Single-cell generalized trend model (scGTM): a flexible and interpretable model of gene expression trend along cell pseudotime

Elvis Han Cui [1,†], Dongyuan Song [2,†,*], Weng Kee Wong [1] and Jingyi Jessica Li [1,3,4,*]

[1]Department of Biostatistics, University of California, Los Angeles, CA 90095-1772, USA, [2]Bioinformatics Interdepartmental Ph.D. Program, University of California, Los Angeles, CA 90095-7246, [3]Department of Statistics, University of California, Los Angeles, CA 90095-1554, USA, and [4]Department of Computational Medicine, University of California, Los Angeles, CA 90095-1766, USA.

[†] These authors contributed equally to this work. *To whom correspondence should be addressed.

Abstract

Motivation: Modeling single-cell gene expression trends along cell pseudotime is a crucial analysis for exploring biological processes. Most existing methods rely on nonparametric regression models for their flexibility; however, nonparametric models often provide trends too complex to interpret. Other existing methods use interpretable but restrictive models. Since model interpretability and flexibility are both indispensable for understanding biological processes, the single-cell field needs a model that improves the interpretability and largely maintains the flexibility of nonparametric regression models.

Results: Here we propose the single-cell generalized trend model (scGTM) for capturing a gene's expression trend, which may be monotone, hill-shaped, or valley-shaped, along cell pseudotime. The scGTM has three advantages: (1) it can capture non-monotonic trends that are still easy to interpret, (2) its parameters are biologically interpretable and trend informative, and (3) it can flexibly accommodate common distributions for modeling gene expression counts. To tackle the complex optimization problems, we use the particle swarm optimization algorithm to find the constrained maximum likelihood estimates for the scGTM parameters. As an application, we analyze several single-cell gene expression data sets using the scGTM and show that it can capture interpretable gene expression trends along cell pseudotime and reveal molecular insights underlying the biological processes.

Availability and implementation: The Python package scGTM is open-access and available at https://github.com/ElvisCuiHan/scGTM.

Contact: jli@stat.ucla.edu

1 Introduction

Pseudotime analysis is one of the most important topics in single-cell transcriptomics. There has been fruitful work on inferring cell pseudotime (Magwene et al., 2003; Bendall et al., 2014; Trapnell et al., 2014; Shin et al., 2015; Ji and Ji, 2016; Qiu et al., 2017; Street et al., 2018; Mondal et al., 2021; Cao et al., 2019) and constructing statistical models for gene expression along the inferred cell pseudotime (Campbell and Yau, 2017; Bacher et al., 2018; Van den Berge et al., 2020; Ren and Kuan, 2020; Song and Li, 2021). Informative trends of gene expression along cell pseudotime may reflect molecular signatures in the biological processes. For instance, a gene may over time exhibit a *hill-shaped* trend (i.e. first-upward-then-downward) (Fig. 1b) or a *valley-shaped* trend (i.e. first-downward-then-upward) (Fig. 1c) trend and either of the trend may indicate the occurrence of some biological event. Hence, it is of great interest to have a statistical model that can capture the different gene expression trends accurately along cell pseudotime.

There are two types of statistical methods that have been developed to model the relationship between a gene's expression in a cell (or a sample) and the cell pseudotime. The latter is the sample's physical time and so

modeling issues encountered here are similar to those commonly seen in fitting a regression model to a data set in statistics. Methods of the first type are based on statistical regression models, such as generalized linear models (GLM) and generalized additive models (GAM), where frequently parameters in these models do not have direct relevance to gene expression dynamics. Specifically, the GLM used in the Monocle3 method (Cao et al., 2019) assumes that a gene's log-transformed expected expression in a cell is a linear function of the cell pseudotime, making it unable to capture hill- and valley-shaped trends that linear trends cannot approximate well. Consequently, most methods use nonparametric regression models, such as the GAM and piece-wise linear models, to capture complex trends. For example, Storey et al. (2005) applied basis regression; Trapnell et al. (2014) considered the GAM with the Tobit likelihood; Ren and Kuan (2020) applied the GAM with Bayesian shrinkage dispersion estimates; Van den Berge et al. (2020) proposed tradeSeq using the spline-based GAM. More recently, Song and Li (2021) proposed the PseudotimeDE method, which fixes the p-value calibration issue in tradeSeq and also uses the spline-based GAM with spline functions. Additionally, Bacher et al. (2018) used a piecewise linear model, which is more restrictive than the GAM. Although these nonparametric regression methods can fit complex gene expression trends, they are prone to over-fitting without proper hyper-parameter tuning (as we will show in Section 3) and their parameters either do not directly inform the shape of a trend (e.g., hill-shaped) or carry biological meanings.

Unlike the first type, methods of the second type use models with direct relevance to gene expression dynamics, and notable methods include ImpulseDE/ImpulseDE2 (Chechik and Koller, 2009; Sander et al., 2017; Fischer et al., 2018) and switchDE (Campbell and Yau, 2017). Specifically, ImpulseDE2 estimates a gene expression trend using a double-logistic curve to capture the non-monotone trends; however, even though the parameters have biological interpretations, they do not intuitively inform the shape of a trend. In contrast, switchDE uses a restrictive model with parameters that directly inform the shape of a trend (e.g., a gene's activation time) but is unable to detect non-monotonic trends.

The above review suggests that there is no current model that can capture monotone, hill-shaped, and valley-shaped trends with biologically interpretable and trend-informative parameters. To this end, we propose the scGTM that (i) can capture both hill- and valley-shaped trends and monotone trends, (ii) has interpretable and trend-informative parameters, and (iii) has flexible modeling for count data.

To estimate the scGTM parameters, we apply particle swarm optimization (PSO) to find the constrained maximum likelihood estimates (MLE) of the model parameters (Fig. S21). PSO has several advantages that make it suitable for our optimization problem: (i) it does not require the objective function to be convex or differentiable; (ii) it can handle boundary constraints and discrete parameters without having to re-formulate the objective function, and (iii) unlike the Newton-type algorithms used in Trapnell et al. (2014); Wood (2017); Campbell and Yau (2017), PSO is gradient-free. In addition, PSO codes are widely and freely available, easy to implement and its successes in tackling complex optimization problems are already well documented in computer science and engineering.

The rest of the paper is organized as follows. In Section 2, we introduce the scGTM and briefly review the PSO algorithm. In Section 3, we compare the scGTM with the GLM, GAM, ImpulseDE2, and switchDE and show its advantages in capturing informative, interpretable gene expression trends in two real data sets. Section 4 contains a discussion and future work.

