015; Haghverdi et al., 2016; Setty
et al., 2016).

A prominent feature of current pseudotime algorithms is that they
emphasise an "unsupervised" approach. The high-dimensional molecular
profiles for each cell are projected on to a reduced dimensional space
by using a (non)linear transformation of the molecular features. 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 including gene selection, dimensionality
reduction technique, and cell ordering algorithm, all of which 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
behaviours are for the correct developmental time period. Crucially, this
allows the pseudotime inference procedure to be understood in terms of
Copyright © 2018 Ouija-Bioinf. — ouija-bioinf.org — page 2 — #2

3.2 Input data normalisation
We index $N$ cells by $n \in 1, \ldots, N$ and $G$ genes by $g \in 1, \ldots, G$. Let $y_{ng} = [Y]_{ng}$ denote the log-transformed non-negative observed
cell-by-gene expression matrix. In order to make the strength parameters
comparable between genes we normalise the gene expression so the
approximate half-peak expression is 1 through the transformation $y_{ng} \rightarrow
y_{ng} = y_{ng} / \sigma_g$ where $\sigma_g$ is a gene-specific size factor defined by

$$ s_g = \frac{1}{|D_g|} \sum_{n \in D_g} y_{ng} $$

(2)

and $D_g = \{y_{ng} : y_{ng} > 0\}$.

3.3 Noise model
Our statistical model can be specified as a Bayesian hierarchical model
where the likelihood is given by a bimodal distribution formed from a
mixture of zero-component (dropout) and an non-zero expressing cell
population. If $\mu(t_n, \Theta_g)$ is the mean for cell $n$ and gene $g$ (evaluated
at pseudotime $t_n$ with gene-specific parameters $\Theta_g$) then

$$ \beta_0, \beta_1 \sim \text{Normal}(0, 1) $$

$$ \pi_{ng} \sim \text{Bernoulli}(\text{logit}^{-1}(\beta_0 + \beta_1 \mu(t_n, \Theta_g))) $$

$$ p(y_{ng} | \sigma_{ng}, \mu_{ng}, \pi_{ng}) = \pi_{ng} \phi(y_{ng}) $$

$$ + (1 - \pi_{ng}) T_s(y_{ng} \mu(t_n, \Theta_g), \sigma_{ng}^2) $$

(3)

where $\pi_{ng}$ is the probability of observing a dropout (zero-count) in
cell $n$ gene $g$ and $T_s$ is the density function of the Student-t distribution
with $v$ degrees of freedom.
The relationship between dropout rate and expression level is expressed
as a logistic regression model (Kharchenko et al., 2014). Furthermore,
we impose a mean-variance relationship of the form $\sigma_{ng}^2 = 1 + \phi \mu(t_n, \Theta_g) + \epsilon$ where $\phi$ is the dispersion parameter with prior $\phi \sim
\text{Gamma}(a_\phi, b_\phi)$, which is motivated by empirical observations of
marker gene behaviour (Supplementary Text 4.1).

3.4 Mean functions
We then need to specify the form of the mean functions $\mu(t_n, \Theta_g)$, for
which we consider both sigmoidal and transient gene behaviour. For genes
we expect to be a priori switch-like we model

$$ \mu(t_n, \Theta_g) = \frac{2 \theta_g}{1 + \exp \left(-k_g (t_n - t_{\theta_g}^{(0)})\right)} $$

(4)

where $k_g$ and $t_{\theta_g}^{(0)}$ denote the activation strength and activation time
parameters for each gene and $\theta_g$ the average peak expression with priors $\eta_g \sim \text{Gamma}(5/2, 1/2)$, $k_g \sim \text{Normal}(1/9, 1/9^2)$, $t_{\theta_g}^{(0)} \sim \text{TruncNormal}(0, 1)(\theta_g^{(1)}, 1/\theta_g^{(1)})$. If available, user-supplied prior
information can be encoded by specifying priors on the parameters
$\mu(g_k), \tau_k, \mu_\theta^{(1)}, \tau_\theta^{(1)}$. Otherwise, inference can be performed using
uninformative priors on these parameters. Specifying $\mu(g_k)$ encodes
an a priori belief in the strength and direction of the activation of gene $g$ along

2

Campbell and Yau

Page dimensions: 708.7x1000.6
where we take \( \theta_k \) (versely-) representing the strength of this belief. Similarly, specifying \( \mu \) \( \Theta \) encoding a prior belief of where in the trajectory gene \( g \) exhibits behaviour (either turning on or off) with \( \tau \) \( \tau \) encoding the strength of this belief.

