Supporting material for “Order Restricted Inference for Oscillatory Systems for Detecting Rhythmic Signals”

Yolanda Larriba¹, Cristina Rueda¹, Miguel A. Fernández¹, and Shyamal D. Peddada ²

¹Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Paseo de Belén 7, Valladolid, 47011, Spain
²Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences (NIEHS), Alexander Dr., RTP, NC 27709, USA

This supporting materials text is divided in three sections. In Section 1 we describe the statistical methodology underlying ORIOS, in Section 2 we provide interesting time course plots of some genes which were referenced in the main article and in Section 3 we offer some additional numerical results dealing with the computation time of the methods considered and their performance for sparse time course data.

1 Theoretical details on the testing procedure

The algorithm ORIOS described in the main manuscript to detect rhythmic genes in oscillatory systems consists of two steps, namely, the filtering step and the classification step. Both these steps involve testing some statistical hypotheses and therefore in this section we describe the theoretical details of those statistical tests.

For a given gene, let $Y = (Y'_1, Y'_2)'$ denote its expression in the two periods of interest, where $Y_j = (Y_{j1}, Y_{j2}, \ldots, Y_{jn})'$ denotes its expression at the $n$ time points in the $j$th period with $j = 1, 2$. We further assume that the sampling variance is constant at all time points and that the gene expression at each time point follows a normal distribution and the expressions are uncorrelated at all time points. In other words, for each gene, its observed expression data are modelled by a signal plus error model $Y_j = \mu_j + \epsilon_j$, where $\epsilon_j \sim N_n(0, \sigma^2 I)$ independent and $j = 1, 2$.

Although various statistical hypotheses described in this section can be tested using the likelihood ratio test (LRT), the LRT can be computationally intensive when the hypotheses of interest are constrained by mathematical inequalities. This is particularly true as the number of inequalities is large, as is in the present case of circadian clock or cell-cycle gene expression studies. For this reason, we instead use
conditional tests (CT) which are well-known in the statistical literature (cf. [1, 2, 3, 4, 5, 6]). The CT are a modification to LRT where the critical value depends upon the number of level sets in the estimator of the underlying parameter. This results in computational simplicity and often substantial gains in the power.

1.1 A. The Filtering Step

Step A1: For each gene we estimate the location of the trough $L$ and that of the peak $U$, using the method described in the main paper. Note that this landmark estimation is common to both periods. In other words, we are combining the data from the two periods when estimating these landmarks. Since the alternative hypotheses depend upon the estimated values of $L$ and $U$, therefore the following “statistical hypotheses tests” should be viewed more as a pattern recognition algorithm rather than formal statistical tests since the hypotheses involve random variables (i.e. estimated $L$ and $U$). Note that the ORIOGEN software [7] also addresses the problem of detecting genes with cyclical patterns and is also based on order restricted inference based methods. The fundamental difference between ORIOGEN and ORIOS is that, unlike in the present case, ORIOGEN does not pre-estimate $L$ and $U$ but determines these values as part of its testing problem. However, as noted in [7] and in [8], as the number time points increases, such as in the case of circadian clock experiments, ORIOGEN is likely to be underpowered because it is testing at all possible locations of $L$ and $U$.

Step A2: For each gene, we test the following pairs of hypotheses:

$$
\begin{align*}
H_{10} : \mu_{11} = \cdots = \mu_{1n} & \quad & H_{20} : \mu_{21} = \cdots = \mu_{2n} \\
H_{11} : \mu_1 \in C_{LU} & \quad & H_{21} : \mu_2 \in C_{LU}
\end{align*}
$$

Note that each pair of the above hypotheses is tested separately using the data in the corresponding period, i.e., we test $H_{10}$ vs. $H_{11}$ in the first period and $H_{20}$ vs. $H_{21}$ is tested using the data in the second period. For each period $j = 1, 2$, $\mu_{H_{j0}} \in H_{j0}$ can be represented by $\mu^0 = (\mu(1, 1, \ldots, 1)')$ for some unknown scalar $\mu$. We construct the LRT statistics $T_{j0} = \frac{||\hat{\mu}_{H_{j1}} - \mu_{H_{j0}}||^2}{\sigma^2}$ along the lines of [9]. Under suitable regularity conditions, conditional on the level sets, for each $j = 1, 2$, the distribution of $T_{j0}$ under the null hypothesis $H_{j0}$, is known to be $\chi^2_{d_{j0}}$ (see [10] and [11]). Therefore, the corresponding $\alpha^*$-level conditional test (CT) rejects $H_{j0}$ when $T_{j0} \geq c(d_{j0})$ where $c(d_{j0})$ is defined as the $1-\alpha'$ percentile of the $\chi^2_{d_{j0}}$ such that:

$$
\alpha' = Pr(\chi^2_{d_{j0}} \geq c(d_{j0})) = \frac{\alpha^*}{1 - Pr_{\mu^0}(T_{j0} = 0)} \tag{1}
$$

where $d_{j0} = m_{j0} - 1$ and $m_{j0}$ is the number of level sets of $\mu_{H_{j1}}$. It was demonstrated in [12] that $\mu^0$ is the least favourable configuration (LFC) of parameters for the usual LRT. This fact guarantees that the
conditional test is asymptotically an $\alpha^*$-level test, and allows to obtain a $p$-value from a $\chi^2_{d_{j0}}$ as follows:

$$p_{j0} = Pr(T_{j0} \geq c(d_{j0})) [1 - Pr_{\mu^0}(T_{j0} = 0)] \quad (2)$$

For each gene, the above result assumes that the variance in the gene expression is constant $\sigma^2$ for all time points. However, since $\sigma^2$ is unknown in practice, we therefore replace $\sigma^2$ in $T_{j0}$ by its ANOVA based estimator $\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \sum_{j=1}^{2} (Y_{ij} - \bar{Y}_i)^2}{2n}$, which consistent for large samples.

