Large scale maximum average power multiple inference on time-course count data with application to RNA-seq analysis

Meng Cao1 | Wen Zhou1 | F. Jay Breidt1 | Graham Peers2

1Department of Statistics, Colorado State University, Fort Collins, Colorado
2Department of Biology, Colorado State University, Fort Collins, Colorado

Correspondence
Wen Zhou, Department of Statistics, Colorado State University, Fort Collins, CO 80523.
Email: riczw@stat.colostate.edu

Funding information
Department of Energy, Grant/Award Number: DE-SC0018344; National Science Foundation, Grant/Award Number: IIS-1545994

Abstract
Experiments that longitudinally collect RNA sequencing (RNA-seq) data can provide transformative insights in biology research by revealing the dynamic patterns of genes. Such experiments create a great demand for new analytic approaches to identify differentially expressed (DE) genes based on large-scale time-course count data. Existing methods, however, are suboptimal with respect to power and may lack theoretical justification. Furthermore, most existing tests are designed to distinguish among conditions based on overall differential patterns across time, though in practice, a variety of composite hypotheses are of more scientific interest. Finally, some current methods may fail to control the false discovery rate. In this paper, we propose a new model and testing procedure to address the above issues simultaneously. Specifically, conditional on a latent Gaussian mixture with evolving means, we model the data by negative binomial distributions. Motivated by Storey (2007) and Hwang and Liu (2010), we introduce a general testing framework based on the proposed model and show that the proposed test enjoys the optimality property of maximum average power. The test allows not only identification of traditional DE genes but also testing of a variety of composite hypotheses of biological interest. We establish the identifiability of the proposed model, implement the proposed method via efficient algorithms, and demonstrate its good performance via simulation studies. The procedure reveals interesting biological insights, when applied to data from an experiment that examines the effect of varying light environments on the fundamental physiology of the marine diatom Phaeodactylum tricornutum.

KEYWORDS
false discovery rate control, Gaussian mixture, latent negative binomial model, maximum average power, RNA-seq experiments, time course data of counts

1 | INTRODUCTION

Studying transcriptomes through the sequencing of RNA (RNA-seq) has revolutionized biology and medical science. RNA-seq experiments have been employed to detect allele-specific expression and novel biomarkers, to assist medical prognosis, and to explore how global expression profiles alter in different biological environments. RNA-seq is affordable and allows for more samples; in particular, longitudinal studies are feasible, and can reveal dynamical biological patterns for thousands of genes simultaneously.
As with traditional RNA-seq analysis (Anders and Huber, 2010; Love et al., 2014; Robinson et al., 2010), the major goal of time-course RNA-seq is to detect differentially expressed (DE) genes across treatments. Genes differentially expressed over time are of particular interest. An early and popular analysis for time-course RNA-seq data is to model logarithmic fold change as a function of time points as categorical factors (Love et al., 2014; Robinson et al., 2010). This approach does not account for possible smoothness in the mean dynamics and may fail to identify DE genes with complex temporal profiles (Sun et al., 2016). An extension of the original model for time-course data in microarray experiments is the negative binomial employed by maSigPro-GLM, which accounts for temporal structure by a polynomial in time (Nueda et al., 2014). Similarly, splineTC uses cubic splines to model time-course transcriptome data and applies empirical Bayes moderated F-statistics for DE analysis (Michna et al., 2016). Oh et al. (2013) accounts for temporal structure with a hidden Markov model. Äijö et al. (2014) introduces an MCMC sampling procedure to identify temporally DE genes. The one-sample problem for time-course RNA-seq data, which focuses on identifying genes or gene sets with significant temporal dynamics, has also been studied (Leng et al., 2015; Agniel and Hejblum, 2017). Based on a family of flexible parametric functions, ImpulseDE2 performs both one-sample and two-sample analyses on time-course RNA-seq data (Fischer et al., 2018). Using a negative binomial mixed-effect model (NBMM), Sun et al. (2016) analyze time-course read counts of genes at the exon level and identify DE genes using the Kullback-Leibler distance ratio. They also discuss the importance of testing a variety of composite hypotheses other than the overall temporal pattern; namely, the nonparallel differentially expressed (NPDE) and parallel differentially expressed (PDE) genes, which are modeled through time by treatment interactions. Other methods for analyzing time-course RNA-seq data include Heinonen et al. (2015), Luo et al. (2017), Topa and Honkela (2018), and see Spies et al. (2019) for a review.

