BIOINFORMATICS

Vol. 00 no. 00
Pages 1–10

Bayesian detection of non-sinusoidal periodic patterns in circadian expression data

Darya Chudova [a]*, Alexander Ihler [a], Kevin K Lin [b], Bogi Andersen [b], Padhraic Smyth [a]

[a] Department of Computer Science
[b] Departments of Medicine and Biological Chemistry, Sprague Hall, Room 206
University of California, Irvine, Irvine CA 92697

ABSTRACT

Motivation: Cyclical biological processes such as cell division and circadian regulation produce coordinated periodic expression of thousands of genes. Identification of such genes and their expression patterns is a crucial step in discovering underlying regulatory mechanisms. Existing computational methods are biased towards discovering genes that follow sine-wave patterns.

Results: We present an ANOVA periodicity detector and its Bayesian extension that can be used to discover periodic transcripts of arbitrary shapes from replicated gene expression profiles. The models are applicable when the profiles are collected at comparable time points for at least two cycles. We provide an empirical Bayes procedure for estimating parameters of the prior distributions and derive closed-form expressions for the posterior probability of periodicity, enabling efficient computation. The model is applied to two data sets profiling circadian regulation in murine liver and skeletal muscle, revealing a substantial number of previously undetected non-sinusoidal periodic transcripts in each. We also apply quantitative real-time PCR to several highly ranked non-sinusoidal transcripts in liver tissue found by the model, providing independent evidence of circadian regulation of these genes.

Availability : Matlab software for estimating prior distributions and performing inference is available for download from `ftp://ftp.ics.uci.edu/pub/dchudova/periodicity`.

Contact: dchudova@gmail.com

INTRODUCTION

Identifying periodic transcripts in large time-course gene expression experiments is an important step in studying diverse biological systems, including the cell cycle, hair growth cycle, mammary cycle, and circadian rhythms. The data from these studies are often characterized by a large number of genes with relatively coarse sampling in time (for example, a few time points per cycle) and only a few measurements at each time point. The objective is to identify or rank which of these genes are most likely to be periodically regulated. In this paper, we propose a simple probabilistic mixture model for identifying periodic expression in cyclic processes where cycle length is known a prior and expression levels can be profiled at comparable time points in multiple cycles[1]. Such data sets are generated, for example, in experiments profiling circadian regulation in peripheral tissues (see Storch et al. (2002); Rudic et al. (2005); Miller et al. (2007) among others).

Existing techniques for detecting periodic expression patterns fall into two major categories: time domain and frequency domain analyses. Typical frequency domain methods compute the spectrum of the average expression profile for each probe, and test the significance of the dominant frequency against a suitable null hypothesis such as uncorrelated noise. However, frequency domain analysis is most effective on long time series and is not well suited for short time-courses (Tai and Speed, 2007).

In time domain analysis, most methods rely on the identification of sinusoidal expression patterns (Straume, 2004; Andersson et al., 2006; Wijnen et al., 2006). These detectors are popular due to their simplicity and computational efficiency, but are not effective at finding periodic signals which violate the sinusoidal assumption. While this assumption can be appropriate for some data (such as the cell cycle), a significant number of profiles with non-sinusoidal shapes have been identified in the control of hair cycling (Lin et al., 2004) and in the circadian rhythms of *Drosophila* (Keegan et al., 2007). More general shapes could be modeled using, for example, B-spline representations (Luan and Li, 2004), but such approaches require a set of "guide genes" to define the possible shapes of periodic patterns, which in practice may be unavailable or incomplete.

In this paper we propose a general statistical framework for detecting periodic profiles from time-course microarray data by analyzing the similarity of observed profiles across the cycles. Using this framework, we identify a significant number of previously undetected circadially-regulated genes with non-sinusoidal profiles in peripheral mouse tissues. For example, Fig. 1 shows profiles of several probe sets that were among those ranked most likely to be periodically expressed (in the top 25 profiles) by our proposed approach but were ranked much lower by a more traditional sine-wave detection algorithm (Miller et al., 2007). Notably, two of these probe sets (*Nr1d1* and *Arntl*) correspond to well established clock-control genes. In addition, circadian regulation of *Cyp2a4* in liver has been established in Lavery et al. (1999), and *Mknk2* has been identified as circadially-regulated in liver in an independent

*to whom correspondence should be addressed

[1] In systems where it is only possible to profile a single synchronous cycle, more domain-specific methods are required for identifying periodic profiles (Lin et al., 2004; Rudolph et al., 2003).

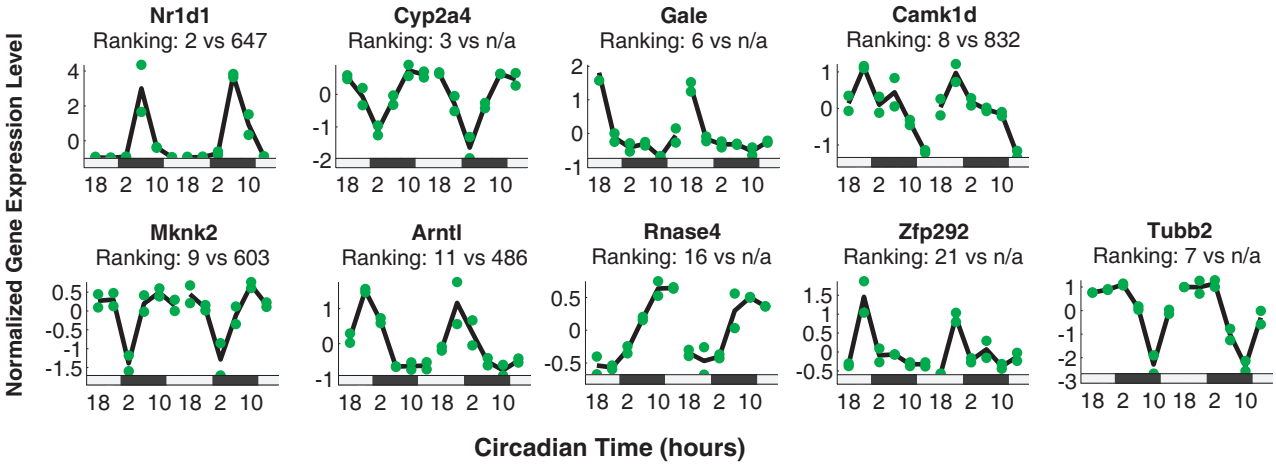

Fig. 1. Examples of non-sinusoidal periodic patterns in the circadian profiling of liver tissues. Shown are the profiles of 9 probe sets that are ranked among the top 25 probe sets by the proposed approach but ranked below 400 by a sine wave detector. Rank 'n/a' indicates ranking below the 848 published probe sets in Miller et al. (2007) based on the sine wave detector. The dots indicate individual replicate observations, and the line shows the empirical means at each time point. The measurements have been log-transformed and normalized to zero mean across time for each probe set. The x-axis shows circadian time, and the light/dark bands underneath the bar plot denote the light/dark experimental conditions.

microarray study by Oishi et al. (2003). Our quantitative PCR experiments validate circadian cycling for seven out of eight tested genes in this figure[2], demonstrating that these are likely true positives missed by previous analyses (see the Experimental Results section). Overall we detect significant numbers of non-sinusoidal patterns that were missed by the original analyses using existing detection algorithms.

