Methodology article

Phase analysis of circadian-related genes in two tissues
Delong Liu1,2, Shyamal D Peddada1, Leping Li1 and Clarice R Weinberg*1

Address: 1Biostatistics Branch, National Institute of Environmental Health Sciences, MD: A3-03, 111 TW Alexander Dr, Research Triangle Park, NC 27709, USA and 2CIIT Centers for Health Research, 6 Davis Drive, Research Triangle Park, NC, 27709, USA

Email: Delong Liu - dliu@ciit.org; Shyamal D Peddada - peddada@niehs.nih.gov; Leping Li - li3@niehs.nih.gov; Clarice R Weinberg* - weinberg@niehs.nih.gov

* Corresponding author

Abstract

Background: Recent circadian clock studies using gene expression microarray in two different tissues of mouse have revealed not all circadian-related genes are synchronized in phase or peak expression times across tissues in vivo. Instead, some circadian-related genes may be delayed by 4–8 hrs in peak expression in one tissue relative to the other. These interesting biological observations prompt a statistical question regarding how to distinguish the synchronized genes from genes that are systematically lagged in phase/peak expression time across two tissues.

Results: We propose a set of techniques from circular statistics to analyze phase angles of circadian-related genes in two tissues. We first estimate the phases of a cycling gene separately in each tissue, which are then used to estimate the paired angular difference of the phase angles of the gene in the two tissues. These differences are modeled as a mixture of two von Mises distributions which enables us to cluster genes into two groups; one group having synchronized transcripts with the same phase in the two tissues, the other containing transcripts with a discrepancy in phase between the two tissues. For each cluster of genes we assess the association of phases across the tissue types using circular-circular regression. We also develop a bootstrap methodology based on a circular-circular regression model to evaluate the improvement in fit provided by allowing two components versus a one-component von-Mises model.

Conclusion: We applied our proposed methodologies to the circadian-related genes common to heart and liver tissues in Storch et al. [2], and found that an estimated 80% of circadian-related transcripts common to heart and liver tissues were synchronized in phase, and the other 20% of transcripts were lagged about 8 hours in liver relative to heart. The bootstrap p-value for being one cluster is 0.063, which suggests the possibility of two clusters. Our methodologies can be extended to analyze peak expression times of circadian-related genes across more than two tissues, for example, kidney, heart, liver, and the suprachiasmatic nuclei (SCN) of the hypothalamus.

Background

Circadian rhythms (or the biologic clocks that control them) have stimulated interest in recent years due to their importance in orchestrating physiological behavior, biological processes, and adaptability of biological systems to changes in environment [1-3]. Many circadian-related genes have been explored using high-throughput DNA microarray technology [1-3]. These studies also have stim-
ulated efforts to apply and develop methodologies in circular/directional statistics to elucidate important characteristics of circadian gene expression and also compare their patterns of peak expression times (phase angles) across different tissue types, to help elucidate their diverse tissue-specific functions [1,2,4,5].

As periodic oscillation characterizes the expression pattern of both circadian genes and cell cycle genes, many correlation-based and Fourier-based methodologies [1-3] proposed for analyzing cell cycle gene expression can be directly applied to circadian gene expression analysis. However, there are some distinct differences between studies in cell cycle gene expression and circadian gene expression. First, most cell cycle gene expression patterns are based on cell cultures studied in vitro, while most circadian gene expressions are based on various tissues or organs in vivo. Consequently, circadian gene expression may be more complex or tissue/cell-specific. Second, the four phases of a cell cycle, namely, $G_1$, $S$, $G_2$, and $M$ phases, have been well characterized through intensive research over the last thirty years, and more than 54 mammalian [6] and 104 yeast cell cycle genes have been identified [7]. In contrast, to date, less is known about circadian genes: only eight core mammalian circadian genes have been identified: Csnk1e, Cry1, Cry2, Per1, Per2, Per3, Clock, and Bmal1 [8]. In addition, it is not clear whether these known circadian genes and any other circadian-related genes identified from high-throughput microarrays can be assigned to a few functional phases, in analogy to the phases ($G_1$, $S$, $G_2$, $M$) in cell cycle. Note that many studies on cell cycle gene expression based on microarray were on same organism [6] and cell-type [7] under different experimental conditions. Therefore, we expect that a set of cell cycle genes commonly expressed in various conditions are consistent in their peak expression/activation time [9]. However, it is an opening question whether phases or peak expression times for a set of circadian-related genes commonly expressed in multiple tissues, such as heart, liver, kidney, and SCN of the hypothalamus in phase with Bmal1 and in anti-phase with Per2 oscillation. These studies suggest that the coordinated temporal expression of circadian genes in-phase and anti-phase in different tissues is an interesting but a complex biological phenomenon. Statistical analysis tools for studying this type of interesting biological questions arising in recent genomics studies are needed.