The aforementioned methods for analyzing time-course RNA-seq data may suffer from low power in practice. Though some approaches, such as the local regression model employed by Anders and Huber (2010) to estimate the dispersion parameters, adopt the idea of borrowing information across genes from the traditional RNA-seq analysis, the improvement in power is not completely theoretically justified and is unsatisfactory in practice. To illustrate, we considered the fission yeast data set used by Love et al. (2014) to show the performance of DESeq2 in analyzing time-course RNA-seq data. Besides DESeq2, we also fitted the expression of each gene to a negative binomial model and identified DE genes using the resulting likelihood ratio statistics and the standard procedure by Benjamini and Hochberg (1995). We refer to this procedure as LRT. With the nominal false discovery rate (FDR) controlled at 0.05, our proposed method identified 128 temporally DE genes out of 7039 genes, while DESeq2 identified 85 DE genes and the LRT only identified 27 DE genes. See Web Appendix D for further discussion of this example. In addition to the lack of theoretical guarantees on the power to detect DE genes using time-course RNA-seq data, the FDR control of existing methods is not well studied. Simulation and empirical studies like Kvam et al. (2012), Law et al. (2014), Chu et al. (2015), and Van den Berge et al. (2017) show that both DESeq and edgeR are conservative in some cases while liberal in others for traditional RNA-seq analysis as well as for analyzing time-course RNA-seq data (Sun et al., 2016; Nguyen, 2018). Our comprehensive simulations support the same conclusions.

Furthermore, most existing methods, such as maSigPro-GLM, splineTC, DEseq2, and edgeR, are mainly designed for overall differential patterns across time among conditions. But other differential expression profiles, such as PDE and NPDE genes (Sun et al., 2016), can provide a more subtle and accurate characterization of underlying biological processes. Section 5 describes analysis of time-course RNA-seq data collected from a novel light physiology experiment (Sindt et al., 2018) on the marine diatom Phaeodactylum tricornutum (P.t.). The experiment investigates the gene transcriptional response of P.t. during the process of photoaclimation to better understand how photosynthetics are regulated during the shifts of light conditions from low to high levels. Using our proposal, we detected more DE genes than existing methods and identified different types of DE profiles, as displayed in Figure 1. Specifically, gene Phatr3_J49108 primarily exhibits a relative overall mean shift on the expression between the high and low light levels, gene Phatr3_EG01131 displays altered temporal patterns between these two light levels while the overall mean levels remain unchanged, and gene Phatr3_J34003 possesses a more complex DE profile.

To address the above challenges systematically, motivated by Hwang and Liu (2010), Storey (2007), and Si and Liu (2013), we develop an optimal test in Section 2 for time-course RNA-seq data analysis: the test achieves maximum power averaged across all genes for which null hypotheses are false, while controlling the FDR. Use of the test relies on a model, which we develop in Section 3. Read counts for a gene at each time point have negative binomial distributions, with the mean temporal profile generated by a latent mixture Gaussian process. By modeling different types of DE profiles with mixture components, we are able to draw inference
on a variety of composite hypotheses other than the simple overall DE hypothesis. In addition, the model naturally adapts to non-equally spaced time points across conditions. We carefully establish the identifiability of the proposed mixture model (Web Appendix A) and implement the proposed test by developing an efficient algorithm to estimate the model parameters, using the gradient expectation-maximization (EM) algorithm and quasi-Monte Carlo integration (Web Appendix B). We perform comprehensive simulation studies in Section 4 and Web Appendix C to demonstrate the advantages of our method in comparison to existing methods. In Section 5, we analyze the Phaeodactylum light data using our method. Some biologically interesting genes and critical pathways, which are potentially related to photosynthesis regulation, are discovered. Discussion follows in Section 6.

2 A MAXIMUM AVERAGE POWER TEST

Let \( Y_g \) denote the vector (combined across all experimental conditions, time points, and replicates) of read counts for gene \( g \) (\( g = 1, \ldots, G \)) from an RNA-seq experiment. Assume that \( Y_g \) follows a distribution parameterized by generic \( \eta_g \) and \( \tau_g \), where \( \eta_g \) models the potential differential expression across conditions. Specifically, considering \( \Delta_0 \) and \( \Delta_1 \) as the null and alternative sets for \( \eta_g \), we identify DE genes by testing, for each \( g \),

\[
H^g_0: \eta_g \in \Delta_0 \text{ vs } H^g_1: \eta_g \in \Delta_1. \tag{1}
\]

Sets \( \Delta_0 \), \( \Delta_1 \) can be defined flexibly to reflect a variety of biological questions of interest. Without loss of generality, suppose that the first \( G_0 \) null hypotheses are true while the others are false: with respect to a common dominating measure \( \nu(\cdot) \), the densities \( f(\cdot|\eta_g, \tau_g) \) are null for \( g = 1, 2, \ldots, G_0 \) and alternative for \( g = G_0 + 1, G_0 + 2, \ldots, G \). Employing these densities and Neyman-Pearson arguments, Storey (2007) developed an optimal discovery procedure (ODP) to identify DE genes, in which null hypothesis \( H^g_0 \) is rejected if and only if \( S(Y_g) \leq \lambda \) where \( S(Y) = \sum_{g=1}^{G_0} f(Y|\eta_g, \tau_g) \sum_{g=G_0+1}^{G} f(Y|\eta_g, \tau_g) \). Let the expected false and true positives be \( \text{EFP}(\Gamma) = \sum_{g=1}^{G_0} \int f(Y|\eta_g, \tau_g) d\nu(Y) \) and \( \text{ETP}(\Gamma) = \sum_{g=G_0+1}^{G} \int f(Y|\eta_g, \tau_g) d\nu(Y) \), where \( \Gamma = \{ Y_g : S(Y_g) \leq \lambda \} \) is the significance region with respect to \( \lambda \). Storey (2007) shows that the ODP achieves maximum EFP among all procedures controlling the EFP. However, ODP implementation relies on estimates of \( G_0 \) and identification of the nulls in advance, which may be challenging for time-course RNA-seq data. Furthermore, ODP needs \( G^2 \) likelihood evaluations, which is computationally demanding.