The rest of the paper is organized as follows. In the next section, we describe our probabilistic model in detail and describe how it can be used to infer, for each probe set, the probability of its observed expression pattern being periodic. We also describe two simplified versions of the model, a (non-Bayesian) ANOVA test and a simplified Bayesian model which can be implemented using the Bioconductor *timecourse* package (Tai and Speed, 2007). We then provide experimental validation by analysing two data sets profiling circadian regulation in different peripheral tissues, and using independent experiments to confirm our findings. Finally, we discuss potential extensions of the model and present our conclusions.

METHODOLOGY

Our model for detecting periodicity is similar to existing methods for detecting differential expression. These methods typically assume that observed data can be described by a mixture distribution with two components: one component corresponds to genes that change their expression levels in response to changes in experimental conditions (*differentially expressed* genes), the other corresponds to genes that remain constant throughout the experiment (*background* genes). To model periodic phenomena, we include an additional third component that encodes *coordinated* expression across multiple cycles, see Fig. 2. Our task of identifying periodicity then reduces to a probabilistic inference problem: given the observed expression profiles, compute the posterior probability that a given probe set was generated by the periodic component.

[2] The annotation information for the *Tubb2* probe set was not available at the time of our experiments and so was not included in the PCR evaluation.

A Probabilistic Model for Periodicity

Consider a time course experiment that profiles expression of N probe sets over C cycles of known length. Each cycle is represented by the same grid of T time points, indexed from 1 to T. Profiling is typically done using multiple observations or replicates at a given time point (for example, 2 or 3) using a cross-sectional study design, i.e., all of the replicates at all of the time points originate from different biological subjects. We denote the number of replicate observations for probe set $i \in \{1 \dots N\}$ at time point $j \in \{1 \dots T\}$ of cycle $c \in \{1 \dots C\}$ by n_{ij}^c. Note that this number may be zero; for example, we may not make any observations at time j in some cycle c, in which case n_{ij}^c will be zero for all i. We use Y_{ijk}^c to denote the expression intensity value for a particular probe set i, time point j, and replicate k for cycle c, and let Y_i be the entire set of observations for probe set i. We assume that the intensity values Y_{ijk}^c have been estimated from raw data using a standard approach such as that of Wu et al. (2004), log-transformed and shifted to zero mean for each probe set's profile.

Our probabilistic model for expression, then, consists of three components: background (b), differentially expressed but aperiodic (d), and periodically expressed profiles (p). Let $Z_i \in \{b, d, p\}$ denote the component associated with probe set i. The forward or generative model is simple: to simulate an expression profile, one selects one of the three components according to their respective probabilities $[\pi_b, \pi_d, \pi_p]$, then samples a collection of observations according to the associated component model. Each of the three component models consists of a Normal/Inverse Gamma (NIG) prior distribution (Gelman et al., 1995) on the latent profile and additional Normal (i.e., Gaussian) noise on the observations. The components differ in the structure of latent profiles and in the parameters of their (NIG) model.

The NIG prior is a flexible and computationally convenient distribution commonly used as a prior model for latent expression levels and replicate variability (e.g., Smyth, 2004; Tai and Speed,

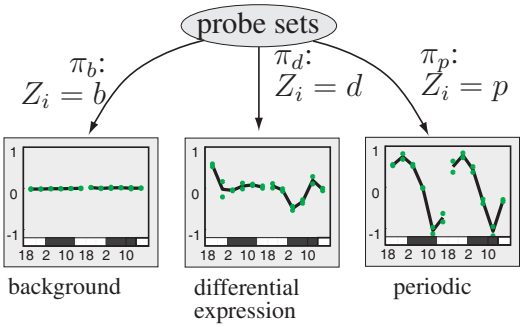

Fig. 2. We model the data using a mixture of three components for background, differentially, and periodically expressed profiles, with probabilities $[\pi_b, \pi_d, \pi_p]$ respectively.

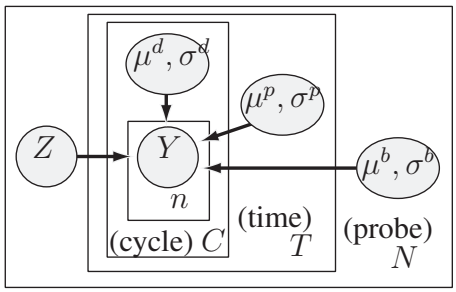

Fig. 3. A graphical model describing the observed profiles Y and latent (unobserved) variables Z (component identity) and $\{\mu, \sigma\}$ for each component using plate notation.

2006, 2007). In general, scalar variables (μ, σ) are distributed as NIG with parameters $(\nu, \eta\,;\, a, b)$ if

$$P(\mu, \sigma) = \mathrm{N}(\mu \,|\, \nu, \sigma/\eta)\; \Gamma^{-1}(\sigma \,|\, a, b)$$

where $\mathrm{N}(x \,|\, \nu, s)$ denotes a Gaussian distribution with mean ν and variance s and $\Gamma^{-1}(x \,|\, a, b)$ denotes an inverse Gamma distribution with a degrees of freedom and scale parameter b, evaluated at x.

Note that in what follows, we refer to three types of unknown quantities. The first are the *prior parameters*, denoted Θ, which we determine via an empirical Bayesian procedure (details later) and are subsequently treated as known and fixed. The other two types are probe set specific hidden variables: the latent profiles (consisting of a mean and variance) for each component, and the component identity Z_i, indicating from which component the data were generated.

Components of the Mixture Model

Our model is shown as a graphical model using plate notation in Fig. 3 (Jordan, 2004). The plates, or rectangles, are used to group together variables that are repeated in the model as many times as shown in the right-bottom corner of the plate. For example, the outermost plate corresponds to a single probe set and all variables within it are repeated N times, once for each of the N probe sets (indexed by i in the text). This structure implies conditional independence of the probe sets given fixed prior parameters Θ, since there are no shared dependencies. While in reality periodic or differentially expressed genes may share similar profiles, the assumption of conditional independence of probe sets is a reasonable first-order approximation and is computationally convenient. More realistic alternatives to this assumption are briefly described in the Discussion section.

The Background Component Model We model "background" probe sets as having a constant expression over the experiment (denoted by μ_i^b), with small fluctuations in the actual observations due to technical errors (variance σ_i^b). These variables are given a NIG prior shared by all background probe sets and parameterized by four scalars $\Theta_b = \{\nu^b, \eta^b\,;\, a^b, b^b\}$.