2 Methods

2.1 The scGTM formulation

Let $\mathbf{Y} = (y_{gc})$ be an observed $G \times C$ gene expression count matrix, where G is the number of genes, C is the number of cells (i.e., the number of pseudotime values), and y_{gc} is the (g, c)-th element indicating the observed expression count of gene $g = 1, \ldots, G$ in cell $c = 1, \ldots, C$. We consider gene expression counts as random variables whose randomness comes from experimental measurement uncertainty, so y_{gc} is a realization of the random count variable Y_{gc}. Given a particular gene g, for notation simplicity, we drop the subscript g and denote Y_{gc} as Y_c and y_{gc} as y_c. We denote by $t_c \in [0, 1]$ the inferred (normalized) pseudotime of cell c. In the scGTM, $t_1, \ldots, t_C$ are treated as fixed values of pseudotime and serve as the covariate vector of interest.

Given t_c, the scGTM can model the count variable Y_c using four count distributions commonly used for gene expression data: the Poisson, negative binomial (NB), zero-inflated Poisson (ZIP), and zero-inflated negative binomial (ZINB) distributions.

For a hill-shaped gene, the scGTM is

$$Y_c \overset{\text{ind}}{\sim} F\left(\tau_c,\, \phi,\, p_c\right), \quad c = 1, \ldots, C, \tag{1}$$

$$\log(\tau_c + 1) = \begin{cases} b + \mu_{\text{mag}} \exp\left(-k_1 (t_c - t_0)^2\right) & \text{if } t_c \le t_0 \\ b + \mu_{\text{mag}} \exp\left(-k_2 (t_c - t_0)^2\right) & \text{if } t_c > t_0 \end{cases}, \tag{2}$$

$$\log\left(\frac{p_c}{1 - p_c}\right) = \alpha \log(\tau_c + 1) + \beta, \tag{3}$$

where $F\left(\tau_c,\, \phi,\, p_c\right)$ in (1) represents one of the four common count distributions. The most general case is when $F\left(\tau_c,\, \phi,\, p_c\right) = \text{ZINB}\left(\tau_c,\, \phi,\, p_c\right)$ with mean parameter $\tau_c \ge 0$, dispersion parameter $\phi \in \mathbb{Z}_+ := \{1, 2, 3, \cdots\}$ and zero-inflated parameter $p_c \in [0, 1]$. As special cases, $F\left(\tau_c,\, \phi,\, 0\right) = \text{NB}\left(\tau_c,\, \phi\right)$, $F\left(\tau_c,\, \infty,\, p_c\right) = \text{ZIP}\left(\tau_c,\, p_c\right)$, and $F\left(\tau_c,\, \infty,\, 0\right) = \text{Poisson}\left(\tau_c\right)$.

We design the parametric form (2) for the following reasons. First, on the left-hand side, $\log(\tau_c + 1)$ is motivated by the logarithmic link function used in the generalized linear models (GLM) and GAM. The addition of 1 is to ensure that $\log(\tau_c + 1) \ge 0$ so we can use $b \ge 0$ as the baseline of the hill-shaped trend (empirically, b is set to 0 and works well). Second, on the right-hand side, two partial Gaussian functions are adopted to model the trend's increasing and decreasing parts separately, so that the trend is allowed to be asymmetric (e.g., the increasing trend may be steeper or flatter than the decreasing trend). We choose to piece two partial Gaussian functions at the maximum into one function for two reasons: (1) the function is smooth (differentiable everywhere) and has a zero derivative at the maximum; (2) the function has tails that converge to the baseline b as the pseudotime t_c moves away from the mode t_0, a pattern that agrees with many biological processes.

Fig. 1a shows the roles of the four parameters μ_{mag}, k_1, k_2, and t_0 in (2) for modelling a hill-shaped trend. For a valley-shaped trend, there are four similar parameters and we note that a monotone increasing trend is a special case of a hill-shaped trend with the increasing part only. The four parameters in the Fig. 1a are the maximum log expected expression μ_{mag}, the activation strength k_1, the repression strength k_2, and the change time t_0 where the expected expression stops increasing. Fig. 1b–c show the scGTM fits to the gene (*Tmsb10* in the GYRUS data set and another gene *NFKBIA* in the LPS data set from the Supplementary Table S1). Their trends reveal a hill- and valley-shaped trend., respectively.

In this hill-shaped scGTM, we assume that the gene's expression count Y_c in cell c has mean parameter τ_c and zero-inflation parameter p_c, and both depend on the pseudotime t_c of cell c. In (2), we link τ_c to t_c by assuming that $\log(\tau_c + 1)$ is a non-negative transformation that compresses extremely large values of τ_c using a two-part Gaussian

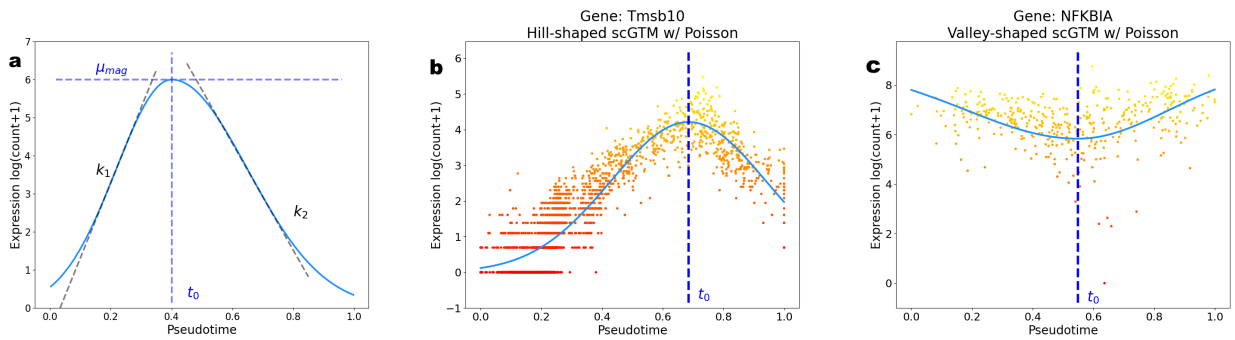

Fig. 1. Illustration of the scGTM. (**a**) Four parameters of the scGTM in Equation (2) for a hill-shaped trend: the maximum log expected expression μ_{mag} (horizontal blue line), the activation strength k_1 (absolute value of the left tangent line's slope), the repression strength k_2 (absolute value of the right tangent line's slope), and the change time t_0 (vertical blue line). (**b**) A hill-shaped trend of gene *Tmsb10* (in the GYRUS dataset) fitted by the scGTM with counts modeled by the Poisson distribution. (**c**) A valley-shaped trend of gene *NFKBIA* (in the LPS dataset) fitted by the scGTM with counts modeled by the Poisson distribution. In b–c, the scatter points indicate gene expression levels, and the curves are the trends fit by the scGTM.

function corresponding to $t_c \leq t_0$ and $t_c > t_0$; we choose the Gaussian function for its good mathematical properties and interpretability. We link p_c to t_c in (3) using a logistic regression, with predictor $\log(\tau_c + 1)$, i.e., the logistic transformation of p_c is a linear function of $\log(\tau_c + 1)$ (with slope α and intercept β) and thus a function of t_c.