To circumvent these challenges with ODP, we adopt the idea of borrowing information across genes, which is commonly used in traditional DE analysis to improve power (Cui et al., 2005; Hwang and Liu, 2010). Specifically, consider \( (\eta_g, \tau_g) \)'s to be independent random vectors with finite second moments from a common probability distribution \( \mu_0 \) if \( H^g_0 \) is true and \( \mu_1 \) otherwise. Assume that \( G_0 \to \infty, G - G_0 \to \infty \) with \( G_0/G \to p_0 < 1 \). Then

\[
\text{EFP}(\Gamma)/G_0 \overset{P}{\to} \int \int f(Y|\eta, \tau) d\nu(Y) d\mu_0(\eta, \tau)
\]

\[= \mathbb{P}(Y \in \Gamma|\Delta_0)\]

and

\[
\text{ETP}(\Gamma)/(G - G_0) \overset{P}{\to} \int \int f(Y|\eta, \tau) d\nu(Y) d\mu_1(\eta, \tau)
\]

\[= \mathbb{P}(Y \in \Gamma|\Delta_1),\]
which are referred as the average type I error and the average power by Hwang and Liu (2010).

Analogous to the ODP, we propose a test statistic

\[
\delta^*_\text{MAP}(Y) = p_0 \int \int f(Y|\eta, \tau) d\mu_0(\eta, \tau) \left((1 - p_0) \int \int f(Y|\eta, \tau) d\mu_1(\eta, \tau)\right)^{-1}
\]

for (1) and \( H_1^\lambda \) is rejected if and only if \( \delta^*_\text{MAP}(Y) \leq \lambda \) for some \( 0 \leq \lambda < \infty \). Statistic \( \delta^*_\text{MAP}(Y) \) mimics \( S(Y) \) in the ODP and maximizes the average power while controlling the average type I error, and is also equivalent to the maximum average power testing statistic introduced by Hwang and Liu (2010) and Si and Liu (2013).

Furthermore, assume that \((\eta_\tau, \tau_\eta)\) are independent and identically distributed from a common mixture \( \pi(\eta, \tau) = p_0 \pi_0(\eta, \tau) + (1-p_0) \pi_1(\eta, \tau) \). Let \( \Theta = \Delta_0 \cup \Delta_1 \) and \( \Omega \) denote the parameter spaces for \( \eta \) and \( \tau \), respectively. Assume that \( \pi_0(\eta, \tau) \) has support \( \Delta_0 \times \Omega \) and \( \pi_1(\eta, \tau) \) has support \( \Delta_1 \times \Omega \), and \( \Delta_0 \) is zero-measure under \( \pi_1(\eta, \tau) \) and \( \Delta_1 \) is zero-measure under \( \pi_0(\eta, \tau) \). Then, equivalent to \( \delta^*_\text{MAP}(Y) \), statistic

\[
\delta_{\text{MAP}}(Y) = \left\{ \int_{\Delta_0 \times \Omega} f(Y|\eta, \tau) d\mu(\eta, \tau) \right\} \times \left\{ \int_{\Theta \times \Omega} f(Y|\eta, \tau) d\mu(\eta, \tau) \right\}^{-1}
\]

(2)

provides a natural way to estimate FDR. Theorem 3 in Hwang and Liu (2010) states that among all tests controlling the FDR, a Neyman-Pearson test maximizes the average power. Hence, tests based on \( \delta^*_\text{MAP}(Y) \) or (2) also maximize the average power with FDR controlled.

Given \( \pi(\eta, \tau) \) above, \( \delta_{\text{MAP}}(Y) = P(Y|\Delta_0) P^{-1}(Y) = P(\Delta_0|Y) \) is the posterior probability of \( \Delta_0 \) given \( Y \). The (posterior) expected number of false positives for a given rejection region \( \Gamma \) is \( \sum_\gamma \delta_{\text{MAP}}(Y_\gamma) I(Y_\gamma \in \Gamma) = \sum_\gamma P(\Delta_0|Y_\gamma) I(Y_\gamma \in \Gamma) \). As is standard (Storey, 2007; Hwang and Liu, 2010), we estimate the FDR by the ratio of expected false positives and expected positives, which is \( \text{FDR}_\Gamma = \sum_\gamma \delta_{\text{MAP}}(Y_\gamma) I(Y_\gamma \in \Gamma)/\sum_\gamma I(Y_\gamma \in \Gamma) \). Therefore, the rejection region \( \Gamma(\alpha) = \{ Y_\gamma : \delta_{\text{MAP}}(Y_\gamma) \leq \lambda_\alpha \} \) is chosen with \( \text{FDR}_\Gamma \leq \alpha \) for a nominal level \( \alpha \) and it defines the maximum average power test for (1).