Since μ_i^b and σ_i^b are shared across time, they are shown outside the cycle and time plates in Fig. 3. The observations Y_i are modeled as independent samples from a Gaussian distribution with mean and variance (μ_i^b, σ_i^b):

$$P(Y_i|\mu_i^b, \sigma_i^b, Z_i = b) = \prod_{c=1}^{C} \prod_{j=1}^{T} \prod_{k=1}^{n_{ij}^c} \mathrm{N}(Y_{ijk}^c \,|\, \mu_i^b, \sigma_i^b)$$

where the products are over the C cycles, the T time-points within cycle c, and the observed replicate expression measurements for time point t in cycle c, respectively.

The Differentially Expressed Component Model For differentially expressed genes, the true expression levels vary as a function of time. Accordingly, we let μ_i^d and σ_i^d be $(C \times T)$-dimensional vectors characterizing the expression value and replicate variance at each of the time points. These variables are shown inside the cycle and time plates in Fig. 3. We let the expression at each time point vary independently from the other time points, so that the prior distribution for this component is defined by four $(C \times T)$-dimensional parameters, $\Theta_d = \{\nu^d, \eta^d\,;\, a^d, b^d\}$:

$$P(\mu_i^d, \sigma_i^d|\Theta_d) = \prod_{c=1}^{C} \prod_{j=1}^{T} \mathrm{NIG}(\mu_{ij}^{d,c}, \sigma_{ij}^{d,c} \,|\, \nu_j^{d,c}, \eta_j^{d,c}, a_j^{d,c}, b_j^{d,c})$$

The independence assumption works well for relatively sparse sampling of the time axis, a common situation with expression data measurements in practice[3]. Since the replicates are assumed to originate from different experimental units (cross-sectional design), we model observations as being independent given (μ_i^d, σ_i^d) :

$$P(Y_i|\mu_i^d, \sigma_i^d, Z_i = d) = \prod_{c=1}^{C} \prod_{j=1}^{T} \prod_{k=1}^{n_{ij}^c} \mathrm{N}(Y_{ijk}^c \,|\, \mu_{ij}^{d,c}, \sigma_{ij}^{d,c})$$

The Periodic Component Model The periodic component assumes repeated expression of the same pattern across multiple cycles. The true, latent expression level at a single time point gives rise to the observed intensities in cycles 1 through C. We let μ_i^p and σ_i^p be T-dimensional variables encoding expression levels and replicate variability in the "ideal" cycle. These variables are shown inside the

[3] For more densely-sampled data, one could extend this approach by adding dependency between the means, for example by introducing covariance structure into the prior.

time plate but outside the cycle plate in Fig. 3. Assuming sparsity of the time grid, we use independent NIG priors for each time point:

$$P(\mu_i^p, \sigma_i^p | \Theta_p) = \prod_{j=1}^{T} \mathrm{NIG}(\mu_{ij}^p, \sigma_{ij}^p \mid \nu_j^p, \eta_j^p \mid a_j^p, b_j^p)$$

The periodic component is parameterized by four T-dimensional parameters $\Theta_p = \{\nu^p, \eta^p\ ;\ a^p, b^p\}$. Due to the cross-sectional study design we again assume conditional independence of observations:

$$P(Y_i | \mu_i^p, \sigma_i^p, Z_i = p) = \prod_{c=1}^{C} \prod_{j=1}^{T} \prod_{k=1}^{n_{ij}^c} \mathrm{N}(Y_{ijk}^c \mid \mu_{ij}^p, \sigma_{ij}^p)$$

The complete set of prior parameters Θ includes the prior component probabilities π_z (corresponding to the relative frequencies of background, differentially expressed, and periodic probe sets), and prior parameters for each of the component models: $\Theta = \{(\pi_z, \Theta_z), z \in \{b, d, p\}\}$.

Inference

Given the model, we can detect periodic expression by computing the posterior probability of the periodic component $p(Z_i = p | Y_i, \Theta)$ conditioned on the prior parameters Θ and the observed profile Y_i:

$$P(Z_i = p | Y_i, \Theta) = \frac{\pi_p P(Y_i | \Theta_p, Z_i = p)}{\sum_z \pi_z P(Y_i | \Theta_z, Z_i = z)} \qquad (1)$$

Each of the three marginal likelihood terms in the denominator, for $z \in \{b, d, p\}$, is computed by averaging over our uncertainty about the latent profiles μ and replicate variances σ. Since the priors for (μ, σ) are conjugate to the Gaussian likelihood of Y_i, the marginal likelihood can be computed in closed form as shown in Appendix 1.

An ANOVA Periodicity Detector

Our Gaussian mixture model, and its resulting inferential test for periodicity, is quite close to a simplified, non-Bayesian test based on analysis of variance (ANOVA). We can construct a one-way ANOVA test for periodicity by dividing the data into groups, or factor levels, by their associated time point regardless of cycle number, so that all replicates Y_{ijk}^c for $c = 1 \ldots C$ and $k = 1 \ldots n_{ij}^c$ fall into the same group. We then test whether the data support separation into these groups, i.e., whether the amount of variation between groups is significantly larger than the variation found within the groups. High values of the ratio of these quantities indicates that most of the variability in observations can be explained using a time-dependent, cycle-independent profile, i.e., that the profile appears periodic.

Like our Bayesian test, the ANOVA test has a number of desirable properties; for example, it considers both similarity among the raw replicate observations and the magnitude of overall changes (the average profile) over time. Both quantities are important – replicate variability is useful in assessing similarity among cycles relative to inherent biological variability, while the magnitude of change helps differentiate signals from random noise. The ANOVA test is also easy to implement using any standard statistical package.

However, there are also a number of disadvantages to the ANOVA test. For it to work as expected, we require a balanced experiment design in which the number of replicates is unchanged over time ($n_{ij}^c = n_i$). It implicitly assumes that the data are Gaussian, with equal variance among the groups (i.e., over time). One can view our model as a Bayesian extension of the ANOVA test: both discriminate based on the amount of variance in the data under models of different complexity, but the Bayesian model relaxes the assumption of equal variances over time and adds a prior term which regularizes the variance estimates when there are few data. Moreover, it can handle a variable number of replicates at each time – an important feature when the data may suffer from missing observations or insufficient replication at certain time points.

Estimating Parameters of the Prior Distribution

Following Newton et al. (2004), Smyth (2004), and Tai and Speed (2006, 2007), we develop an empirical Bayes procedure to determine the prior parameters Θ for our model. We first determine a tentative assignment of probe sets to each component, then use this assignment to find approximate maximum likelihood estimates of the location scale η and parameters of the inverse Gamma distribution (a, b); we set the location mean ν to 0 in all three components.

To find a tentative initial assignment of probe sets for estimating prior parameters, we run one-way ANOVA detectors of differential expression and periodicity. Probe sets that vary significantly over time according to the first test (p-value less than 0.01) are used to define parameters of the component for differential expression, while probe sets which fail this test (p-value above 0.1) are used to define the parameters of the background component. Similarly, we use the described ANOVA periodicity detector to identify probe sets for estimating the prior parameters of the periodic component. Choosing those probe sets with p-value less than 0.001 results in a number of probe sets similar to that previously identified in the literature (Miller et al., 2007). The prior component probabilities π are set to the fraction of probe sets that were assigned to each component using this procedure.