Besides $\phi \in \mathbb{Z}_+$ and $\alpha, \beta \in \mathbb{R}$, the following parameters of the hill-shaped scGTM shown in Fig. 1a need to be estimated for biological interpretations:

- $\mu_{\text{mag}} \geq 0$: magnitude of the hill,
  i.e., $\mu_{\text{mag}} = \max_{c \in \{1,\ldots,C\}} \log(\tau_c + 1)$;
- $k_1 \geq 0$: activation strength (how fast the gene is up-regulated);
- $k_2 \geq 0$: repression strength (how fast the gene is down-regulated);
- $t_0 \in [0, 1]$: change time (where the gene reaches the maximum expected expression). It is within $[0, 1]$ since the pseudotime has been re-scaled.

For a valley-shaped gene, the scGTM is the same except that we replace (2) by

$$\log(\tau_c + 1) = \begin{cases} b - \mu_{\text{mag}} \exp\left(-k_2 (t_c - t_0)^2\right) & \text{if } t_c \leq t_0 \\ b - \mu_{\text{mag}} \exp\left(-k_1 (t_c - t_0)^2\right) & \text{if } t_c > t_0 \end{cases}, \quad (4)$$

where b indicates the baseline (maximum) log-transformed (expected expression $+ 1$) of the valley-shaped gene. The interpretation of the four key parameters of the valley-shaped scGTM becomes

- $\mu_{\text{mag}} \in [0, b]$: magnitude of the valley, i.e., $b - \mu_{\text{mag}} = \min_{c \in \{1,\ldots,C\}} \log(\tau_c + 1)$;
- $k_1 \geq 0$: activation strength (how fast the gene is up-regulated);
- $k_2 \geq 0$: repression strength (how fast the gene is down-regulated);
- $t_0 \in [0, 1]$: change time (where the gene reaches the minimum expected expression).

Compared to the hill-shaped scGTM, the valley-shaped scGTM has an additional baseline parameter b that needs to be estimated. For simplicity, we estimate b by a plug-in estimator $\hat{b} = \max_{c \in \{1,\ldots,C\}} \log(y_c + 1)$, where $y_1, \ldots, y_C$ are the observed counts of a valley-shaped gene. This estimate is justified by the fact that the maximum likelihood estimate (MLE) of the upper bound parameter of a domain is the maximum order statistic; i.e., if $x_1, \ldots, x_n$ are randomly sampled from a distribution with domain $[a, b]$, then $\hat{b} = \max_{i \in \{1,\ldots,n\}} x_i$ is the MLE of b. For the common parameters of the hill- and valley-shaped scGTMs in Section 2.2, we next discuss how PSO can provide constrained likelihood estimates for these parameters.

2.2 Constrained MLE and the PSO algorithm

To fit the scGTM to a gene, we first need to ascertain whether the gene is hill- or valley-shaped: we first need to ascertain whether the gene is hill- or valley-shaped: we fit both hill- and valley-shaped models to the gene's data and choose the model that has the smaller Akaike information criterion (AIC) value (see Supplementary Information S9 for the model selection results for the two genes in Fig. 1b–c). Next, based on the trend shape, we estimate the scGTM parameters. For a hill-shaped gene, we estimate the scGTM parameters $\Theta = (\mu_{\text{mag}}, k_1, k_2, t_0, \phi, \alpha, \beta)^\mathsf{T}$ from the observed expression counts $\boldsymbol{y} = (y_1, \ldots, y_C)^\mathsf{T}$ and cell pseudotimes $\boldsymbol{t} = (t_1, \ldots, t_C)^\mathsf{T}$ using the constrained maximum likelihood method, which respects each parameter's range and ensures the estimation stability. Let $\log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t})$ be the log likelihood function and the optimization problem is:

$$\begin{aligned} &\max_{\Theta} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) \text{ such that} \\ &\min_{c \in \{1,\ldots,C\}} \log(y_c + 1) \leq \mu_{\text{mag}} \leq \max_{c \in \{1,\ldots,C\}} \log(y_c + 1)\,, \\ &k_1, k_2 \geq 0\,, \quad \min_{c \in \{1,\ldots,C\}} t_c \leq t_0 \leq \max_{c \in \{1,\ldots,C\}} t_c\,, \ \phi \in \mathbb{Z}_+\,, \end{aligned} \quad (5)$$

where

$$\begin{aligned} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}) &= \log \left[\prod_{c=1}^{C} \mathbb{P}(Y_c = y_c \mid t_c) \right] \\ &= \sum_{c=1}^{C} \log \left[(1 - p_c) f(y_c | t_c) + p_c \, \mathbb{I}(y_c = 0) \right] \end{aligned} \quad (6)$$

and

$$f(y_c | t_c) = \frac{\tau_c^{y_c}}{y_c!} \frac{\Gamma(\phi + y_c)}{\Gamma(\phi)(\phi + \tau_c)^{y_c}} \frac{1}{\left(1 + \frac{\tau_c}{\phi}\right)^{\phi}}\,,$$

which can be further specified as a function of Θ based on (2) and (3).

For a valley-shaped gene, the constrained MLE problem is similar, and we omit the discussion for space consideration.

There are two difficulties in the optimization problem (5). First, the likelihood function (6) is neither convex nor concave. Second, the constraint is linear in μ_{mag}, k_1, k_2, and t_0 but ϕ is a positive integer-valued variable. Hence, conventional optimization algorithms such as P-IRLS in GAM (Wood, 2011, 2017) and L-BFGS in switchDE (Van Loan and Golub, 1996; Campbell and Yau, 2017) are difficult to apply in this case. Metaheuristics is a class of assumptions-free general purpose optimization algorithms that is widely and increasingly used to tackle challenging

and high-dimensional optimization problems in the quantitative sciences (Whitacre, 2011a,b; Yang, 2017). PSO is an exemplary metaheuristic algorithm, and it has effectively solved various types of optimization problems. Korani and Mouhoub (2021) is a recent review of metaheuristic algorithms and their applications across various disciplines.

PSO first generates a swarm of candidate solutions (known as particles) to the optimization problem (5). At each iteration, particles change their positions within the constraints, and the algorithm finds the best solution among all particle trajectories. We summarize the vanilla PSO algorithm (Bratton and Kennedy, 2007) for the constrained MLE of the scGTM in Algorithm 1, and provide further details of PSO in the Supplementary Information.