section 3 | METHODOLOGY

3.1 | Data model

Let \( Y_{gij}(t) \) denote the number of reads mapped to gene \( g \) from the \( j \)th replicate in treatment group \( i \) at time point \( t \), where \( g = 1, ..., G \), \( i = 1, 2 \), \( j = 1, ..., n_i \), and \( t = t_1, ..., t_T \) with integers \( T_i > 0 \). A widely applicable model for traditional RNA-seq analysis is the negative binomial (NB), which provides extra flexibility to model count data with large variations yet includes the popular Poisson model as a special case (Robinson and Smyth, 2008; Si and Liu, 2013). We therefore model \( Y_{gij}(t) \mid \lambda_{gij}(t) \) as independent \( \text{NB}(\lambda_{gij}(t), \phi_g) \) across genes \( g \), treatments \( i \), replicates \( j \), and time points \( t \), where \( \phi_g \) is the dispersion parameter and

\[
\lambda_{gij}(t) = \mathbb{E}\{Y_{gij}(t) \mid \lambda_{gij}(t)\} = S_0 \exp[\eta_{g11} + \mathbf{B}'(t) \eta_{g12} + \mathbf{B}'(t) \tau_g]
\]

with some \( q \)-dimensional orthogonal basis functions \( \mathbf{B}'(t) = (b_1(t), ..., b_q(t)) \), \( \eta_{g11} = \eta_{g11}^{[i=2]}, \eta_{g12} = \eta_{g12}^{[i=2]} \) with \( \eta_{g12}, \tau_g \in \mathbb{R}^q \). In (3), \( \mathbf{B}'(t) \tau_g \) models the mean temporal pattern for time-course expressions and \( S_0 \) are normalization factors (treated as known constants in practice) that adjust for varying sequencing depths and other technical effects across replicates.

Here, \( \eta_{g1} \) is the relative mean shift of gene expression between treatments and \( \eta_{g2} \) characterizes potential interactions between the temporal patterns and treatments. If \( \eta_{g2} \) is not zero, the relative differences between \( \lambda_{g1j}(t) \) and \( \lambda_{g2j}(t) \) display NPDE patterns, as introduced in Sun et al. (2016). FDR may hold when \( \eta_{g2} = 0 \), for which the relative differences between \( \lambda_{g1j}(t) \) and \( \lambda_{g2j}(t) \) do not vary over time (Sun et al., 2016).

Like the negative binomial model for time series of counts in Davis and Wu (2009), model (3) has a smooth mean temporal pattern \( \lambda_{gij}(t) \), conditional on a latent, zero-mean Gaussian process \( \mathcal{G}(t) \). By Mercer’s theorem (Williams and Rasmussen, 2006), \( \mathcal{G}(t) \) admits a series representation that converges almost surely. In practice, a finite expansion \( \sum_{k=1}^{K} Z_k b_k(t) \) approximates \( \mathcal{G}(t) \) well for even relatively small \( K \), where \( b_k(t) \)’s are the eigenfunctions corresponding to the covariance structure of \( \mathcal{G}(t) \) and eigenvalues \( Z_k \)’s are independent normal random variables, which motivate model (3). This link to general Gaussian processes reflects the flexibility of the proposed model. As \( \eta_{g11}, \eta_{g21}, \text{and} \tau_g \) mimic the eigenvalues \( Z_k \), they are similarly modeled using normal distributions in Section 3.2.

3.2 | Latent model and hypotheses

Under model (3), testing (1) reduces to testing \( H_0^\lambda: \eta_{g1}, \eta_{g2} \in \Delta_0 \) vs \( H_1^\lambda: \eta_{g1}, \eta_{g2} \in \Delta_1 \). To employ the proposed maximum average power test defined by \( \Gamma(\alpha) \) in Section 2, the derivation of \( \delta_{\text{MAP}}(Y) \) in (2) assumes information on \( \pi(\eta, \tau) \), where \( \eta = (\eta_\tau, \tau_\eta) \) for our model. Motivated from the discussion in Section 3.1, we consider \( (\eta_{g11}, \eta_{g21}, \tau_g) \) as independent normal random vectors.
Given the large number of tests to be performed, we balance the computational burden and model flexibility by specifying

$$
(\eta_{g1}, \eta_{g2}, \tau_g) \sim \pi(\eta, \tau) = \sum_{k=1}^{K} p_k \mathcal{N}(\mu^{(k)}, \Lambda^{(k)}),
$$

where $\mu^{(k)} = (\mu^{(k)}_1, 0, ..., 0)'$ and $\Lambda^{(k)} = \text{diag}(\sigma^2_{\eta(k)}$, $\sigma^2_{\mu(k)}$, $\sigma^2_{\tau})$ with diagonal matrices $\Sigma^{(k)}$ and $\Sigma$. Together, (3) and (4) are the proposed $K$-component latent Gaussian-Negative Binomial model. By varying the number of components $K$, proportions $p_k$, and parameters of the components, the proposed model possesses ample flexibility.