For the transient case we have

\[
\mu(t_n, \Theta_k) = 2\eta_k \exp (-\lambda \theta_k (t_n - \tau)^2),
\]

where we take \( \lambda = 10 \) to be a constant and with prior structure \( \eta_k \sim \text{Gamma}(3/2, 1/2) \), \( \theta_k \sim \text{TruncNorm}(0,1,1/\delta) \), \( \tau \sim \text{TruncNorm}(0,\infty,1/\delta) \), where informative priors may be placed on \( \rho \) and \( \beta \) as before.

### 3.5 Inference

Under this framework learning the single-cell trajectory becomes Bayesian inference of \( \mu(t, \Theta|Y) \), the joint posterior distribution of the pseudotimes and gene behaviour parameters given the expression data. We performed posterior inference using Markov Chain Monte Carlo (MCMC) stochastic simulation algorithms, specifically the No U-Turn Hamiltonian Monte Carlo approach (Homan and Gelman, 2014) implemented in the STAN probabilistic programming language (Carpenter et al., 2015). The parameter \( \epsilon = 0.01 \) is used to avoid numerical issues in MCMC computation. For larger marker gene panels, such as in the cell cycle analysis section, we used stochastic gradient variational Bayes implemented in Edward (Tran et al., 2016) to perform approximate Bayesian inference.

### 4 Results & Discussion

#### 4.1 Pseudotime inference from small marker gene panels

The transcriptomes of both single cells and bulk samples exhibit remarkable correlations across genes and transcripts. Such concerted regulation of expression is thought to be due to pathway-dependent transcription (Tegge et al., 2012; Braun et al., 2008) and is necessary for the field of network inference from gene expression data (Langfelder and Horvath, 2008). An example of such transcriptome-wide correlations can be seen in Figure 2A for the Trapnell et al. (2014) dataset, where hierarchical clustering reveals a block-diagonal structure, implying an intrinsic low-dimensionality of the data that can be efficiently compressed using techniques such as principal components analysis (Supplementary Figure 1).

This redundancy of expression is often exploited by statistical models of single-cell RNA-seq data. Examples include Heimberg et al. (2016) where the intrinsic low-dimensionality is used to reconstruct transcriptome-wide expression from ultra-shallow read depths; Cleary et al. (2017) who propose a column subset selection procedure where a small number of genes are chosen to represent the full transcriptome. The compressibility of transcriptome data is likewise exploited by many single-cell pseudotime inference algorithms. A heterogeneous dispersed noise model with dropout is used to accurately model scRNA-seq data. Each gene's expression over pseudotime is modelled either through a sigmoidal shape (capturing both linear and switch-like behaviour) or through a Gaussian shape (capturing transient expression patterns). These models include several interpretable types of gene behaviour. A heteroskedastic dispersed noise model with dropout is used to accurately model scRNA-seq data. B The posterior distributions over the switch and peak times can be inferred as a sigmoidal shape (capturing both linear and switch-like behaviour) or through a Gaussian shape (capturing transient expression patterns). These models include several interpretable types of gene behaviour. A heteroskedastic dispersed noise model with dropout is used to accurately model scRNA-seq data. B Each gene's expression over pseudotime is modelled either through a sigmoidal shape (capturing both linear and switch-like behaviour) or through a Gaussian shape (capturing transient expression patterns). These models include several interpretable types of gene behaviour. A heteroskedastic dispersed noise model with dropout is used to accurately model scRNA-seq data.
Fig. 2. Transcriptome-wide pseudotimes can be inferred from small marker gene panels. A A gene-by-gene correlation matrix for the Trapnell et al. (Trapnell et al., 2014) dataset reveals similarities in the transcriptional response of hundreds of genes. The redundancy of expression implies the information content of the transcriptome may be compressed through techniques such as principal components analysis (PCA) or by picking informative marker genes. B Comparison of pseudotimes fitted using Ouija on a small panel of marker genes to transcriptome-wide fits (using the 500 most variable genes) across five datasets using the algorithms Monocle 2, DPT, and TSCAN. The marker gene fits show high correlation to the transcriptome-wide fits with the exception of the Shin et al. (Shin et al., 2015) dataset. C Gene expression profiles for two marker genes ID1 and MYOG from the Trapnell et al. (Trapnell et al., 2014) dataset. The solid red line denotes the maximum a posteriori (MAP) Ouija fit while the grey lines show draws from the posterior mean function. D Gene expression profiles for the same genes for the algorithms DPT, Monocle 2, and TSCAN show similar expression fits, demonstrating equivalent pseudotemporal trajectories have been inferred. The solid red line denotes a LOESS fit.