Algorithm 1 PSO for the constrained MLE for the scGTM

Input data: a gene's expression counts and cell pseudotime values

$\boldsymbol{y}$: a $C \times 1$ gene expression count vector;

$\boldsymbol{t}$: a $C \times 1$ cell pseudotime vector;

Input parameters:

F: count distribution: Poisson, NB, ZIP, or ZINB;

H: number of iterations in PSO; set to $H = 100$ by default;

w, c_1, and c_2: hyperparameters of PSO; set to $w = 0.9$, $c_1 = 1.2$, and $c_2 = 0.3$ by default;

Algorithm:

1. Randomly initialize Θ with B particles: $\Theta_1^0, \Theta_2^0, \ldots, \Theta_B^0$;

2. Randomly initialize velocity vectors for the B particles: $\boldsymbol{v}_1^0, \boldsymbol{v}_2^0, \cdots, \boldsymbol{v}_B^0$;

3. For $h = 0$ to H:

(i) Update the best solution of each particle i

$$\widehat{\Theta}_i^h = \underset{\Theta \in \mathcal{A}_i^h}{\arg\max} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t}),$$

where $\mathcal{A}_i^h = \{\Theta_i^k : k = 0, \cdots, h\}$, $i = 1, \ldots, B$;

(ii) Update the global best solution

$$\widehat{\Theta}^h = \underset{\Theta \in \cup_{i=1}^B \mathcal{A}_i^h}{\arg\max} \log L(\Theta \mid \boldsymbol{y}, \boldsymbol{t});$$

(iii) Update velocity of each particle i

$$\boldsymbol{v}_i^{h+1} = w\boldsymbol{v}_i^h + c_1 r_{i1}^h \left(\widehat{\Theta}_i^h - \Theta_i^h\right) + c_2 r_{i2}^h \left(\widehat{\Theta}^h - \Theta_i^h\right),$$

where r_{i1}^h and r_{i2}^h are independently generated from Unif$(0, 1)$, $i = 1 \ldots, B$;

(iv) Update each particle i

$$\Theta_i^{h+1} = \Theta_i^h + \boldsymbol{v}_{i+1}^h, \quad i = 1, \ldots, B;$$

4. Set $\widehat{\Theta} = \widehat{\Theta}^H$;

5. Calculate 95% approximate confidence intervals of key parameters based on $\widehat{\Theta}$ (Section 2.3).

Output:

$-\log L(\widehat{\Theta} \mid \boldsymbol{y}, \boldsymbol{t})$: fitted negative log likelihood value;

$\widehat{\Theta} = (\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0, \hat{\phi}, \hat{\alpha}, \hat{\beta})^\mathsf{T}$: estimated parameters;

$[\hat{\mu}_{\text{mag}}^{\text{lb}}, \hat{\mu}_{\text{mag}}^{\text{ub}}]$, $[\hat{k}_1^{\text{lb}}, \hat{k}_1^{\text{ub}}]$, $[\hat{k}_2^{\text{lb}}, \hat{k}_2^{\text{ub}}]$, and $[\hat{t}_0^{\text{lb}}, \hat{t}_0^{\text{ub}}]$: 95% approximate confidence intervals.

2.3 Approximate confidence intervals of the four key parameters in the scGTM

The estimated parameters $\widehat{\Theta} = (\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0, \hat{\phi}, \hat{\alpha}, \hat{\beta})^\mathsf{T}$ are next used to construct approximate confidence intervals for μ_{mag}, k_1, k_2, and t_0 using the maximum likelihood theory. Specifically, we calculate the plug-in asymptotic covariance matrix of $(\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0)^\mathsf{T}$ as the inverse of the partial Fisher information matrix of the four parameters evaluated at $(\hat{\mu}_{\text{mag}}, \hat{k}_1, \hat{k}_2, \hat{t}_0)^\mathsf{T}$ (detailed derivation in the Supplementary Information). Then we use the diagonal elements of this matrix as the plug-in asymptotic variances of $\hat{\mu}_{\text{mag}}$, $\hat{k}_1$, $\hat{k}_2$, and $\hat{t}_0$, and denote them by $\widehat{\text{Var}}(\hat{\mu}_{\text{mag}})$, $\widehat{\text{Var}}(\hat{k}_1)$, $\widehat{\text{Var}}(\hat{k}_2)$, and $\widehat{\text{Var}}(\hat{t}_0)$, respectively. We then obtain a 95% approximate confidence interval for each of the parameters: $[\hat{\mu}_{\text{mag}}^{\text{lb}}, \hat{\mu}_{\text{mag}}^{\text{ub}}]$, $[\hat{k}_1^{\text{lb}}, \hat{k}_1^{\text{ub}}]$, $[\hat{k}_2^{\text{lb}}, \hat{k}_2^{\text{ub}}]$, and $[\hat{t}_0^{\text{lb}}, \hat{t}_0^{\text{ub}}]$, where

$$\hat{\mu}_{\text{mag}}^{\text{lb}} = \max\left(0,\ \hat{\mu}_{\text{mag}} - 1.96\sqrt{\widehat{\text{Var}}(\hat{\mu}_{\text{mag}})}\right),\ \hat{\mu}_{\text{mag}}^{\text{ub}} = \hat{\mu}_{\text{mag}} + 1.96\sqrt{\widehat{\text{Var}}(\hat{\mu}_{\text{mag}})},$$

$$\hat{k}_1^{\text{lb}} = \max\left(0,\ \hat{k}_1 - 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_1)}\right),\ \hat{k}_1^{\text{ub}} = \hat{k}_1 + 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_1)},$$

$$\hat{k}_2^{\text{lb}} = \max\left(0,\ \hat{k}_2 - 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_2)}\right),\ \hat{k}_2^{\text{ub}} = \hat{k}_2 + 1.96\sqrt{\widehat{\text{Var}}(\hat{k}_2)},$$

$$\hat{t}_0^{\text{lb}} = \max\left(0,\ \hat{t}_0 - 1.96\sqrt{\widehat{\text{Var}}(\hat{t}_0)}\right),\ \hat{t}_0^{\text{ub}} = \min\left(\hat{t}_0 + 1.96\sqrt{\widehat{\text{Var}}(\hat{t}_0)},\ 1\right).$$

3 Results

3.1 scGTM outperforms GAM, GLM, LOESS, switchDE, and ImpulseDE2 in capturing informative and interpretable trends