To address the above questions, we propose a few steps in the following sections. Given a set of circadian-related genes common to two tissues, we first fit a random-epochs model [13] to the time-course expression for each gene individually in each of two tissues, to estimate its phase angles along with periods and amplitudes. The angular difference between the two phases for each gene can be represented as an angle, i.e. as a point on a circle. Using a mixture of two von Mises distributions, we cluster angular differences of the genes into two groups: genes whose expressions are synchronized (mean difference is close 0) in the two tissues and those whose expressions are different in the two tissues. The identified clusters may provide a hint on association of circadian genes specific to these tissues. We then assess the association of each set of genes common to two tissues using Down and Mardia’s circular-circular regression model [14]. In addition, we propose a new circular-circular regression-based boot-
strap method to assess the mixture of two homogeneous phase distributions for the two tissues. We illustrate the proposed methodologies using the heart and liver circadian-related gene expression data sets from Storch et al. [2].

### Methods

#### Phase estimation

One characterizing feature of circadian-related genes expression is the periodically oscillating pattern. Sinusoidal functions have been used to model the circadian gene expression level [1-3]. We apply the "random-periods model" [13] to estimate the phase angles, period, and amplitude together for a given circadian gene expression using nonlinear least-squares regression. While there is no attenuation in circadian gene expression, the sinusoidal component of the "random-periods model" is reduced to a simple sinusoidal function $K g \cos(2\pi t / T g + \phi g)$, where $K g$, $T g$, and $\phi g$ are the amplitude, period, and phase (angle) of gene g. The phase parameter $\phi g$ indicates when the expression of the g gene reaches its maximum.

### A mixture of two von Mises distributions for circular paired-difference data

After estimation of activation times or phase angles for a set of circadian-related genes that expressed in two tissues or organs, we are interested in examining whether the
phase angles are synchronous or not. Let \( \hat{x}_g \) and \( \hat{y}_g \) denote the estimated phase angles of a circadian-related gene \( g, g = 1, 2, \ldots, n \), in the two tissues \( x \) and \( y \), where \( -\pi \leq \hat{x}_g \leq \pi \) and \( -\pi \leq \hat{y}_g \leq \pi \). Further, we model the distribution of the angular difference
\[
\Delta_g = \hat{y}_g - \hat{x}_g, \quad -\pi \leq \Delta_g \leq \pi
\]
as a mixture of two von Mises distributions. One component will correspond to a subset of the \( n \) genes having the same phase angle in the two tissues and the other will correspond to genes having unequal phase angles in the two tissues. Thus the probability density function is given by:
\[
f(\Delta_g) = \sum_{i=1}^{2} p_i f_i(\Delta_g), \quad \sum_{i=1}^{2} p_i = 1,
\]
where \( f_i(\Delta_g) = \frac{1}{2\pi I_0(\kappa_i)} \exp(\kappa_i \cos(\Delta_g - \mu_i)), i = 1, 2; 0 \leq \kappa_i; -\pi < \mu_i \leq \pi \). Here, \( p_i \) is the mixing parameter, \( \mu_i \) is the mean direction for distribution \( i \), \( \kappa_i \) is the concentration parameter characterizing the variability of the estimated differences \( \Delta_g \) about \( \mu_i \), and \( I_0(\kappa_i) \) is the modified Bessel function of the first kind and order zero. We expect that one von Mises distribution has mean close to 0 radians, because it consists of a concordant subset of genes having the same phase in the two tissues, whereas the other distribution contains a set of “discordant” genes. The variation in shift characterizing genes of the second set can be measured by summing \( 1 - \cos(\Delta_g) \). [11].