The other parameters are then determined using a greedy maximum likelihood method. Briefly, the inverse Gamma parameters (a, b) are chosen to maximize the likelihood of the observed sums of squared deviations under an F-distribution. After the parameters a and b are fixed, η is chosen to maximize the likelihood of the observations Y under a Normal-inverse Gamma prior. While the resulting estimates do not necessarily maximize the joint likelihood with respect to η, a, and b, due to the two-step nature of the procedure, these estimates are fast to compute and we have found them to work well in practice. More details on this estimation process can be found in Appendices 2 and 3.

Implementation via Tai and Speed's Framework

We note in passing that a Bayesian model similar to our own can be implemented using the framework of Tai and Speed (2007) and the *timecourse* package in Bioconductor. Like the ANOVA test, we use only two hypotheses: periodic versus background, and again group together all replicates from the same relative time point regardless of cycle. We then apply the test from Tai and Speed (2007) for analysis of differential expression in one-sample cross-sectional experiments to the grouped data. Any aperiodic yet differentially expressed signals should have high "in-group" variation due to combining data across cycles, causing only periodic profiles to be ranked highly.

We believe this technique is less intellectually satisfying than our three-component Bayesian model, since it groups two sets of apparently different behaviors (background and aperiodic differential

expression) under a single Gaussian model. However, in empirical comparisons the methods often behaved similarly, and both models provide useful alternatives to traditional analyses that rely on identifying sinusoidal expression changes.

EXPERIMENTAL RESULTS

In this section we demonstrate that our model can effectively identify both sinusoidal and non-sinusoidal periodic expression patterns in data sets profiling circadian expression in peripheral tissues, including the automatic discovery of genes which were not previously known to exhibit circadian patterns. It is widely believed that $5 - 10\%$ of transcribed genes in these tissues may be under circadian regulation (Storch et al., 2002), with some studies suggesting a higher proportion - up to 50% in murine liver (Ptitsyn et al., 2006). Different studies and computational methods are not consistent in identifying the exact set of such genes, with the exception of a few core clock control genes.

The data sets analyzed in this paper contain gene expression profiles of liver and skeletal muscle tissues in mice (Miller et al., 2007). The data are available through GEO repository, accession *GSE3751*. The microarray experiments used a custom-made Affymetrix platform with 33143 probe sets representing 20110 different genes. This study profiled wild-type male C57BL/6J mice and age-matched Clock/Clock homozygous mutants with the goal of studying the effects of disrupting the circadian clock. Two independent biological replicates were sampled every 4 hours for 2 complete circadian cycles in wild-type mice, and every 4 hours for a single circadian cycle (7 time points) in the Clock mutant. The raw intensity values were pre-processed using *gcRMA* software (Wu et al., 2004), log-transformed and normalized to zero-mean for each of the wild-type profiles.

Sine-wave detection The original analysis of this data (Miller et al., 2007) used the sine-wave matching algorithm of Straume (2004). They identified 848 distinct rhythmic probe sets in liver and 383 such probe sets in skeletal muscle. The authors filtered out probe sets below a threshold value of intensity, resulting in a final ranked list of 714 probe sets in the liver and 252 probe sets in the skeletal muscle. A subsequent analysis of the skeletal muscle data using the same sine-wave matching algorithm but with a more stringent cut-off threshold resulted in 215 probe sets (McCarthy et al., 2007).

Model-based detection Using our model we ranked the probe sets by their posterior probability of belonging to the periodic component (see Appendix 1). The posterior probabilities inferred for each of the probe sets are available in the Supplement. Among the top 25 probe sets there are 9 that were not among the top 400 ranked by sine-wave matching. Many of their profiles (Fig. 1) peak or drop at a single time point, and are poorly matched to a sinusoid shape. The fact that two of these are known core clock genes (*Arntl* and *Nr1d1*) suggests that such non-smooth measurements may be observed in true circadian genes due to the sparse sampling in time. The reverse list of probe sets, those ranked above 25 by the sine wave method but below 400 by the model, contains just the single probe set *Tns3*. The profile conforms to the sine-wave pattern but possesses a very small amplitude, and is assigned to the background component by the model. All of the other probe sets that were so highly ranked by the sine-wave method received posterior probabilities of periodicity above 0.9 from our model.

PCR Validation We used quantitative real-time PCR to estimate fold changes over time of the nine probe sets with known gene identities from the combined difference sets. Eight of these genes correspond to probe sets ranked highly by our model but not by the sine-wave method, and the ninth (*Tns3*) was the gene ranked highly by the sine-wave method but not by our method. Details of the PCR experiment are described in Appendix 4.

Fig. 4 shows estimated log-fold change at each of the 12 time points covering 2 complete circadian cycles. The ordering of genes in the panels is the same as in Fig. 1, except that the gene *Tubb2* (which was unidentified at the time we performed PCR) is replaced with the *Tns3* gene. The PCR results for *Tns3* indicate that the signal-to-noise ratio is smaller than 1: the variance of its mean profile over time (0.014) is smaller than the average replicate variability (0.0192) and thus quantitative PCR does not support circadian changes in this gene. This example illustrates how an explicit background model can use replicate variability to filter out noisy profiles that may appear periodic to methods that do not weigh the magnitude of changes in the averaged profiles against the variability of the replicates.

In contrast, all of the genes identified by the model except for *Zfp929* show profiles consistent with circadian regulation. They change significantly over time and the changes are consistent across the cycles. P-values (from an ANOVA periodicity detector) for these 7 profiles are below 2.13×10^{-6}; the largest value corresponds to *Rnase4*. The profile of *Zfp929* shows substantially smaller variation over time than the other 7 genes, and little similarity across the cycles (p-value 0.082). In the microarray experiment, this gene peaks at a single time point within each cycle (see Fig. 1) and may be an example of a false positive arising from the random background process.

FDR analysis We estimate the false discovery rate (FDR) to characterize the number of false positive probe sets that exceed a particular threshold on the posterior probability of periodicity. Assuming a correct model, the FDR for a given threshold can be estimated directly as the average of the posterior probability of non-periodicity, taken over probe sets above the threshold (Newton et al., 2004). A threshold of 0.9 selects 468 probe sets in the liver and 97 in the skeletal muscle corresponding to an estimated FDR of 2.28% and 2.23% respectively.