As an example, we use the *MAOA* gene in the WANG dataset (Wang et al., 2020) (Supplementary Table S1) to compare the fitted trends of the scGTM, GAM, GLM, LOESS, switchDE, and ImpulseDE2. In the original study, the gene was reported to have a hill-shaped trend. Our comparison results have several interesting observations. First, we show that the scGTM provides more informative and interpretable gene expression trends than the GAM and GLM do when the count outcome comes from the Poisson, ZIP, NB, and ZINB distributions. Fig. 2a shows that the scGTM robustly captures the hill-shaped trends for the four distributions and consistently estimates the change time around 0.75, which is where the *MAOA* gene reaches its expected maximum expression. While the GAM also estimates the maximum expression around 0.75, its estimated trends are much more complex. This is likely due to possible overfitting (despite the use of penalization) and consequently, more difficult to interpret them than the scGTM trends (Fig. 2b). Unlike the scGTM and GAM, the GLM only allows for capturing monotone trends, making it unable to detect the possible existence of expression change time (Fig. 2c). Second, we compare the scGTM with the two existing methods, switchDE and ImpulseDE2, that use models with direct relevance to gene expression dynamics. Although switchDE estimates the activation time around 0.75, similar to the scGTM's estimate change time, it cannot capture the downward expression trend as the cell pseudotime approaches 1.00 due to its monotone nature (Fig. 2d). ImpulseDE2 can theoretically capture a hill-shaped trend, but it only fits a monotone increasing trend for the *MAOA* gene (Fig. 2e). A likely reason is that the method was designed for time-course bulk RNA-seq data. Third, we compare the scGTM with the LOESS method commonly used for exploratory data analysis. While LOESS outputs a reasonable, though less smooth trend (Fig. 2f), it is not probability-based and thus does not have a likelihood. Hence, LOESS does not allow likelihood-based model selection, a functionality of the scGTM. To summarize, the scGTM fits outperform those from GAM, GLM, LOESS, switchDE and ImpulseDE2 by providing a more informative and interpretable trend with less concern on model overfitting.

In addition to the *MAOA* gene, Wang et al. (2020) reported 19 other exemplary genes that define menstrual cycle phases and exhibit hill-shaped expression trends along the cell pseudotime. Supplementary Figs. S1-S19 compare the various model fits for the 19 genes and and we observe that

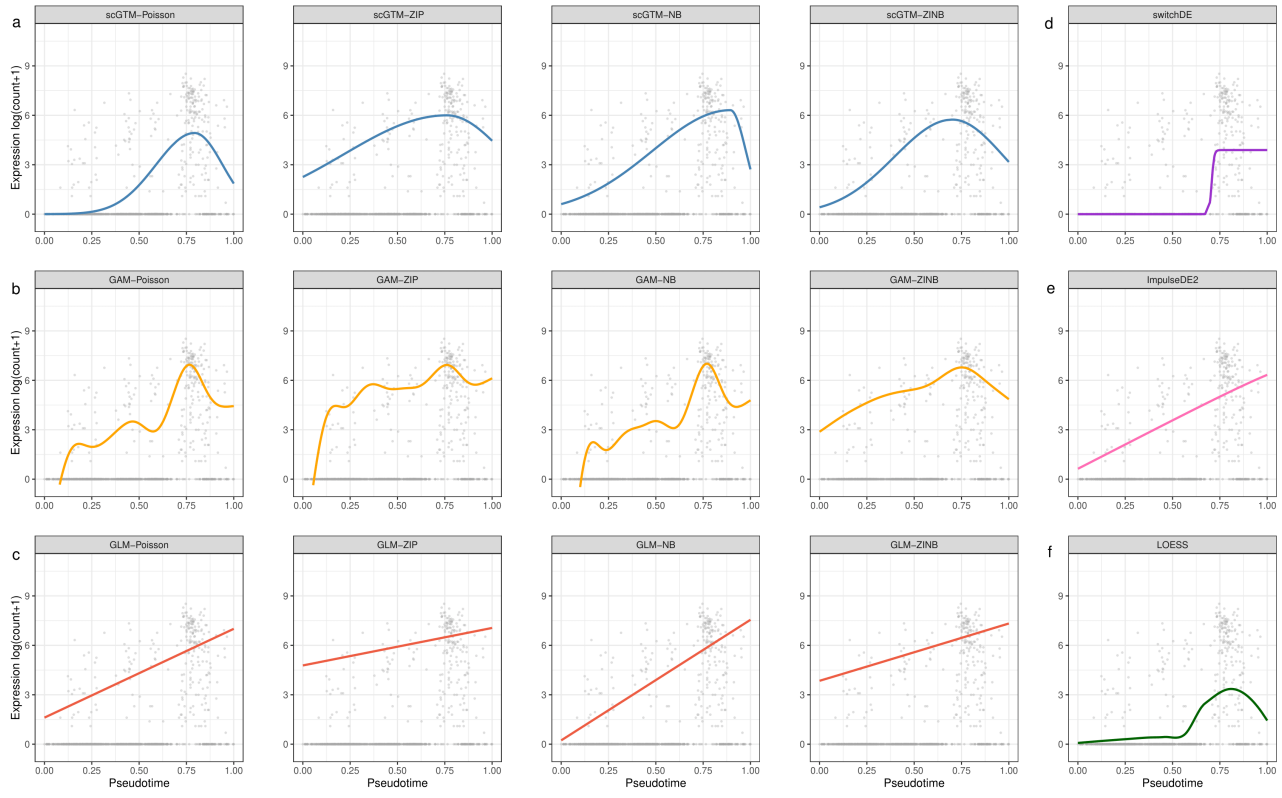

Fig. 2. Comparison of the scGTM with GAM, GLM, LOESS, switchDE, and ImpulseDE2 for fitting the expression trend of gene *MAOA* in the WANG dataset (Wang et al., 2020) (Supplementary Table S1). In the first four columns, the three rows correspond to (a) scGTM, (b) GAM, and (c) GLM. From left to right, the first four columns correspond to Poisson, ZIP, NB, and ZINB as the count distribution used in the scGTM, GAM, and GLM. The fifth column corresponds to (d) switchDE, (e) ImpulseDE2 and (f) LOESS. Each panel shows the same scatterplot of gene *MAOA*'s log-transformed expression counts vs. cell pseudotime values, as well as a model's fitted trend. With all four count distributions, the scGTM robustly captures the gene expression trend and estimates the change time around 0.75. In contrast, GLM, switchDE and ImpulseDE2 only describe the trend as increasing; GAM overfits the data and does not output trends as interpretable as the scGTM does; LOESS outputs a reasonable trend, but it does not allow likelihood-based model selection like the scGTM.

the scGTM consistently provides more informative, interpretable trends than the other models.

Besides visually inspecting the fitted expression trends, we compare the AIC values of the scGTM, GAM, and GLM used with the four count distributions fitted to the aforementioned 20 genes. Note that a lower AIC value indicates a model's better goodness-of-fit with the model complexity penalized. Supplementary Fig. S20 shows that the scGTM has comparable or even lower AIC values than the GAM's AIC values, confirming that the scGTM fits well to data despite its much simpler model than GAM's. Based on Fig. 2 and Supplementary Figs. S1–S20, we use the scGTM with the Poisson distribution in the following applications for its goodness-of-fit and model simplicity. This choice is consistent with previous research on modeling sequencing data (Silverman et al., 2020) and other count data (Warton, 2005; Campbell, 2021).