In Ouija, we exploit the high gene-gene correlations by modelling a small number of marker genes that are representative of the whole transcriptome. Such an approach is advantageous as by modelling the data directly rather than a reduced-dimension representation we can understand the pseudotimes for each cell in terms of the behaviour of genes through time rather than abstract notions of manifolds embedded in high-dimensional space. This takes the form of a nonlinear factor analysis model, departing from previous models that have relied upon linear factor analysis (Pierson and Yau, 2015; Campbell and Yau, 2017) by introducing sigmoidal nonlinearities and transient expression functions, both of which have been successfully applied previously in post-processing of single-cell trajectories (Campbell and Yau, 2016a,b; Sander et al., 2017).

We then turn to the question of how to choose the small number of marker genes in order to fit the pseudotimes. In single-cell pseudotime studies, the cells under examination undergo a known biological process such as differentiation or cell cycle. Importantly, key marker genes associated with these processes are usually known a priori by investigators. These marker genes act as positive controls whose behaviour is used post-hoc to confirm the validity of the transcriptome-wide pseudotime fit. Examples include the markers of myoblast differentiation MYH3, MEF2C, and MYOG in Trapnell et al. (2014); the markers of neurogenesis Gfap and Sox2 in Shin et al. (2015); and in Li et al. (2016) the authors tabulate the marker genes they expect to be involved in the process along with their expected behaviour along the differentiation trajectory. Given both the widespread a priori knowledge of such markers and their requirement to validate transcriptome-wide pseudotime fits, we therefore propose to derive pseudotimes directly from such markers.

We first sought to test whether our model applied to small panels of marker genes could accurately recapitulate the transcriptome-wide pseudotimes inferred by popular pseudotime methods. We applied Monocle 2, DPT, and TSCAN to five publicly available single-cell RNA-seq datasets (Trapnell et al., 2014; Shin et al., 2015; Zhou et al., 2016; Dulken et al., 2017; Chu et al., 2016) using the 500 most variable genes as input (the default in packages such as Scater (McCarthy et al., 2017) for PCA representations). For each dataset, we then inferred pseudotimes using Ouija based only on a small number of marker genes reported in each paper (ranging from 5 to 12), and compared the Pearson correlation between the Ouija pseudotimes and the pseudotimes reported for each dataset (Figure 2B). There was good agreement between the marker-based pseudotimes inferred using Ouija and the transcriptome-wide pseudotimes inferred using existing algorithms, with the correlation exceeding 0.75 in the majority of comparisons.

Noting that the correlation will not be 1 unless the algorithms are identical, we sought to compare Ouija’s correlation to transcriptome-wide
4.2 Gene regulation timing from marker gene-based pseudotime

Having demonstrated Ouija can accurately recapitulate transcriptome-wide pseudotimes using just small marker gene panels, we next sought to show how it allows for marker-driven interpretable inference of such trajectories. We applied Ouija to a single-cell time-series dataset of human embryonic stem cells differentiating into definitive endoderm cells (Chu et al., 2016). The authors examined the expression of key marker genes over time and found 9 to exhibit approximately switch-like behaviour (POUSF1, NANOG, SOX2, EOMES, CER1, GATA4, DKK4, MYCT1, and PRDM1) with a further two exhibiting transient expression (CDX1 and MSX2). We applied Ouija using informative priors over the behaviour parameters with no information about the capture times of the cells included.