However, this estimate of the FDR is likely to be optimistic, since it assumes a correct model. As an alternative, we can estimate the FDR using a permutation test. We simulate data from the background distribution by permuting the time labels of our original data within each cycle. This permutation removes correlations in time but preserves the overall magnitude and observed replicate variability. Both the original and permuted data are then scored under the model. The FDR estimate is defined as the ratio of the number of permuted probe sets that exceed the posterior threshold to the number of the original probe sets that exceed the same threshold (Keegan et al., 2007); see Appendix 5. As expected, the FDR estimates based on time-permuted data are higher than those computed directly using the posterior probabilities, and suggests that for our threshold of 0.9 we can expect an FDR of approximately 14% (liver) and 17% (skeletal muscle). These rates are consistent with the PCR findings in the previous section, where there is evidence that one out of eight detections from our model is a false positive and the other seven are

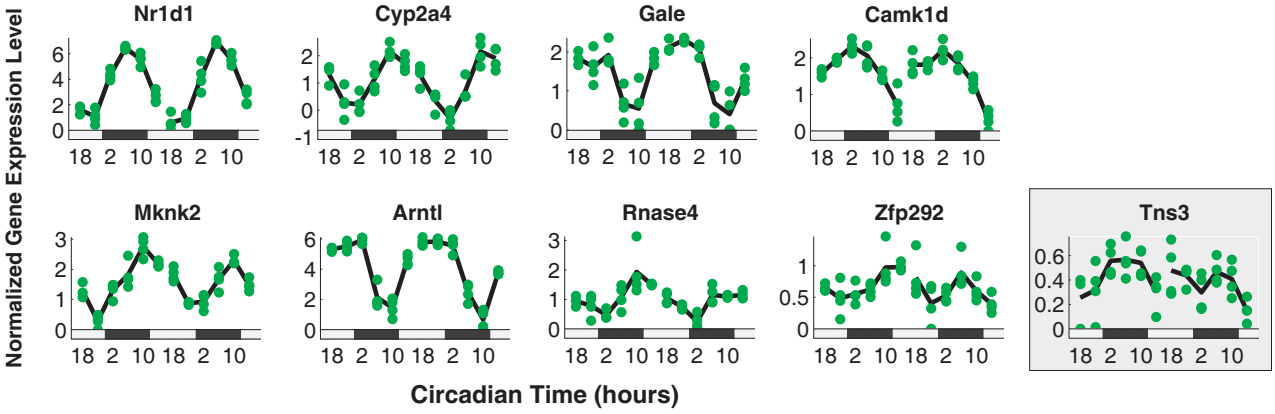

Fig. 4. Quantitative real-time PCR analysis of genes in mouse liver tissue, for eight genes ranked highly by the model and one (gray) ranked highly by a sine detector. PCR results support periodicity in all but two (*Zfp292* and *Tns3*).

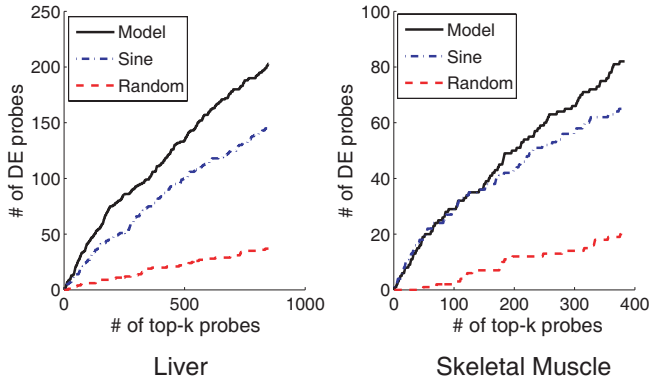

Fig. 5. Number of probe sets differentially expressed between the wild-type and the Clock mutant among those identified as rhythmic by the model and the sine-wave approach.

likely to be true positives. That these rates are not closer to zero may be due to the short and sparse nature of the time course data sets.

Comparison with mutant time series data To further validate the rankings from our model we evaluate the influence of the Clock mutation on the top ranking profiles. Our hypothesis is that a large fraction of genes under circadian regulation will change their expression patterns in response to the mutation. We perform a two-sample comparison between wild-type and mutant time series, using the Bioconductor *timecourse* package (Tai and Speed, 2007) and comparing mutant time points 1 through 7 to their corresponding circadian times in the wild type (time points 2 through 8). In this analysis we do not normalize the individual time courses to zero mean, so that a shift in absolute intensity can be detected as well.

Each plot in Fig. 5 shows how many probe sets ranked in the top k by periodicity are also in the top 5% ranked according to differential expression in mutants, in the liver (left) and skeletal muscle (right). Our model consistently selects more probe sets with altered expression patterns between the wild type and the mutant than the sine-wave method. Since temporal profiles of non-rhythmic genes are also affected by the mutation (Miller et al., 2007; McCarthy et al., 2007), this evaluation should be interpreted with caution. Nonetheless, in the absence of ground truth, these results provide additional (albeit indirect) evidence to indicate that the Bayesian model is able to consistently extract more relevant information from the data than a sine-wave approach.

DISCUSSION

Our Bayesian model for detecting periodic expression has a number of inherent simplifying assumptions which ensure a fast estimation process. Primarily, these assumptions are:

- All probe sets are independent
- Latent expression profiles (μ, σ) are Gaussian and independent across time, except as constrained by the component type.
- Replicate measurements are Gaussian given the latent profile

However, there are a number of possible extensions to the model which could lead to more robust detection, at the expense of increased computational costs.

Distributional assumptions. Conjugate prior distributions such as the Normal-inverse Gamma form assumed here ensure closed-form computation. However, some authors have suggested that non-Gaussian forms such as Gamma-Gamma models are more appropriate for expression data (Newton et al., 2004; Lewin et al., 2007). Our model can be easily re-cast with alternative priors, but may require numerical approximations in the computation of posterior probabilities.

Dependence across time. The assumption of a sinusoidal shape regularizes (or smooths) the estimated true profiles of periodic patterns. In contrast, the periodic component in our model does not have any such regularization (it treats sequential time points independently). Adding a non-diagonal covariance structure (correlation in time) such as that in Tai and Speed (2007) might increase the specificity of the model in detecting periodic probe sets with lower magnitude changes. For very sparsely sampled time points such as those in our data sets, however, this seems to be unnecessary.

Shared expression patterns. Computation is greatly simplified by assuming that all genes are independent, but many genes share similar patterns of expression (Do et al., 2005). Including a higher-level mixture model which groups similar periodic profiles together could help identify weak patterns that appear in many expression profiles by sharing information across genes. Although this change is easy

in principle, it greatly complicates the inference process. In this case, the estimates of periodicity for each gene become coupled and must be computed jointly rather than individually, and requires more complex methods such as Markov chain Monte Carlo.

CONCLUSIONS

In this paper we present an alternative to sinusoid or frequency-based testing for identifying periodic patterns in gene expression time series data. We argue that in typical experiments with only a small number of samples per cycle, we should test for arbitrary patterns which are repeated between cycles, rather than parametric shapes. To this end, we propose a Bayesian mixture model for identifying patterns of unconstrained shape, which stand out as both differentially and periodically expressed. The algorithm is computationally fast and easy to implement due to the conjugate nature of the underlying Bayesian model.

Using two experimental data sets we showed that our proposed method identifies a number of patterns, many with sharp transitions compared to the sampling rate that would be missed by a conventional sine-wave detector. Moreover, the Bayesian model identifications are supported by subsequent real-time PCR experiments and comparison to Clock-mutant expression profiles. This suggests that these detections are true positives missed by the analysis methods in common use.