3.2 scGTM recapitulates gene expression trends of endometrial transformation in the human menstrual cycle

The WANG dataset contains 20 exemplar genes that exhibit temporal expression trends in unciliated epithelia cells in the human menstrual cycle Wang et al. (2020). The original study also ordered the 20 genes by the estimated pseudotime at which they achieved the maximum expression (Fig. 3a; genes ordered from top to bottom), and it was found that the ordering agreed well with the menstrual cycle phases (Fig. 3a; the top bar indicates the phases). Comparing the fitted expression trends of the 20 genes by the scGTM, switchDE, and ImpulseDE2, we observe that only the scGTM trends agree well with the data (Fig. 3). Additionally, we evaluate the 20 genes' estimated change times (i.e., t_0) by the scGTM and their estimated activation times by the switchDE. Although the change times and estimation times are both expected to correlate well with the gene ordering in the original study, only the change times estimated by the scGTM fulfills this expectation (Fig. 3b–c). Compared with the scGTM, switchDE miscalculates the activation times for many hill-shaped genes whose maximum expression occurs in the middle of the cycle; this is likely due to the fact that switchDE can only capture monotone trends (Fig. 3c). Similarly, ImpulseDE2 cannot well capture the trends of those hill-shaped genes (Fig. 3d). Unlike switchDE and ImpulseDE2, the scGTM estimates the change times reasonably for almost all genes. For instance, the *GPX3* gene has an estimated change time at 0.88, consistent with its role as a secretory middle/late phase marker gene Wang et al. (2020).

Besides the 20 exemplar genes, we apply the scGTM, switchDE, and ImpulseDE2 to fit the expression trends of all $1{,}382$ menstrual cycle genes reported in Wang et al. (2020). Supplementary Fig. S28 shows that the scGTM still outperforms switchDE and ImpulseDE2 for capturing these genes' expression trends. In summary, the scGTM provides useful summaries for gene expression trends in the human menstrual cycle.

3.3 scGTM identifies informative gene expression trends after immune cell stimulation

As the second real data application, we use the scGTM to categorize gene expression trends in mouse dendritic cells (DCs) after stimulation

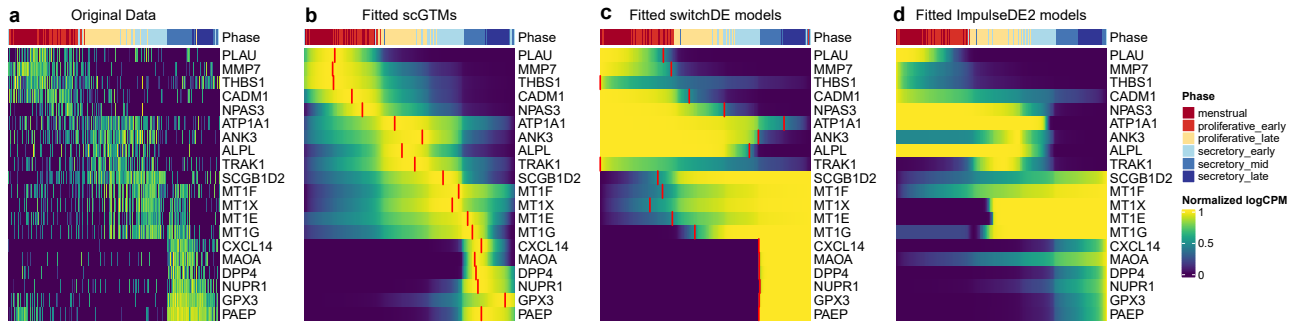

Fig. 3. Fitted expression trends by the scGTM, switchDE, and ImpulseDE2 for 20 exemplar genes in the WANG dataset (Wang et al., 2020) (Supplementary Table S1). All panels are ordered by cell pseudotime values from 0 (left) to 1 (right). The top color bars show the endometrial phases defined in the original study. (a) The original expression values along pseudotime. (b) The fitted trends of the scGTM, with the red segments highlighting the estimated change times t_0. (c) The fitted trends of switchDE, with the red segments highlighting the estimated activation times. (d) The fitted trends of ImpulseDE2.

with lipopolysaccharide (LPS, a component of gram-negative bacteria) Shalek et al. (2014). First, we apply the likelihood ratio tests to screen the genes that the scGTM fits significantly better than the null Poisson model (in which τ_c and p_c in (1) do not depend on cell pseudotime t_c). Assuming that the likelihood ratio statistic of every gene follows χ^2_3 as the null distribution, we retain 2405 genes whose Benjamini-Hochberg (BH) adjusted p-values ≤ 0.01.

Second, we use the scGTM's confidence levels of the three parameters t_0, k_1, and k_2 to categorize the 2405 genes into three types: (1) *hill-shaped & mostly increasing genes*: $t_0^{\text{lb}} > 0.5 + 0.1$ (change time occurs at late pseudotime) and $k_1^{\text{lb}} > 1$ (strong activation strength), (2) *hill-shaped & mostly decreasing genes*: $t_0^{\text{ub}} < 0.5 - 0.1$ (change time occurs at early pseudotime) and $k_2^{\text{lb}} > 1$ (strong repression strength), and (3) *valley-shaped genes*. To demonstrate that this categorization is biologically meaningful, we perform gene ontology (GO) analysis on the three gene types and compare the enriched GO terms. Fig. 4a shows that the three gene types are enriched with largely unique GO terms, verifying their functional differences. Notably, the hill-shaped & mostly increasing genes are related to immune response processes, showing consistency between their expression trends (activation after the LPS stimulation) and functions (immune response). Further, we visualize 5 illustrative genes from each gene type (Fig. 4b) and observe that the scGTM's fitted trends agree well with the data. In conclusion, the scGTM can help users discern genes with specific trends by its trend-informative parameters.

Besides the above three real data applications, we conduct a simulation study to verify the robustness of the scGTM to gene expression trends not generated from the scGTM assumptions. The simulation results also show that, beyond good interpretability, the scGTM is flexible enough to fit various trends to a similar extent as the GAM does (Supplementary Information S3). Moreover, we use a bootstrap analysis to show that the fitted scGTM trend has a smaller variance than the fitted GAM trend does (Supplementary Information S3), at the cost of a larger bias.

4 Discussion

We propose the scGTM as a flexible and interpretable statistical model for studying single-cell gene expression trend along cell pseudotime. Using four count distributions and two real data sets, we demonstrated that the scGTM has interpretable parameters that can directly inform a trend for gene expression counts. The scGTM parameters are estimated by the constrained maximum likelihood estimation via PSO, which is one of the most popular metaheuristic algorithms for function optimization. We showed that scGTM has distinct advantages over the classic models GLM and GAM and the two recent methods switchDE and ImpulseDE2 in that it can uniquely capture robust, informative, and interpretable trends. In contrast, the GLM and switchDE can only estimate monotonic trends; the GAM often provides trends that are too complex to interpret, and ImpulseDE2 (a method designed for bulk RNA-seq data) does not have stable performance on single-cell data. The estimated parameters and confidence intervals from the scGTM are then used to characterize the expression trends of the genes.