The log likelihood for the mixture of two von Mises distributions in (2) is
\[
\log L(\Delta, \theta) = \sum_{g=1}^{n} \log \left( \sum_{i=1}^{2} \frac{p_i}{2\pi I_0(\kappa_i)} \exp(\kappa_i \cos(\Delta_g - \mu_i)) \right)
\]
The parameters in the vector \( \{p_1, \kappa_1, \mu_1, \kappa_2, \mu_2\} \) in the mixture model (3) can be estimated using the Newton-type optimization method in the Matlab optimization toolbox. To ensure convergence to the global solution, we use fifty random starting points. A comparison of the performance of various estimators can be found in [15]. We chose the Newton-type optimization method in the estimation due to its simplicity and flexibility of converting unconstrained searching to constrained optimization by adding constraints on the mixing parameter \( p_1 \), i.e., \( 0.15 < p_1 < 0.85 \) or the concentration parameter \( \kappa_1 \), or \( \kappa_2 \), i.e., \( \kappa_1 < 10 \) and \( \kappa_2 < 10 \). Upon obtaining the five estimated parameters \( \{\hat{p}_1, \hat{\kappa}_1, \hat{\mu}_1, \hat{\kappa}_2, \hat{\mu}_2\} \) in the mixture model (2), we statistically assign each of the \( \Delta_g \) to one of the two components based on its relative likelihood. That is, gene \( g \) is assigned to cluster 1 if \( \hat{p}_1 f_1(\Delta_g) > \hat{p}_2 f_2(\Delta_g) \), otherwise to cluster 2.

**Circular-circular regression**

In a recent article [9] we described the notion of association between the phase angles of a set of cell-cycle genes from a pair of experiments using the circular-circular regression model of Downs and Mardia [14]. Within each cluster obtained above, we shall apply the methodology described in [9] to examine the association between the estimated phase angles of the genes in the two tissues.

Consider a pair of angular random variables for cluster \( i \) as \( (\hat{x}_g, \hat{y}_g), i = 1, 2, g = 1, \ldots, n \), with mean directions \( \alpha_i \) and \( \beta_i \) respectively. Further, suppose \( \eta_{ig} \) denotes the mean direction of \( \hat{x}_g \) given \( \hat{y}_g \). In the present context, this would be the mean estimated phase angle of a gene in one tissue, conditional on its estimated phase angle in the other tissue. Downs and Mardia [16] introduced the fol-
lowing flexible circular regression model to regress $\hat{\phi}_g^y$ on $\hat{\phi}_g^x$

$$\tan \frac{\eta_g - \beta_i}{2} = \omega_i \tan \frac{\phi_g^x - \alpha_i}{2},$$

where $\omega_i$ denotes the "slope" parameter of the regression and $\eta_g$ is the mean direction of $\hat{\phi}_g^y$ conditional on $\hat{\phi}_g^x$.