The resulting pseudotime fit demonstrates we can understand single-cell pseudotime in terms of the behaviour of particular genes. Figure 3A shows a heatmap of the 9 switch-like genes (top) and 2 transient genes (bottom), ordered by the posterior switch time of each gene. It can be seen that the early trajectory is characterised by the expression of NANOG, SOX2, and POUSF1, which then leads to a cascade of switch-like activation of the remaining genes as the cells differentiate.

While transcriptome-wide pseudotime algorithms could provide similar heatmaps if the marker genes were known in advance, the key departure of Ouija is that we can quantitatively associate each gene with a region of pseudotime at which its regulation (switch time or peak time) occurs. This is illustrated in Figure 3B-C showing the posterior values for the regulation timing along with the associated uncertainty. In essence, Ouija allows us to order genes along trajectories as well as being able to order the cells, which provides insight into gene regulation relationships.

To approach such questions of gene regulation timings in a quantitative and rigorous manner we constructed a Bayesian hypothesis test to find out whether one gene is regulated before another given the noise in the data. If \( t_a < t_b \) are the regulation timings of genes A and B respectively, we calculate the posterior distribution \( p(t_a < t_b | \text{data}) \), and if both the lower and upper bounds of the 95% posterior credible interval fall outside 0 we say the two genes are regulated at significantly different times. We applied this to the pseudotime fit in the Chu et al. dataset, the results of which can be seen in Figure 3D for a subset of genes. The model suggests that EOMES is downregulated before DKK4 and MYCT1 is downregulated after PRDM1. Furthermore, it suggests the switch-like downregulation of DKK4 occurs after the transient peak-time of CDX1. However, it suggests the difference in regulation timings of DKK4 and MYCT1 are not significantly different from zero, which could imply co-regulation.
4.3 Ouija is robust to gene behaviour misspecification

A potential disadvantage of our model is the requirement to pre-specify genes as having switch-like or transient behaviour over pseudotime, which may result in biased or erroneous pseudotimes. We noticed such an effect in the Li et al. (2016) dataset, where the authors pre-specified how they expected several marker genes to behave over pseudotime. Upon fitting the pseudotimes using Ouija, we noted that the genes Mef2c and Pkd3/2 exhibited the correct upregulation over pseudotime (Supplementary Figure 4A), but that Scd1 that was supposed to exhibit transient, peaking expression was effectively constant along the trajectory (Supplementary Figure 4B).

We therefore sought to identify the existence of these discrete cell types along the continuous developmental trajectory. As Ouija uses a probabilistic model and inference we were able to obtain a posterior ordering “consistency” matrix (Figure 4A) where an entry in row i column j denotes the empirical probability that cell i is ordered before cell j.

Performing PCA on this matrix gives a rank-one representation of cell-cell continuity, which is then clustered using a Gaussian mixture model to find discrete cell states along the continuous trajectory (where the number of states is chosen such that the Bayesian information criterion is maximised).
4.5 Scalable pseudotime inference using TensorFlow

Finally, we wanted to consider a study composed of a large panel of putative marker genes to determine if Ouija could automatically identify genes satisfying its behavioural constraints. We identified a single-cell RNA-seq study (Kowalczyk et al., 2015) that examined variation between individual hematopoietic stem and progenitor cells from two mouse strains (C57BL/6 and DBA/2) as they age. Principal component analysis for each cell type and age showed a striking association of the top principal components with cell cycle-related genes (Figure 5A), indicating that transcriptional heterogeneity was dominated by cell cycle status. They scored each cell for its likely cell cycle phase using signatures based on functional annotations (of the Gene Ontology Consortium et al., 2009) and profiles from synchronized HEK cells (Whitfield et al., 2002) for the G1/S, S, G2, and G2/M phases.