The number of components $K$ needs to be specified for estimation and inference with (3) and (4). Whereas some practical guidance is discussed in the literature (Khalili and Chen, 2007; Wang and Zhu, 2008; Si and Liu, 2013), we make the choice of $K$ from the important classification of DE genes into PDE and NPDE, as in Sun et al. (2016). PDE and NPDE genes have been observed in real time-course RNA-seq studies, such as the light physiology experiment discussed in Sections 1 and 5. To model these two types of differential expression, we consider $K = 4$ components: (a) genes without DE, $\eta_{g1} = 0$ and $\eta_{g2} = 0$, that is $\sigma^2_{\eta(g1)} = 0$ and the diagonals of $\Sigma^{(1)}$ are zeros, so that $(\eta_{g1}, \eta_{g2})$ have degenerate marginal distributions with point mass at zeros; similarly, (b) NPDE genes with only time-by-treatment interaction, $\eta_{g1} = 0$ and $\eta_{g2} \sim \mathcal{N}(0, \Sigma)$; (c) PDE genes, $\eta_{g1} \sim \mathcal{N}(\mu_1, \sigma^2_1)$ and $\eta_{g2} = 0$; and (d) NPDE genes with both treatment and time-by-treatment effects, $\eta_{g1} \sim \mathcal{N}(\mu_2, \sigma^2_2)$ and $\eta_{g2} \sim \mathcal{N}(0, \Sigma)$. In view of $\delta_{\text{MAP}}(Y)$ in (2), the null set $\Delta_0$ in (1) can then be specified to correspond to components of the mixture. For example,

$$
\Delta^\text{DE}_0 = \Delta^\text{Mean}_0 \cap \Delta^\text{NPDE}_0
$$

(5)

correspond to the first and second components, the first and third components, and only the first component in the above model, respectively. The alternatives to these nulls correspond to biologically interesting hypotheses: any mean shift (with or without NPDE) is the alternative to $\Delta^\text{Mean}_0$, any NPDE is the alternative to $\Delta^\text{NPDE}_0$, and any DE is the alternative to $\Delta^\text{DE}_0$.

Estimates of the unknown parameters $\mu_1, \sigma^2_1$, the diagonal entries of $\Sigma$, and the proportions $p_k$ are needed to conduct the test. In the Web Appendix B, we detail a quasi-Monte Carlo integration-assisted gradient EM algorithm for estimation. Specify $\Delta_0$ as in (5) and let $\beta = (\eta_{g1}, \eta_{g2}, \tau)$. To perform the test, we compute the plug-in estimate $\hat{\delta}_\text{MAP}(Y) = \int_{\Theta} \int_{\Theta} f(Y | x^T \beta, \Phi) \; d\delta(\beta) / \int_{\Theta} \int_{\Theta} f(Y | x^T \beta, \Phi) \; d\delta(\beta)$, where $f(-; \beta, \Phi)$ is the negative binomial density with $x_{(2p+1) \times (T_1 + T_2)} = [\{0_{T_1 \times (1)}^T \beta(t) \{0_{T_2 \times 1}^T \beta(t) \{ B_2(t) \}']$ for which $B_1(t) = (B_1(t), ..., B_1(t))_{t=1}^{T_1}$, estimated dispersion $\hat{\Phi}$, and $d\delta(\beta) = \sum_{k=1}^4 \pi_k \Phi(\beta | \mu^{(k)}, \Lambda^{(k)}) d\beta$ with multivariate normal density $\Phi$. Then, as in Section 2, $\lambda_\alpha$ is chosen, so that $-2 \hat{LR}_\alpha \leq \alpha$ with $\delta_{\text{MAP}}(Y)$ replaced by $\delta_{\text{MAP}}(Y)$ for each $g$. Using $\hat{\Gamma}(\alpha) = \{ Y_g; \delta_{\text{MAP}}(Y) \geq \lambda_\alpha \}$, we can identify DE genes for the hypothesis defined by $\Delta_0$ and related biological questions.

The proposed $K$-component latent Gaussian-Negative Binomial model in (3) and (4) is a finite mixture model for which identifiability must be established to draw meaningful statistical conclusions (Teicher, 1963; Yakowitz and Spragins, 1968; Chen, 1995). In the Web Appendix A, we show that our proposed model admits finite identifiability (Yakowitz and Spragins, 1968) provided the number of basis functions $q$ satisfies $2q + 1 \leq T_1 + T_2$.

## 4 | MONTE CARLO EVIDENCE

In this section, we examine the performance of the proposed method and other existing approaches, including DESeq2, edgeR (Lun et al., 2016), ImpulseDE2, maSigPro-GLM, and splineTC, using two extensive sets of simulation studies. We generate time-course count data according to (3), (4) and a variety of $\Delta_0$'s.

### 4.1 | Simulation Setting

For settings A and B, detailed below, we simulate 100 independent datasets of time-course count data from the proposed $K$-component latent Gaussian-Negative Binomial model in (3) and (4), with each dataset containing $G = 1000$ genes, two treatment groups, $T = 6$ or 10 time points, and $r=3$ or 6 replicates. Three types of basis functions with $q = 2$ or 3 are considered for estimating (3): the basis functions for the traditional Gaussian kernel (see Williams and Rasmussen, 2006 and the Web Appendix) denoted GA$_2$ and GA$_3$; the orthogonal Fourier basis functions, denoted FO$_2$ and FO$_3$; and the orthogonal polynomial bases, denoted PL$_2$ and PL$_3$. Also, we set $S_{ij} \equiv 1$.