The above model allows for estimating not only the rotational angle $\theta_i = \beta_i - \alpha_i$, but also the slope parameter $\omega_i$. As in Downs and Mardia [16], to avoid multiple solutions, we restrict $-1 \leq \omega_i \leq 1$ and $-\pi \leq \alpha_i \leq \pi$ and $-\pi \leq \beta_i \leq \pi$. We model the conditional distribution of $\hat{\phi}_g^y$ given $\hat{\phi}_g^x$ as a von Mises with concentration parameter $\kappa_i^c$, i.e.,

$$\hat{\phi}_g^y | \hat{\phi}_g^x \sim M\left( \eta_g(\hat{\phi}_g^x; \alpha_i, \beta_i, \omega_i) \kappa_i^c \right).$$

As shown in Downs and Mardia [16], the angular error $\hat{\phi}_g^y - \eta_g(\hat{\phi}_g^x; \alpha_i, \beta_i, \omega_i)$ is von Mises with mean 0 and concentration parameter $\kappa_i^v$, where

$$\eta_g(\hat{\phi}_g^x; \alpha_i, \beta_i, \omega_i) = \beta_i + 2 \tan^{-1} \left( \omega_i \tan \frac{1}{2} (\hat{\phi}_g^x - \alpha_i) \right).$$

The association from one tissue to the other of genes in each cluster $(\hat{\phi}_g^y, \hat{\phi}_g^x)$, $i = 1, 2$, can then be assessed using the $F$-test derived by Downs and Mardia [16].
A bootstrap test for number of clusters
To assess whether there are two clusters in the mixture of \( \Delta_g \), \( g = 1, \ldots, n \), in (1), in the following we propose a bootstrap methodology to test the null hypothesis that \( \Delta_g \)’s are a random sample from a single von-Mises distribution against the alternative hypothesis that they are from a mixture of two independent von-Mises distributions.

Let \( cv = \sum_{g=1}^{n} (1 - \cos(r_g)) \) denote an estimate of the circular variance for the combined sample of \( n = n_1 + n_2 \) observations, based on residuals from a single circular regression, while \( cv_1 = \sum_{g=1}^{n_1} (1 - \cos(r_g)) \) and \( cv_2 = \sum_{g=1}^{n_2} (1 - \cos(r_g)) \) denote the estimates of the circular variances for the two individual clusters separately based on residuals from two circular regressions.

The proposed bootstrap procedure is described in the following steps:

1) Regress phase angles in \( y \) tissue on phase angles in \( x \) tissue using the circular-circular regression model (4) and compute the circular variance \( cv \) based on the residuals \( r_g \), \( g = 1, \ldots, n \), from this single circular regression;

2) Compute for each gene \( g \) the difference \( \Delta_g = \hat{\phi}_g^y - \hat{\phi}_g^x \), where \( \hat{\phi}_g^y \) and \( \hat{\phi}_g^x \) are the estimated phase angles for gene \( g \) in tissues \( y \) and \( x \);

3) Fit the phase differences \( \Delta_g, g = 1, \ldots, n \), to a two-component mixture of von-Mises model, obtaining two sepa-
rate clusters with \( n_1 \) and \( n_2 \) genes in provisional clusters 1 and 2, respectively;

4) Regress \( n_1 \) and \( n_2 \) phase angles of \( y_1 \) on \( x_1 \) and \( y_2 \) on \( x_2 \) separately for the two clusters, and obtain the residuals \( r_{\text{cluster}1} \) and \( r_{\text{cluster}2} \) from each of the two regressions;

5) Compute the circular variances \( cv_1 \) and \( cv_2 \) for each of the two-cluster sets of residuals \( r_{\text{cluster}1} \) and \( r_{\text{cluster}2} \) from the regressions carried out in step 4);

6) Calculate the test statistic: \( T = cv - cv_1 - cv_2 \)

7) Compute the absolute values of \( r_s \) obtained from Step 1, and randomly assign a +/- sign to each \( r_s \), call it \( r_s^* \), then obtain bootstrapped data \( r_{s,\text{bootstrap}} \) from \( r_s^* \);

8) Obtain each pseudo phase angle for the heart data by \( \eta_r + r_{s,\text{bootstrap}} \Rightarrow \phi_y^{\text{bootstrap}} \), where \( \eta_r \) is the predicted angle in tissue \( y \) based on regression performed in Step 1;

9) Repeat the loop from step 1) to 8) 3000 times using \( \phi_x^{\text{bootstrap}} \), the bootstrap data \( \theta_g^{y,\text{bootstrap}} \) to replace \( \phi_y^{\text{bootstrap}} \) and replacing \( r_s \) by \( r_{s,\text{bootstrap}} \). For the bootstrap data denote the \( T \) in step 6 by \( T_{\text{bootstrap}} \).