We investigated if Ouija could be used to identify cell cycle phase, treating the inferential problem as a continuous pseudotime process and assuming all genes as candidate switch genes. We applied Ouija to 1,008 C57BL/6 HSCs using 374 GO cell cycle genes that satisfied gene selection criteria used in the original study. This large number of genes and cells makes inference using Hamiltonian Monte Carlo (HMC) slow so we implemented a second version of Ouija (termed Ouijaflow) which suggests that these cell types are not completely distinct in terms of expression. When examining the inferred pseudotime progression of each marker gene (Figure 4C), these three metastable states corresponded to the activation of all genes at the beginning of pseudotime, the complete inactivation of all the marker genes at the end of the pseudotime and a intervening transitional period as each marker gene turns off. Each metastable state clearly associates with a particular cell type with Npy2 and N12/2 exhibiting early down-regulation and Npy1, Hey1, Ephb2 and Ephb4 all exhibiting late down-regulation. Using this HSC formation system as a proof-of-principle it is evident that, if a small number of switch-like marker genes are known, it is possible to recover signatures of temporal progression using Ouija and that these trajectories are compatible with real biology.

To show the widespread applicability of this method we applied it to two further publically available datasets. Dutken et al. (Dutken et al., 2017) examined the trajectory of quiescent neural stem cells (qNSCs) as they differentiate into activated neural stem cells (NSCs) and neural progenitor cells (NPCs). Applying Ouija’s clustering-along-pseudotime revealed seven distinct clusters (Supplementary Figure 8; Supplementary Table 1) with clusters 1-2 corresponding to early and late qNSCs, cluster 3 defining the qNSC to NSC transition, clusters 4-6 corresponding to early to late aNSCs and cluster 7 defining the aNSC to NPC transition. We similarly applied this method to the Chu et al. dataset of time-series scRNA-seq that identified 8 distinct clusters along pseudotime (Supplementary Figure 9, Supplementary Table 2). Clusters 1-4 track the cells as the progress through the 4 stages from 0 hours to 36 hours, while clusters 5-8 track the 3 stages from 36 hours to 96th hours but with much more heterogeneity within each cluster, which is expected due to the longer time-scales considered.

The estimated pseudotime progression given by Ouija recapitulates the trajectory observed in principal component space (Figure 5A). The estimated pseudotime distribution correlates well with the cell cycle phase categorisation given in the original study (Figure 5C). Furthermore, we identified 88 genes with large activation strengths indicating strong switching-on behaviour (Figure 5D). Ordering the genes by activation time demonstrates a cascade of expression activation across these 88 genes over cell cycle progression with the quiescent (G0) indicated by complete inactivation of all 88 genes (Figure S2F). The explicit parametric model assumed by Ouija makes this gene selection and ordering process simple and quantitative compared to a non-parametric approach that would require some retrospective analysis or visual inspection.

5 Conclusion

We have developed a novel approach for pseudotime estimation based on modelling switch-like and transient expression behaviour for a small panel of marker genes chosen a priori. Our strategy provides an orthogonal and complementary 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, allowing us to explicitly determine a potential ordering of gene (de)activation and peaking events over pseudotime. The posterior ordering uncertainty can also be used to identify homogeneous metastable phases of transcriptional activity that might correspond to transient, but discrete, cell states. In summary, Ouija provides a novel contribution to the increasing plethora of pseudotime estimation methods available for single cell gene expression data.

Funding

KRC is supported by a doctoral studentship from the UK Medical Research Council and a postdoctoral fellowship from the Canadian Statistical Sciences Institute. CY is supported by the UK Medical Research Council.

References


Fig. 5. Cell cycle phase prediction. Principal component representation of hematopoietic stem cells coloured according to (A) the original cell cycle progression score (Kowalczyk et al., 2015) and (B) Ouija - cell cycle classes indicated are based on original study classifications. (C) Distribution of Ouija inferred pseudotime versus the original cell cycle classifications. (D) Estimated activation strengths for the 374 cell cycle gene panel. (E) Gene expression profile for 88 switch-like genes with cells ordered by pseudotime and (F) genes ordered by activation time.


Marker gene single-cell pseudotime inference