**Setting A:** In (3), $q = 2$ and the true basis functions are PL$_2$. Parameters $\eta_{g1}, \eta_{g2},$ and $\tau$ are drawn from (4), where $\mu_1 = 2$ or 4, $\sigma^2_1 = 1$ or 2, $\Psi = \text{diag}(1, 1)$ and $\Sigma = \text{diag}(3, 2)$.

**Setting B:** In (3), $q = 3$ and the true basis functions are PL$_3$. Parameters $\eta_{g1}, \eta_{g2},$ and $\tau$ are drawn from (4),
where $\mu_1 = 2$ or $4$, $\sigma_1^2 = 1$ or $2$, $\Psi = \text{diag}\{1, 1, 1\}$ and $M = \text{diag}\{3, 2, 1\}$.

For each simulated dataset, we fit (3) with each of the six basis types and conduct the test accordingly. We also conduct the true test, using the true basis functions (e.g., PL$_2$ in setting A) and known parameters, and the oracle test, using the true basis functions but unknown parameters. ImpulseDE2 and splineTC use their own mean models; we use the true basis functions for all other competing methods. Dispersion parameters are estimated, by borrowing information across genes, if applicable, for competing methods except maSigPro-GLM, for which the true dispersion parameters are used, and splineTC, for which no dispersion is assumed. For our proposed method, the moment estimator for dispersion is developed and detailed in the Web Appendix. Extra simulation results for our method with different dispersion estimators yield similar results; see the Web Appendix (Figure S7 in Section C3 and related discussion). Empirical FDRs and powers averaged over 100 replications at nominal levels 0.01, 0.025, 0.05, 0.075, and 0.1 are reported for each method.

For each setting, 50% of genes are drawn from the null component of (4), 20% from NPDE with only time-by-treatment, 20% from PDE with no time-by-treatment; and 10% from NPDE with both treatment and time-by-treatment. In Web Appendix C, additional results for different proportions of DE genes are reported and show similar patterns (Figure S7).

## 4.2 Results

We consider the three null hypotheses in (5): any temporal DE is the alternative to $\Delta_{0}^\text{DE}$, relative mean shift is the alternative to $\Delta_{0}^\text{Mean}$, and NPDE is the alternative to $\Delta_{0}^\text{NPDE}$. For testing $\Delta_{0}^\text{DE}$, we compare the performance of our proposal to all five competitors. By specifying the corresponding contrasts or likelihood ratio statistics, edgeR and DESeq2 can test the other two composite hypotheses and are compared with our method as well.

Figures 2 and 3 display results for the proposed method using the GA$_3$ basis, true test, oracle test, and the five competing methods. Each point on the figures displays the empirical FDR and power of the corresponding method at a given nominal FDR level, which is marked as a vertical gray dashed line. The closer the point is to the corresponding vertical dashed line, the more the empirical and nominal FDR levels coincide. Results for the proposed method with other basis functions are qualitatively similar; see Figures S1 to S6 in the Web Appendix.

Most points on these plots are to the left of corresponding vertical dashed lines, which suggests that most methods have their empirical FDRs controlled. Compared to our proposed method, all other competitors are less powerful, particularly for a small number of replicates $r$ and lower $\mu$. The true test with known parameters and the oracle test without information on parameters perform the best overall, and are almost indistinguishable for most cases, reflecting the reliability of the estimation procedure described in the Web Appendix. Though the number and type of the basis functions are all misspecified, our proposal still provides satisfactory results, reflecting the expected flexibility and robustness of the method (Figures S1-S6 in the Web Appendix). In addition, as the number of replicates $r$ increases, the deviations between the proposed method with misspecified bases and the true test quickly diminish.

For setting A with PL$_2$ as the bases, all methods have the nominal FDRs controlled, as shown in Figure 2. For setting B with PL$_3$, Figure 3 shows that maSigPro-GLM may not provide a satisfactory control on the FDR for all levels as it essentially models the mean dynamics by a less flexible quadratic regression. It is interesting to observe that in both figures, compared to edgeR and ImpulseDE2, DESeq2, and splineTC have smaller empirical FDRs but comparable powers. Similar patterns also display in Tables 1 and 2.

Results for testing against $\Delta_{0}^\text{Mean}$ and $\Delta_{0}^\text{NPDE}$ under setting A are reported in Tables 1 and 2, respectively. Those for setting B are referred to the Tables S1 to S4 in the Web Appendix. Empirical FDRs are close to the nominal level for most tests for both hypotheses. The empirical powers for testing against $\Delta_{0}^\text{Mean}$ increase as $\mu_1$ increases from 2 to 4 for all methods, while they are not substantially improved by increasing $T$ and $r$. On the other hand, similar to testing the interactions in functional ANOVA (Gu, 2004; Sun et al., 2016), testing for NPDE is more challenging, and the power is expected to be lower. However, as displayed in Table 2 as well as in Tables S3 to S4 in the Web Appendix, the powers for testing NPDE genes of the proposed method increase in $T$ and $r$ for all bases and are unsurprisingly not influenced by the mean parameter $\mu_1$. Furthermore, the performance of the proposal is not affected much by the choice of the bases. In addition, our proposed method outperforms edgeR and DESeq2 in power, particularly for testing NPDE genes. These numerical results confirm the theoretical guidance in Section 2 on the optimality of the proposed method as well as its flexibility on testing a variety of composite and biologically interesting hypotheses.