ACKNOWLEDGEMENTS

This research was supported by National Institutes of Health-National Institute of Arthritis and Musculoskeletal and Skin Diseases Grant AR 44882 (to B.A.), National Science Foundation Grant NSF IIS-0431085 (to P.S.), and National Library of Medicine-National Research Service Award 5 T15 LM00744 (to K.K.L. and D.C.).

REFERENCES

C. Andersson, A. Isaksson, and M. Gustafsson. Bayesian detection of periodic mRNA time profiles without use of training examples. *BMC Bioinformatics*, sep 2006.

K.-A. Do, P. Müller, and F. Tang. A Bayesian mixture model for differential gene expression. *J. Royal Stat. Soc. C*, 54(3):627–644, June 2005.

A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman & Hall, New York, NY, 1995.

M. Jordan. Graphical models. *Stat. Sci.*, 19(1):140–155, Feb. 2004.

K. P. Keegan, S. Pradhan, J.-P. Wang, and R. Allada. Meta-analysis of *drosophila* circadian microarray studies identifies a novel set of rhythmically expressed genes. *PLoS Comp. Bio.*, 3, Nov. 2007.

D. J. Lavery, L. Lopez-Molina, R. Margueron, F. Fleury-Olela, F. Conquet, U. Schibler, and C. Bonfils. Circadian expression of the steroid 15 alpha -hydroxylase (cyp2a4) and coumarin 7-hydroxylase (cyp2a5) genes in mouse liver is regulated by the par leucine zipper transcription factor dbp. *Mol. Cell. Biol.*, 19: 6488–6499, 1999.

A. Lewin, N. Bochkina, and S. Richardson. Fully Bayesian mixture model for differential gene expression: Simulations and model checks. *Stat. App. Genetics and Mol. Bio.*, 6, 2007.

K. Lin, D. Chudova, G. Hatfield, P. Smyth, and B. Andersen. Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance. *PNAS*, 101(45): 15955–15960, 2004.

Y. Luan and H. Li. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. *Bioinformatics*, 20(3):332–339, 2004.

J. J. McCarthy, J. L. Andrews, E. L. McDearmon, K. S. Campbell, B. K. Barber, B. H. Miller, J. R. Walker, J. B. Hogenesch, J. S. Takahashi, and K. A. Esser. Identification of the circadian transcriptome in adult mouse skeletal muscle. *Physiol. Genomics*, 31:86–95, Sept. 2007.

B. H. Miller, E. L. McDearmon, S. Panda, K. R. Hayes, J. Zhang, J. L. Andrews, M. P. Antoch, J. R. Walker, K. A. Esser, J. B. Hogenesch, and J. S. Takahashi. Circadian and clock-controlled regulation of the mouse transcriptome and cell proliferation. *PNAS*, 104:3342–7, Feb. 2007.

M. Newton, A. Noueiry, D. Sarkar, and P. Ahlquist. Detecting differential gene expression with a semiparametric hierarchical mixture model. *Biometrics*, 5:155–176, 2004.

K. Oishi, K. Miyazaki, K. Kadota, R. Kikuno, T. Nagase, G.-I. Atsumi, N. Ohkura, T. Azama, M. Mesaki, S. Yukimasa, H. Kobayashi, C. Iitaka, T. Umehara, M. Horikoshi, T. Kudo, Y. Shimizu, M. Yano, M. Monden, K. Machida, J. Matsuda, S. Horie, T. Todo, and N. Ishida. Genome-wide expression analysis of mouse liver reveals clock-regulated circadian output genes. *J. Bio. Chem.*, 278:41519–27, 2003. ISSN 00219258.

A. A. Ptitsyn, S. Zvonic, S. A. Conrad, L. K. Scott, R. L. Mynatt, and J. M. Gimble. Circadian clocks are resounding in peripheral tissues. *PLoS Comp. Bio.*, 2(3), Mar. 2006.

R. D. Rudic, P. McNamara, D. Reilly, T. Grosser, A.-M. Curtis, T. S. Price, S. Panda, J. B. Hogenesch, and G. A. FitzGerald. Bioinformatic analysis of circadian gene oscillation in mouse aorta. *Circulation*, 112:2716–2724, 2005.

M. C. Rudolph, J. L. McManaman, L. Hunter, T. Phang, and M. C. Neville. Functional development of the mammary gland: use of expression profiling and trajectory clustering to reveal changes in gene expression during pregnancy, lactation, and involution. *J. Mam. Gland Bio. and Neoplasia*, 8:287–307, July 2003.

G. K. Smyth. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. App. in Genetics and Mol. Bio.*, 3(1), 2004.

K. Storch, O. Lipan, I. Leykin, N. Viswanathan, F. Davis, W. Wong, and C. Weitz. Extensive and divergent circadian gene expression in liver and heart. *Nature*, 417:78–83, May 2002.

M. Straume. DNA microarray time series analysis: Automated statistical assessment of circadian rhythms in gene expression patterning. *Methods in Enzymology*, 383:149–166, 2004.

Y. Tai and T. Speed. A multivariate empirical Bayes statistic for replicated micrarray time course data. *Ann. Stat.*, 34(5):2387–2412, 2006.

Y. C. Tai and T. P. Speed. On the gene ranking of replicated microarray time course data. Technical Report 735, University of California, Berkeley, 2007.

H. Wijnen, F. Naef, C. Boothroyd, A. Claridge-Chang, and M. W. Young. Control of daily transcript oscillations in *drosophila* by light and the circadian clock. *PLoS Genetics*, 2, 2006.

Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer. A model-based background adjustment for oligonucleotide expression arrays. *JASA*, 99(468):909–917, 2004.

1 COMPUTING MARGINAL LIKELIHOOD UNDER THREE MODEL COMPONENTS

To compute the posterior probability of periodicity for a particular probe set i using Equation (1) we need to evaluate the three marginal likelihood terms $P(Y_i|\Theta_z, Z_i = z), z \in \{b, d, p\}$ where the latent mean profile and replicate variance are each integrated out. These marginal likelihood terms can be computed in closed form using Bayes rule and the formula for the posterior probability under the Normal-inverse Gamma prior. In Section A.1 below we describe the computation of marginal likelihood assuming a series of one-dimensional observations X and scalar (μ, σ), and then extend this in section 1.2 to compute the marginal likelihoods under the three model components in our model.

1.1 Marginal likelihood under the Normal-inverse Gamma prior

Let $X = \{X_k, 1 \le k \le n\}$ be a series of one-dimensional observations generated by a Gaussian distribution with a Normal-inverse Gamma prior for scalar mean and variance (μ, σ):

$$P(\sigma) = \Gamma^{-1}(\sigma \mid a, b)$$
$$P(\mu|\sigma) = \mathrm{N}(\mu \mid \nu, \sigma/\eta)$$
$$P(X_k|\mu, \sigma) = \mathrm{N}(X_k|\mu, \sigma)$$

Here (ν, η) are the mean and the scale of the location μ and (a, b) are the degrees of freedom and the scale of the inverse Gamma distribution for variance σ. Let Θ denote the four parameters of the prior $\Theta = \{\nu, \eta, a, b\}$.