Then the bootstrap p-value is the proportion of \( T_{\text{bootstrap}} \) that are greater than the calculated value of test statistic \( T \).

Results

Datasets

We apply our methodologies to 52 circadian-related cycling transcripts that are expressed in mouse heart and
liver tissues, identified by Storch et al. [2]. In Storch’s studies, mice were entrained to a 12 hrs light/dark cycle for more than two weeks, then placed in a constant dim light for more than 42 hrs. The tissue samples were taken from sacrificed mice at 4-hour intervals for 48 hrs, or about two circadian cycles, as in the circadian studies of Panda et al. [1].

Due to the poor fit of our random-periods model [13] to the expression of four transcripts (accession numbers: AI834950, AF003348, AF043288, and AB014494), we excluded them from the list of 52 circadian-related transcripts [2] in our analysis. The estimated phase angles for the remaining 48 transcripts in both heart and liver for clusters 1 and 2 are listed in Tables 1 and 2, respectively. The angular differences (heart, denoted as $\gamma$, minus liver, denoted as $\chi$) $\Delta_\chi$ are plotted in Figure 1.

The 48 circadian-related cycling transcripts were assigned into two clusters based on a mixture of two von Mises distributions, as described above. The first cluster contains 38 genes (Table 1) and the second cluster contains 10 genes (Table 2). The five estimated parameters ($\hat{\mu}_1, \hat{\kappa}_1, \hat{\mu}_2, \hat{\kappa}_2$) = (0.79, 1.50, -0.07, 4.56, 2.16). The distance between the two clusters in mean direction ($\hat{\mu}_1 - \hat{\mu}_2$) is 2.23 rads, suggesting that the ten transcripts in cluster 2 have different points of peak expression in heart and liver. This can also be seen in Figures 2 and 3. The mean direction of 38 phase angles in the cluster 1 is -0.07 rads, suggesting that the peak expression times for the 38 cycling genes in heart and liver are close to synchronized. In contrast, the peak expression times for the 10 genes in heart and liver (in cluster 2) are away from the 0 direction by 2.16 rads, with heart ahead of liver by about 8 hrs. This result suggests that these 10 discordant circadian-related genes may play different roles in the heart and liver.

We estimated the across-tissue association of the genes in each of the two clusters by regressing the phase angles in heart, denoted as $\gamma$, onto those in liver, denoted as $\chi$, using the circular-circular regression model described above (5). The estimated rotational parameters, the slope, and the concentration parameter for the von Mises distribution are ($\hat{\alpha}_1, \hat{\beta}_1, \hat{\omega}_1, \hat{\kappa}_1^1$) = (-0.83, -0.91, 0.58, 1.85) for cluster 1 and ($\hat{\alpha}_2, \hat{\beta}_2, \hat{\omega}_2, \hat{\kappa}_2^2$) = (-2.85, -0.76, 0.86, 9.41) for cluster 2. Figure 2 shows the residuals for all genes, i.e. the angular difference between the phase angle in heart and the prediction based on the same gene's pattern of expression in liver. The residuals for the 10 genes in the second cluster are shown in Figure 3. We can see that the activation times of these transcripts in heart and liver tissues are matched well after a 2.16 radian clockwise rotation of the phase angles in heart relative to in liver. Upon obtaining the ‘slope parameters’ $\hat{\omega}_1$ and $\hat{\omega}_2$, we tested the null hypothesis $H_0: \omega_i = 0$ vs. the alternative hypothesis, $H_1: \omega_i \neq 0, i = 1, 2$, using the test statistic derived by Downs and Mardia [14]. The corresponding p-values for clusters 1 and 2 were less than $5 \times 10^{-7}$ and $1.4 \times 10^{-4}$, respectively, suggesting that the associations of the circadian activation times in heart and liver are strong in both clusters. Here we included the expression plots of the transcripts AI850638 and AI846522 in heart and liver tissues [2] and the fits of our model to the two transcripts in Figures 4 and 5, accordingly. The plots reveal that the fits of our model to the data are reasonably good, and peak times of the two transcripts in liver tissue are markedly lagged relative to heart tissue.