## 5 Analysis of the Phaeodactylum Light Experiment

We apply our method to $P.t.$ transcriptomics data from the light experiment introduced in Section 1. Most algae and cyanobacteria, including $P.t.$, undergo some major changes
FIGURE 2 Empirical FDRs and powers for testing the overall temporal DE genes by the proposed method using GA3 basis (Δ), compared with those of the oracle test (○), the true test (♢), maSigPro-GLM (■), edgeR (●), splineTC (∇), ImpulseDE2 (⊠), and DESeq2 (×) for setting A. Each point on the figures displays the empirical FDR and power of the corresponding method at a given nominal FDR level, which is marked as a vertical gray dashed line. All plots are for T = 10. Results are based on 100 replications. DE, differentially expressed; FDR, false discovery rate.
FIGURE 3  Empirical FDRs and powers for testing the overall temporal DE genes by the proposed method using GA₃ basis (Δ), compared to those of the oracle test (○), the true test (◇), maSigPro-GLM (■), edgeR (●), splineTC (△), ImpulseDE2 (♦) and DESeq2 (×) for setting B. Each point on the figures displays the empirical FDR and power of the corresponding method at a given nominal FDR level, which is marked as a vertical gray dashed line. All plots are for T = 10. Results are based on 100 replications. DE, differentially expressed; FDR, false discovery rate.
in cell biochemistry via a process termed photoacclimation. Growth in high light environments leads to low photosynthetic pigment per cell, accumulation of fats and carbohydrates, and an upregulated oxidative stress response. Growth in low light environments leads to high pigmentation of cells to capture more light and major increase in structural lipids associated with the chloroplast (Falkowski and Raven, 2013).

Sindt et al. (2018) investigated the gene transcriptional response during photoacclimation, to better understand

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Comparison of empirical FDRs and powers for testing DE genes with relative mean shift by the proposed method with different bases, edgeR, and DESeq2 for setting A</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(T, r, σ²)</td>
</tr>
<tr>
<td>μ₁ = 2</td>
<td></td>
</tr>
<tr>
<td>GA₂</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>FO₂</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>FO₃</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>PL₃</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>Oracle</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>True</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>edgeR</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>DESeq2</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>μ₁ = 4</td>
<td></td>
</tr>
<tr>
<td>GA₂</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>GA₃</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>FO₂</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>FO₃</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>PL₃</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>Oracle</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>True</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>edgeR</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
<tr>
<td>DESeq2</td>
<td>FDR</td>
</tr>
<tr>
<td></td>
<td>Power</td>
</tr>
</tbody>
</table>

Note: In simulations, μ₁, T, r, and σ² are displayed in the table. The nominal FDR level is 0.05. The simulation is based on 100 replications.
regulation of photosynthetic and catabolic metabolisms during the shift of a day-night cycle. Cultures of the diatom *P.t.* were submitted to a single step change in light from excess light fluxes to low light fluxes that nearly limit the growth rate. Samples for transcriptomics were taken over a 24-hour period. After mapping to the genome, 12,319 candidate genes were to be analyzed from the two groups with three replicates within each. Data for both groups were collected at 0, 60, 240, and 360 minutes as well as 12 and 24 hours, while data for the
low light group were also collected at 20, 40, 90, and 120 minutes.

We first filtered and normalized the data following standard procedures, described in detail in the Web Appendix D. After filtering, 10,597 genes remain for further analysis.

For DE analysis, the nominal FDR level is set at 0.05, and we use polynomial basis functions with $q = 2$ for $B(t)$. Results for all basis functions discussed in Section 4 yield similar results; for example, the majority of DE genes (~91%) for the hypotheses of interest are identified by all six basis functions. We identify 3869 overall temporally differential expressed genes, within which 3771 genes have a relative mean shift between two groups (reject $\Delta_0^{\text{Mean}}$) and 98 are NPDE (reject $\Delta_0^{\text{NPDE}}$). Figure 4 displays the top 10 selected genes identified with both relative mean shift and NPDE. The top 10 genes selected with only relative mean shifts are displayed in Figure S8 in the Web Appendix.

As a large proportion of predicted gene models for P.t. still encodes genes of unknown function (Bowler et al., 2008), functions of the top 10 differentially regulated genes in Figure 4 are not completely documented. However, gene Phatr3_J47271 is an experimentally verified critical component of the photosynthetic pigment biosynthesis pathway, and our finding on its differential regulation supports the large dynamical change in pigmentation observed during photoacclimation (Falkowski and Raven, 2013). In addition, genes Phatr3_J50183, Phatr3_J49693, and Phatr3_EG01882 are all highly related to photoacclimation. From Figure 4, their temporal differential expression patterns are much more sophisticated than the simple mean shift targeted by the traditional RNA-seq analysis. The new tests have revealed potentially critical pathways for better understanding photoacclimation in general. Follow-up analysis via gene set enrichment for three critical photoacclimation-related pathways is discussed in the Web Appendix D.