To evaluate the marginal likelihood of observations X, we use the following Bayesian identity, evaluated at some arbitrary values (μ^*, σ^*):

$$P(X \mid \Theta) = \frac{P(X|\mu^*, \sigma^*, \Theta)P(\mu^*, \sigma^*|\Theta)}{P(\mu^*, \sigma^*|X, \Theta)} \tag{2}$$

The likelihood is a product of Gaussian densities,

$$P(X|\mu^*, \sigma^*, \Theta) = \prod_{k=1}^{n} \mathrm{N}(X_k|\mu^*, \sigma^*) \tag{3}$$

while the prior is a product of an inverse Gamma for σ^* and a Gaussian distribution for μ^*:

$$P(\mu^*, \sigma^*|\Theta) = \Gamma^{-1}(\sigma^*|a, b)\mathrm{N}(\mu^*|\nu, \sigma^*/\eta) \tag{4}$$

The posterior also has a Normal-inverse Gamma structure (Gelman et al., 1995),

$$P(\mu^*, \sigma^*|X, \Theta) = \Gamma^{-1}(\sigma^*|\tilde{a}, \tilde{b})\mathrm{N}(\mu^*|\tilde{\nu}, \sigma^*/\tilde{\eta}) \tag{5}$$

where the posterior parameters $(\tilde{\nu}, \tilde{\eta}\,;\, \tilde{a}, \tilde{b})$ are given by

$$\tilde{\eta} = \eta + n \qquad \tilde{\nu} = \frac{\eta}{\eta + n}\nu + \frac{n}{\eta + n}\bar{X} \tag{6}$$

$$\tilde{a} = a + \frac{1}{2}n \qquad \tilde{b} = b + \frac{1}{2}S + \frac{1}{2}\frac{k\eta}{\eta + n}\left(\bar{X} - \nu\right)^2 \tag{7}$$

and the empirical means and sum of squares are

$$\bar{X} = \frac{1}{n}\sum_{k=1}^{n} X_k \qquad S = \sum_{k=1}^{n} \left(X_k - \bar{X}\right)^2 \tag{8}$$

Substituting equations (3) through (8) into the Bayesian identity in Equation (2), and using $\mu^* = \nu$ and $\sigma^* = 1$, provides the marginal likelihood $P(X \mid \Theta)$.

1.2 Marginal likelihood for model components

The background component assumes a scalar mean and replicate variance per probe set, and the marginal likelihood is computed directly as in Equation (2) for $X_i^b = \{Y_{ijk}^c, 1 \le T, 1 \le c \le C, 1 \le k \le n_{ij}^c\}$:

$$P(Y_i|\Theta_b, Z_i = b) = P(X_i^b \mid \nu^b, \eta^b, a^b, b^b)$$

The component for differential expression assumes an independent scalar mean and variance for each time point within each of the individual cycles, so the marginal likelihood is a product of likelihood terms for the corresponding observations $X_{ij}^{d,c} = \{Y_{ijk}^c, 1 \le k \le n_{ij}^c\}$:

$$P(Y_i|\Theta_d, Z_i = d) = \prod_{c=1}^{C}\prod_{j=1}^{T} P(X_{ij}^{d,c} \mid \nu_j^{d,c}, \eta_j^{d,c}, a_j^{d,c}, b_j^{d,c})$$

Finally, the periodic component assumes scalar mean and variance for each time point within a single cycle, and the marginal likelihood is a product of likelihood terms corresponding to observations $X_{ij}^p = \{Y_{ijk}^c, 1 \le c \le C, 1 \le k \le n_{ij}^c\}$:

$$P(Y_i|\Theta_p, Z_i = p) = \prod_{j=1}^{T} P(X_{ij}^p \mid \nu_j^p, \eta_j^p, a_j^p, b_j^p)$$

The three equations above for the marginal likelihood under individual components can be combined as in Equation (1) to provide a closed-form estimate of the posterior probability of periodicity for each of the probe sets.

2 ESTIMATING INVERSE GAMMA PARAMETERS IN THE NORMAL-INVERSE GAMMA PRIOR

In this section we describe a maximum likelihood approach for estimating the parameters of the inverse Gamma distribution for variance. Assume that we have a collection of N replicated observations $\{Y_{ik}, 1 \le i \le N, 1 \le k \le n_i\}$ generated by a Normal-inverse Gamma model with parameters $\mathrm{NIG}(\nu, k, a, b)$. To estimate parameters (a, b) of the inverse Gamma distribution, we first derive the F-distribution of the sum of squared errors S under this prior, denoted by $P(S|a, b)$. Once this distribution is known, we can evaluate the likelihood of the observed values of the statistic $P(S_{1\ldots N}|a, b)$ as a function of parameters (a, b) and find the maximum likelihood solution.

Formally, let $\bar{X}_i$ and S_i be the empirical mean and sum of squares for set i, given as in (8). For a given mean and variance (μ_i, σ_i), the observations Y_{ik} have a Gaussian distribution $\mathrm{N}(\mu_i, \sigma_i)$ and the sum of squared errors S_i follows a Gamma distribution with parameters (q_i, θ_i). Defining

$$q_i = (n_i - 1)/2 \qquad \theta_i = 2\sigma_i$$

we have

$$P(S_i|\sigma_i) = \Gamma(S_i|q_i, \theta_i) = \frac{1}{\theta_i^{q_i}\Gamma(q_i)} S_i^{q_i - 1} \exp\left(-\frac{S_i}{\theta_i}\right).$$

To obtain the distribution of S given (a, b), we analytically integrate out the replicate variance and obtain the functional form of an F-distribution:

$$\begin{aligned} P(S|a,b) &= \int_\sigma P(S|\sigma)P(\sigma|a,b)\mathrm{d}\sigma \\ &= \frac{b^a\Gamma(a+q)}{2^q\Gamma(a)\Gamma(q)}\frac{S^{q-1}}{(b+S/2)^{(a+q)}} \end{aligned} \tag{9}$$

The values of each statistic S_i for different observations i are conditionally independent given (a, b), and the likelihood of the entire sample is thus the product

$$P(S_{1\ldots N}) = \prod_{i=1}^{N}\frac{b^a\Gamma(a+q_i)}{2^q\Gamma(a)\Gamma(q_i)}\frac{S_i^{q_i-1}}{(b+S_i/2)^{(a+q_i)}}$$

To find the maximum likelihood solution, we take derivatives of the log-likelihood with respect to (a, b) and set them to zero:

$$\sum_{i=1}^{N}\left[\frac{a}{b}-\frac{a+q_i}{b+S_i/2}\right]=0$$

$$\sum_{i=1}^{N}\left[\psi(a+q)-\psi(a)+\log(b)-\log(b+S_i/2)\right]=0$$

Here $\psi(x)$ denotes the derivative of the logarithm of the Gamma function. The first equation is linear in a, and we solve for a in terms of b and substitute into the second equation, obtaining a single non-linear equation that defines the scale parameter b of the inverse Gamma distribution. Solving this equation for b using any zero-finding algorithm provides maximum likelihood estimates of both parameters.