Based on 3000 bootstraps using the procedure outlined in section of bootstrap test, we found that the estimated p-value for sufficiency of a one-component distribution in the phase difference of heart and liver tissues for the 48 genes was 0.063. Although this is not significant at 5% level of significance, it suggests that a two-component von-Mises distribution for the phase in heart and liver tissues better describes the relationships among the peak expression times for the studied genes.
Based on the single experimental dataset of Storch subsets among the 48 genes. Although our findings are result supports, albeit weakly, the existence of two distinct or 8 hours in heart than in liver. Our bootstrapping test and the other 10 genes express earlier by about 2.23 rads phase or peak expression times in heart and liver tissues, Instead, among the 48 genes, 38 are synchronized in maximally expressed at the same time in the two tissues. In particular, we evaluated the significance of the circular variances of the residuals from circular-circular regression of the phase angles in heart on in liver in one cluster vs. two clusters. In contrast, most studies on mixtures of circular variables have focused on dividing a set of data into two subsets. To the best of our knowledge, no studies have used a mixture of two components for circular datasets for testing heterogeneity of a cyclic pattern.

This work would not have been undertaken without the interesting observations by Panda et al. [1], Storch et al. [2] and Ueda et al. [3] that a few circadian-related genes expressed out of phase across two tissues of mouse. In addition, quantitatively the results of our statistical analysis depend on the approximation of sinusoidal waveform for circadian gene expression, on reasonably accurate estimation of the phase angles, on a relative large sample size of genes common to two tissues, on the approximate validity of the von Mises distribution for each cluster of differences. In previous circadian gene expression studies [1-3], a tissue sample was commonly taken at each 4-hr interval for two circadian cycles, i.e., 12 time points per gene expression. Although our fits to the 48 circadian gene expression in both tissues are reasonably good, as shown in Figures 4, 5, further experimental and simulation studies may be needed to understand the role of sample size and sampling frequency on phase estimation when a sinusoidal waveform is presumed for circadian rhythm. In this work, we considered the difference of two phases for a gene in two tissues as a random variable modeled by a von Mises distribution on a circle. The corresponding uncertainties are captured by the degree to which the von Mises distribution is spread out on the circle.

The 48 circadian related genes expressed in heart and liver of mouse each provide a pair of peak expression times. We have assumed that there are at most two clusters. Further experimental studies are needed for testing whether there might be more than two clusters with for 48 genes. While our method can hypothetically be extended to allow one to test the need for 3 clusters rather than for 2 clusters, the sample size of 48 genes, may not be sufficient to carry out such a test with much power. The results of the two-clusters analysis must be regarded as descriptive. Our analysis of circadian gene expression may serve to stimulate further methodological development in circular/directional statistical analysis of genes that may be expressed differently in phase in two or more tissues. The genes in each cluster would need to be scrutinized separately for further elucidation of their tissue-specific biological and physiological functions [1,2,5].