6 | DISCUSSION

We have proposed an inferential procedure for time-course RNA-seq data that maximizes average power and controls FDR. Novel features of our approach include smooth and flexible modeling of the mean dynamic (3) within genes, conditional on a latent zero-mean Gaussian process; and flexible mixture modeling (4) of coefficients across genes. Taken together, the within-gene/across-gene model (our $K$-component latent Gaussian-Negative Binomial model) allows feasible estimation of unknown model parameters, natural borrowing of information across genes for both the mean and variance instead of
the dispersion only, and straightforward, one-step testing of general composite null hypotheses of great biological interest. By contrast, existing pipelines such as edgeR and DESeq2 have to rely on a two-step procedure for testing some composite hypotheses. Additional simulations in the Web Appendix (see Table S5) show the superior performance of our proposal over existing methods for testing these more general composite null hypotheses.

Though easy to implement, as suggested by a numerical study in the Web Appendix, our method incurs some computational cost on parameter estimation via gradient-EM and quasi-Monte Carlo integration, which can be partly offset by parallel computations in practice. When the true alternatives are not at all smooth, such as an abrupt change in mean for a single time point, our method may lose powers. This can be partly resolved by using the step function as basis in (3) as mentioned in Section D4 in the Web Appendix. In addition to the limitation on smoothness of the alternatives, the proposed method may have a slight inflation in empirical FDR when data have a large proportion of outliers that are not well modeled with the negative binomial distribution; however, our method is only modestly affected and still performs well when the proportion of outliers is small as demonstrated by additional simulations in the Web Appendix C.

Our test is numerically indistinguishable from the oracle and true tests, when the number of replicates \( r \) is relatively large, while the difference is more obvious for small \( r \). One reason for this is the estimation of gene-wise dispersion \( \phi \). We employ a moment estimate for \( \phi \) inducing extra variations for our test. It has been long recognized that the unstable estimation of the dispersion in the small RNA-seq experiments may result in poor control of FDR, while a good estimate of this parameter is challenging in practice (Robinson and Smyth, 2008; Anders and Huber, 2010; Van den Berge et al., 2017). Existing pipelines usually address this challenge by borrowing information across genes via an empirical Bayes approach. In the Web Appendix, additional simulations (see Figure S7) are reported for our test combined with different estimates on the dispersion, including the empirical dispersion estimator in Section B.1.2 (ED), the common dispersion estimator (CD), and the local regression approach used in DESeq2. Although the empirical powers of our test with the three estimates of dispersion are similar with minor differences, the FDR control of ED is slightly inflated when \( r = 3 \) and the DE proportion is small. When \( r \) increases, the performances among different approaches are comparable. This relatively robust performance with respect to different approaches can be possibly explained as follows. Compared to the traditional RNA-seq trials with only \( r \) replicates, dispersion estimation for time-course experiments may leverage more from \( rT \) samples collected at \( T \) time points. Also, edgeR is less conservative than DESeq2 in general, which is due to the robust empirical Bayes method employed in edgeR to estimate the prior degree of freedom for the weighted likelihood as it reduces the informativeness of the prior distribution for outlier genes (Chen et al., 2014; Phipson et al., 2016). In contrast, using a data-adaptive control on the shrinkage, DESeq2 controls the FDR better than edgeR for small \( r \). Employing the edgeR approach to estimate gene-wise dispersion may improve our method for large \( r \) in practice and is left to future study.

Though the proposed method focuses on two-sample problems, it is readily generalized to \( M \geq 3 \) treatments by extending the specification of \( \lambda_{ij}(t) \) in (3) and the mixture distribution on its coefficients in (4). A further extension is to allow the temporal dynamics to vary continuously with respect to some explanatory variable \( Z \), such as age or blood pressure. Extending the mean specification (3) to this continuous case is straightforward, but the resulting continuous mixture distributions on model coefficients lead to more involved questions of identifiability and estimation. We leave these as topics for future investigation.

ACKNOWLEDGMENTS

We thank the coeditor, the associate editor, and two referees for insightful comments that have substantially improved this article. This work is supported in part by grants from the National Science Foundation and the US Department of Energy.

ORCID

Wen Zhou http://orcid.org/0000-0002-7506-3669

REFERENCES


Biometrics website on Wiley Online Library. The R package, MAPTest, implements our proposed method and is publicly available at https://github.com/meca7653/MAPTest, where the illustrative numerical example is also included. Users can specify their own estimates of the normalization factors or dispersion parameters.

How to cite this article: Cao M, Zhou W, Breidt FJ, Peers G. Large scale maximum average power multiple inference on time-course count data with application to RNA-seq analysis. Biometrics. 2019;1–14. https://doi.org/10.1111/biom.13144