Note that the scGTM is extendable by assuming a more complicated mean function, whose estimation can still be achieved by the PSO algorithm (whose major advantage is its flexibility). To demonstrate this functionality of the scGTM, we have added a simulation in Supplementary Information S10, where we use the sine function to generate one gene's true expression trend along the pseudotime. With its mean function set as as the sine function, the scGTM accurately estimates the gene trend (Supplementary Fig. S25). In a future version of the scGTM package, we can allow users to input specified mean functions that reflect the gene expression trends they are interested in. On the other hand, if users do not have a prior preference for the gene expression trends, we would recommend them to use the generalized additive models that can capture flexible trends.

Strictly speaking, the inference of the scGTM has two caveats. First, the parameter estimation includes a double-dipping procedure: the same data are first used to decide whether a trend is hill- or valley-shaped and it is used to estimate the parameters. Second, since only the key parameters μ_{mag}, k_1, k_2, and t_0 are inferential targets, the other parameters ϕ, α and β should be regarded "nuisance" parameters. However, the construction of confidence intervals of the key parameters does not account for these two caveats and would thus result in overly optimistic confidence intervals. We will investigate how to obtain better-calibrated confidence intervals in future research.

In our previous work (Song and Li, 2021), we developed a method PseudotimeDE to account for the uncertainty of inferred pseudotime on the inference of differentially expressed genes along the pseudotime. Note that PseudotimeDE is directly extendable to the scGTM, by just replacing the GAM in PseudotimeDE by the scGTM. However, here our focus is on proposing the scGTM for interpreting a trend, instead of testing whether a trend is different from a horizontal line, i.e., the focus of PseudotimeDE. Hence, we leave the incorpration of the scGTM into PseudotimeDE to future work. Moreover, we have a simulation study to show that the fitted scGTM trends have shapes largely robust to noise added to pseudotime (**Supplementary Information S4**).

The current implementation of the scGTM is only applicable to a single pseudotime trajectory (i.e., cell lineage). A natural extension is to split a

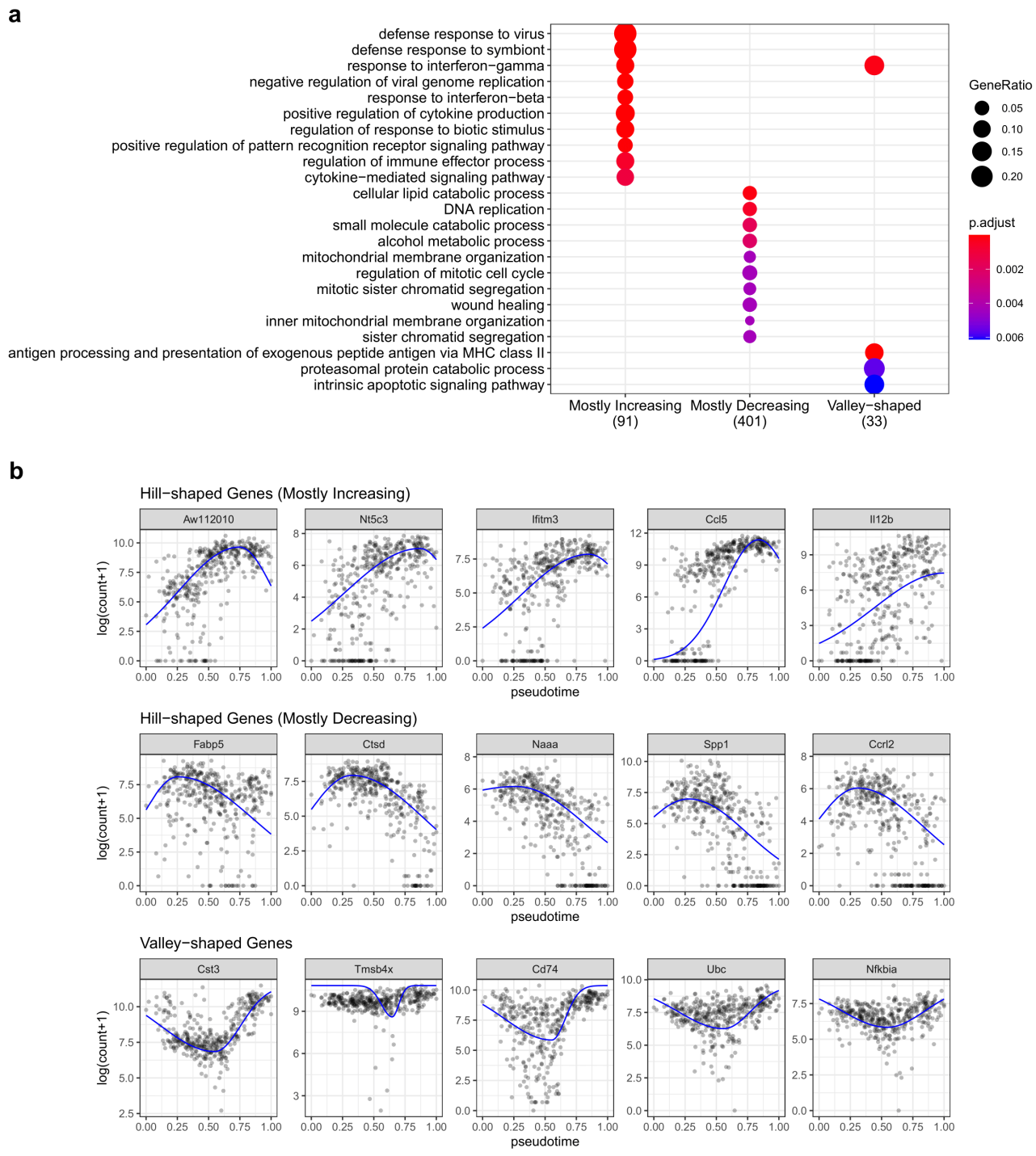

Fig. 4. Three types of gene expression trends characterized by the scGTM parameters in the LPS dataset (Supplementary Table S1). (a) GO enrichment analysis of the three gene types. The top enriched GO terms are different among the three gene types. Notably, the hill-shaped & mostly increasing genes (1st column) are functionally related to immune responses. (b) Visualization of example genes in the three types. The scatter plots show gene expression data; the trends estimated by the scGTM (blue curves) well match the data.

multiple-lineage cell trajectory into single lineages and fit the scGTM to each lineage separately.