Our methodologies can be, in principle, extended to analyzing multiple circadian gene expression data sets across multiple tissues (organs), e.g., kidney, heart, liver, and SCN where investigators are interested in understanding whether there is one set of core circadian-related genes that are similar in their patterns of activation across different organs (or tissues), and others that are differently expressed in different tissues, as suggested by Reppert and Weaver [5]. Because the maximum likelihood approach

![Figure 5](image-url)

Discussion and conclusion

Our analysis of the peak expression times (phases) for a set of 48-circadian-related genes expressed in both heart and liver tissues [2] suggests that not all of the genes are maximally expressed at the same time in the two tissues. Instead, among the 48 genes, 38 are synchronized in phase or peak expression times in heart and liver tissues, and the other 10 genes express earlier by about 2.23 rads or 8 hours in heart than in liver. Our bootstrapping test result supports, albeit weakly, the existence of two distinct subsets among the 48 genes. Although our findings are based on the single experimental dataset of Storch et al. [2], our results are similar to an earlier observation made by Panda et al. [1] that the peak expression times for some genes are not synchronized in suprachiasmatic nuclei (SCN) compared to liver. One implication of our results, giving quantitative support for the conclusion of Storch et al. [2], may be that some commonly expressed circadian-related genes may perform different functions across different organs [1,2].

We have developed a new bootstrap method for assessing the adequacy of one versus the need for two clusters of genes in the two sets of phase angles in heart and liver tissues for the 48 genes. In particular, we evaluated the significance of the circular variances of the residuals from circular-circular regression of the phase angles in heart on in liver in one cluster vs. two clusters. In contrast, most studies on mixtures of circular variables have focused on dividing a set of data into two subsets. To the best of our knowledge, no studies have used a mixture of two compo-
considered in this paper may become challenging due to computational complexity, further methodology development is needed in this area. For such multiple mixture problems, one may want to consider a Bayesian or empirical Bayes approach. However, to the best of our knowledge, Bayes and empirical Bayes methods for mixture problems associated with circular data are not well developed. The present application provides an excellent opportunity for developing such methodology for mixture problems associated with circular data. Secondly, using the standard likelihood approach, we clustered the 48 genes into two clusters on the basis of the phase difference between the two tissues. It would be interesting and useful to derive an estimate of "reliability" of clustering for each gene. One possible approach is to perform a bootstrap by selecting a simple random sample of 48 genes from the list of 48 genes and classify them into two clusters using procedure described in this paper. This procedure could be repeated for a large number of times, say 1000. One then could estimate the proportion of times a gene was classified into one of two clusters. Unfortunately, such a procedure does not work well when the number of genes is small and their "true" cluster memberships are unknown and the numbers of genes in each cluster are highly unbalanced. Nonetheless, these important questions need to be addressed because the number of potential applications for such a procedure is ever growing.

Authors' contributions
DL conceived of the study, performed the calculations, and drafted the manuscript. SDP and CRW suggested the bootstrap method. SDP, CRW, and LL drafted the manuscript. All authors read and approved the final manuscript.

Acknowledgements
We thank David Umbach for reading our manuscript and offering suggestions, and Mei Liu, and Bhaskar Mandavalli for their helpful comments on an earlier version of the manuscript.

References
1. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesch JB: Coordinated transcription of key pathways in the mouse by the circadian clock. Cell 2002, 109:307-320.
2. Storch KF, Lapan O, Leykin I, Viswanathan N, David FC, Wong WH, "Room" CJ: Extensive and divergent circadian gene expression in liver and heart. Nature 2002, 417:78-83.
3. Ueda H, Chen W, Adachi A, Wakamatsu H, Hayashi S, Takasugi T, Nagano M, Nakahama K-I, Suzuki Y, Sugano S, Lino M, Shigeyoshi Y, Hashimoto S: A transcription factor response element for gene expression during circadian night. Nature 2002, 418:534-539.
4. Zyka MJ, Sheaman LP, Weaver DR, Reppert SM: Three period homologs in mammals differential light responses in the suprachiasmatic circadian clock and oscillating transcripts outside of brain. Neuron 1998, 20:1103-1110.
5. Reppert SM, Weaver DR: Coordination of circadian timing in mammals. Nature 2002, 418:935-941.
6. Whitfield ML, Sherlock G, Saldanha AJ, Murray JJ, Ball CA, Alexander KE, Matsej JC, Perou CM, Hurt MM, Brown PO, Bostein D: Identi-