3 ESTIMATING LOCATION SCALE IN THE NORMAL-INVERSE GAMMA PRIOR

In this section we derive maximum likelihood estimates of the location scale η in the Normal-inverse Gamma prior $\mathrm{NIG}(\nu, \eta\,;\, a, b)$, when all other parameters (ν, a, b) are fixed. We begin by writing out the likelihood of the data $Y = \{Y_{ik}, 1 \le i \le N, 1 \le k \le n_i\}$ as a function of η. Using the Bayesian identity as in Equation (2), we have that for arbitrary (μ^*, σ^*),

$$P(Y|\eta) = \prod_{i=1}^{N}\frac{P(Y_i|\mu^*,\sigma^*)P(\mu^*,\sigma^*|\eta)}{P(\mu^*,\sigma^*|Y_i,\eta)}$$

Let us denote the parameters of the posterior Normal-inverse Gamma distribution in (6) and (7) given observations Y_i and location scale η by $(\tilde{\nu}_i(\eta), \tilde{\eta}_i(\eta)\,;\, \tilde{a}_i(\eta), \tilde{b}_i(\eta))$, making their dependence on η explicit. The posterior probability in the denominator is then

$$P(\mu^*,\sigma^*|Y_i,\eta) = \Gamma^{-1}(\sigma^*|\tilde{a}_i(\eta),\tilde{b}_i(\eta))\mathrm{N}(\mu^*|\tilde{\nu}_i(\eta),\sigma^*/\tilde{\eta}_i(\eta))$$

Taking the log of $P(Y|\eta)$ and grouping terms which do not depend on η, we have

$$\begin{aligned} \log(P(Y|\eta)) = \sum_{n=1}^{N}\Big(&\log(\mathrm{N}(\mu^*|\nu,\sigma^*/\eta)) \\ &-\log(\Gamma^{-1}(\sigma^*|\tilde{a}_i(\eta),\tilde{b}_i(\eta))) \\ &-\log(\mathrm{N}(\mu^*|\tilde{\nu}_i(\eta),\sigma^*/\tilde{\eta}_i(\eta)))\Big)+C_1 \end{aligned}$$

which equals

$$\begin{aligned} \log(P(Y|\eta)) = \sum_{n=1}^{N}\Big(&\frac{1}{2}\log(\eta)-\eta\frac{(\mu^*-\tilde{\nu}_i(\eta))^2}{2\sigma^*} \\ &-(a+\frac{1}{2}R)\log(\tilde{b}_i(\eta))+\frac{1}{\sigma^*}\tilde{b}_i(\eta) \\ &-\frac{1}{2}\log(\eta+R)+(\eta+R)\frac{(\mu^*-\tilde{\nu}_i(\eta))^2}{2\sigma^*}\Big)+C_2 \end{aligned}$$

where C_1 and C_2 are constants independent of η. Taking the derivative with respect to η and setting it to zero, we obtain a non-linear equation for the maximum likelihood estimate of η:

$$\begin{aligned} \sum_{n=1}^{N}\Big(&\frac{1}{2\eta}-\frac{(\mu^*-\tilde{\nu}_i(\eta))^2}{2\sigma^*} \\ &-\left(a+\frac{1}{2}R\right)\frac{\tilde{b}_i'(\eta)}{\tilde{b}_i(\eta)}+\frac{1}{\sigma^*}\tilde{b}_i'(\eta)-\frac{1}{2(\eta+R)} \\ &+\frac{(\mu^*-\tilde{\nu}_i(\eta))^2}{2\sigma^*}-(\eta+R)\frac{(\mu^*-\tilde{\nu}_i(\eta))\tilde{\nu}_i'(\eta)}{\sigma^*}\Big)=0 \end{aligned} \tag{10}$$

Here $\tilde{\nu}_i'(\eta)$ and $\tilde{b}_i'(\eta)$ denote the derivatives of $\tilde{\nu}_i(\eta)$ and $\tilde{\nu}_i(\eta)$ with respect to η:

$$\tilde{\nu}_i'(\eta) = \frac{R(\nu-\bar{Y}_i)}{(\eta+R)^2} \qquad \tilde{b}_i'(\eta) = \frac{1}{2}(\bar{Y}_i-\nu)^2\frac{R^2}{(\eta+R)^2}$$

Non-linear zero-finding for the left hand side of Equation (10) provides an estimate of η. Each iteration of the method requires recalculating the parameters of the posterior for each observation i given the new value of η. The time complexity is thus linear in the number of observations, and for a few thousand genes the process of estimating k takes at most a few minutes on a single modern CPU.

4 PCR VALIDATION

To validate circadian cycling of the genes identified as periodic by the proposed approach but not the sine-wave model, we performed quantitative real-time PCR. The validation experiments were conducted on the genes that ranked above 25 according to one of the methods (either the proposed model model or the sine-wave detector (Straume, 2004)), but below 400 using the other method. Mice (strain: C57BL/6, age: 30 days old) were carefully housed under an alternating 12 hours light/dark cycle, and four mice were sacrificed every four hours over the course of 48 hours (52 mice total). Total RNA was extracted from the same region of the liver using the TRIzol method (Invitrogen) and synthesis of cDNA from extracted

RNA (1mg as input) was done using High Capacity cDNA Reverse Transcription Kit (Applied Biosystems) as previously described (Lin et al., 2004). Quantitative real-time PCR was performed using the following TaqMan Gene Expression Assays (Applied Biosystems): *Arntl* (Mm00500226_m1), *Camk1d* (Mm00616508_m1), *Cyp2a4* (Mm00487248_g1), *Gale* (Mm00617772_g1), *Mknk2* (Mm00458-026_m1), *Nr1d1* (Mm00520708_m1), *Rnase4* (Mm00491347_m1), *Tns3* (Mm01192797_g1), and *Zfp292* (Mm00497043_s1). For quantitative real-time PCR, three measurement replicates were used to determine the expression level (critical threshold value) per biological sample, and the expression for each sample was normalized to the endogenous control gene, *Gapdh* (Mm99999915_g1).

5 ESTIMATING FDR

For completeness, we briefly derive the estimate of Keegan et al. (2007). Suppose that we have run a classification test on N data, with N_1 data passing the threshold (predict class 1). Furthermore, we have some means of obtaining new data samples from the null hypothesis (class 0); in our case, this is by permuting data to remove time dependency. If we draw M data from class 0, and find that M_1 pass the threshold, we can estimate the probability of false detection as $p_{FD} \approx \frac{M_1}{M}$. Since fewer than our total N probe sets come from class 0, we should have fewer than approximately $\frac{M_1}{M} N$ probe sets which correspond to false detections. Then,

$$\text{FDR} \lesssim \frac{M_1}{M} \frac{N}{N_1}$$

and if $M = N$, this is simply $\frac{M_1}{N_1}$.