In addition, the vanilla PSO algorithm in this paper handles each parameter's constraint separately. Hence, if we need a constraint on more than one parameter, e.g., k_1/k_2 should be within a a user-specified range, then we have to develop a variant algorithm of PSO or use other metaheuristics algorithms.

5 Data Availability

The code and data for genearing the results are available at zenodo (doi: 10.5281/zenodo.5728342). The scGTM Python package is available at https://github.com/ElvisCuiHan/scGTM.

6 Acknowledgements

The authors would like to thank Dr. Wanxin Wang for providing the data in Wang et al. (2020). The authors also appreciate the comments and feedback from the members of the Junction of Statistics and Biology at UCLA (http://jsb.ucla.edu).

7 Funding

This work was supported by National Science Foundation DBI-1846216 and DMS-2113754, NIH/NIGMS R01GM120507 and R35GM140888, Johnson and Johnson WiSTEM2D Award, Sloan Research Fellowship, and UCLA David Geffen School of Medicine W.M. Keck Foundation Junior Faculty Award (to J.J.L.).

References

Bacher, R., Leng, N., Chu, L.-F., Ni, Z., Thomson, J. A., Kendziorski, C., and Stewart, R. (2018). Trendy: segmented regression analysis of expression dynamics in high-throughput ordered profiling experiments. *BMC bioinformatics*, 19(1):1–10.

Bendall, S. C., Davis, K. L., Amir, E.-a. D., Tadmor, M. D., Simonds, E. F., Chen, T. J., Shenfeld, D. K., Nolan, G. P., and Pe'er, D. (2014). Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development. *Cell*, 157(3):714–725.

Bratton, D. and Kennedy, J. (2007). Defining a standard for particle swarm optimization. In *2007 IEEE swarm intelligence symposium*, pages 120–127. IEEE.

Campbell, H. (2021). The consequences of checking for zero-inflation and overdispersion in the analysis of count data. *Methods in Ecology and Evolution*, 12(4):665–680.

Campbell, K. R. and Yau, C. (2017). switchde: inference of switch-like differential expression along single-cell trajectories. *Bioinformatics*, 33(8):1241–1242.

Cao, J., Spielmann, M., Qiu, X., Huang, X., Ibrahim, D. M., Hill, A. J., Zhang, F., Mundlos, S., Christiansen, L., Steemers, F. J., et al. (2019). The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502.

Chechik, G. and Koller, D. (2009). Timing of gene expression responses to environmental changes. *Journal of Computational Biology*, 16(2):279–290.

Fischer, D. S., Theis, F. J., and Yosef, N. (2018). Impulse model-based differential expression analysis of time course sequencing data. *Nucleic acids research*, 46(20):e119–e119.

Ji, Z. and Ji, H. (2016). Tscan: Pseudo-time reconstruction and evaluation in single-cell rna-seq analysis. *Nucleic acids research*, 44(13):e117–e117.

Korani, W. and Mouhoub, M. (2021). Review on nature-inspired algorithms. *SN Operations Research Forum*, 2(36):https://doi.org/10.1007/s43069–021–00068–x.

Magwene, P. M., Lizardi, P., and Kim, J. (2003). Reconstructing the temporal ordering of biological samples using microarray data. *Bioinformatics*, 19(7):842–850.

Mondal, P. K., Saha, U. S., and Mukhopadhyay, I. (2021). Pseudoga: cell pseudotime reconstruction based on genetic algorithm. *Nucleic Acids Research*.

Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H. A., and Trapnell, C. (2017). Reversed graph embedding resolves complex single-cell trajectories. *Nature methods*, 14(10):979–982.

Ren, X. and Kuan, P.-F. (2020). Negative binomial additive model for rna-seq data analysis. *BMC bioinformatics*, 21(1):1–15.

Sander, J., Schultze, J. L., and Yosef, N. (2017). Impulsede: detection of differentially expressed genes in time series data using impulse models. *Bioinformatics*, 33(5):757–759.

Shalek, A. K., Satija, R., Shuga, J., Trombetta, J. J., Gennert, D., Lu, D., Chen, P., Gertner, R. S., Gaublomme, J. T., Yosef, N., et al. (2014). Single-cell rna-seq reveals dynamic paracrine control of cellular variation. *Nature*, 510(7505):363–369.

Shin, J., Berg, D. A., Zhu, Y., Shin, J. Y., Song, J., Bonaguidi, M. A., Enikolopov, G., Nauen, D. W., Christian, K. M., Ming, G.-l., et al. (2015). Single-cell rna-seq with waterfall reveals molecular cascades underlying adult neurogenesis. *Cell stem cell*, 17(3):360–372.

Silverman, J. D., Roche, K., Mukherjee, S., and David, L. A. (2020). Naught all zeros in sequence count data are the same. *Computational and structural biotechnology journal*, 18:2789.

Song, D. and Li, J. J. (2021). Pseudotimede: inference of differential gene expression along cell pseudotime with well-calibrated p-values from single-cell rna sequencing data. *Genome biology*, 22(1):1–25.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G., and Davis, R. W. (2005). Significance analysis of time course microarray experiments. *Proceedings of the National Academy of Sciences*, 102(36):12837–12842.

Street, K., Risso, D., Fletcher, R. B., Das, D., Ngai, J., Yosef, N., Purdom, E., and Dudoit, S. (2018). Slingshot: cell lineage and pseudotime inference for single-cell transcriptomics. *BMC genomics*, 19(1):1–16.

Trapnell, C., Cacchiarelli, D., Grimsby, J., Pokharel, P., Li, S., Morse, M., Lennon, N. J., Livak, K. J., Mikkelsen, T. S., and Rinn, J. L. (2014). The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386.

Van den Berge, K., De Bezieux, H. R., Street, K., Saelens, W., Cannoodt, R., Saeys, Y., Dudoit, S., and Clement, L. (2020). Trajectory-based differential expression analysis for single-cell sequencing data. *Nature communications*, 11(1):1–13.

Van Loan, C. F. and Golub, G. (1996). Matrix computations (johns hopkins studies in mathematical sciences).

Wang, W., Vilella, F., Alama, P., Moreno, I., Mignardi, M., Isakova, A., Pan, W., Simon, C., and Quake, S. R. (2020). Single-cell transcriptomic atlas of the human endometrium during the menstrual cycle. *Nature Medicine*, 26(10):1644–1653.

Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics: The official journal of the International Environmetrics Society*, 16(3):275–289.

Whitacre, J. M. (2011a). Recent trends indicate rapid growth of nature-inspired optimization in academia and industry. *Computing*, 93:121–133.

Whitacre, J. M. (2011b). Survival of the flexible: Explaining the recent dominance of nature-inspired optimization within a rapidly evolving world. *Computing*, 93:135–146.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(1):3–36.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. CRC press.

Yang, X.-S. (2017). *Nature-inspired algorithms and applied optimization*, volume 744. Springer.