For each gene, let $p_{j0}$, $j = 1, 2$, denote the $p$-value associated with $H_{j0}$ and let $p_0 = \max(p_{10}, p_{20})$. Since we are performing tests on a large number of genes, we adjust $p_0$ for multiple hypotheses testing using the Benjamini-Hochberg (BH) procedure. If the BH adjusted $p$-value is below the nominal FDR level $\alpha$ then we declare the gene to be potentially rhythmic gene (cyclical, quasi cyclical or flat), otherwise the gene is declared as a non-rhythmic gene (flat or non-flat and non-periodic).

1.2 B. Classification Step

For genes that are declared to be potentially rhythmic in the filtering step, in this step we shall further classify them as either cyclical, quasi cyclical or flat gene. Genes that are declared to be non-rhythmic in the filtering step, we shall further classify them as either flat or non-flat and non-periodic.

Depending upon the outcomes of the filtering step, we accomplish the above classifications by testing the following hypotheses in a sequence of two pairs of hypotheses.

$$H_0 : \mu_1 = \cdots = \mu_n$$
$$H_1 : \mu \in C_{LU}$$
$$H_2 : \mu \in \mathbb{R}^n \quad (3)$$

Before we describe this step we first note that the normality assumption in this step considers an identical $n$ dimensional mean vector $\mu$ for both periods, i.e. $\mu_1 = \mu_2 = \mu$. Thus, the observation vector $Y_j$ is distributed as $Y_j \sim ^{\text{iid}} N_n(\mu, \sigma^2 I)$ for $j = 1, 2$. Moreover, in contrast to the filtering step, in this step, when testing the null hypothesis $H_{j-1}$ against the corresponding alternative described in Steps B1 and B2 below, the parameter $\sigma^2$ is consistently estimated by $\hat{\sigma}^2_{H_{j-1}} = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{2} (Y_{ij} - \hat{\mu}_{H_{j-1}})^2$, see [9]. In each case, the LRT statistics is given by $T_j = \frac{2\|\hat{\mu}_{H_j} - \hat{\mu}_{H_{j-1}}\|^2}{\hat{\sigma}^2_{H_{j-1}}}$, see [9].

Similar to the filtering step, we use conditional testing using the LRT. Under $H_{j-1}$, and certain suitable regularity conditions, the conditional distribution of $T_j$, under the appropriate null hypotheses, is known to be a Beta distribution, $B_{a_j, b_j}$, see [10, 11]. The conditional $\alpha^*$-level test rejects when $T_j \geq c(m)$ where
c(m) is defined as the $1-\alpha'$ percentile of the $B_{a_j,b_j}$, such that:

$$\alpha' = Pr(B_{a_j,b_j} \geq c(m)) = \frac{\alpha^*}{1 - Pr_{\mu^0}(T_j = 0)}$$

(4)

where $(a_1, b_1) = ((m - 1)/2, (2n - m)/2)$, $(a_2, b_2) = ((n - m)/2, n/2)$, and $m$ is the number of level sets of $\hat{\mu}_{H_1}$. As before, the parameter configuration $\mu^0$, the mean under the appropriate null hypothesis, is the LFC of parameter under the null hypotheses for the LRT. This guarantees that the conditional test is asymptotically an $\alpha^*$-level test, and allows to obtain a $p$-value from a Beta distribution as follows:

$$p_j = Pr(T_{j0} \geq c(m))[1 - Pr_{\mu^0}(T_{j0} = 0)]$$

(5)

Step B1: For genes identified in the filtering step as potentially non-rhythmic, we test $H_0$ against $H_1$ (or more precisely $H_1 - H_0$). If, for such a gene, $H_0$ is rejected then the gene is declared as non-flat and non-periodic. These are the genes which have flat profiles with unusual peaks (or troughs). Genes for which $H_0$ is not rejected are classified as flat. To do so, $p$-values are adjusted for multiple testing using the BH procedure.

Step B2: For genes identified in the filtering step as potentially rhythmic, we first test the null hypothesis $H_1$ against the alternative $H_2$ (or more precisely $H_2 - H_1$).

Possibility (a): If $H_1$ is not rejected then for that gene we continue to test $H_0$ against $H_1$ (or more precisely $H_1 - H_0$). If $H_0$ is rejected then we classify the gene to be cyclical otherwise it is classified as flat. In order to do that, $p$-values from this last test are adjusted for multiple testing using the BH procedure.

Possibility (b): If $H_1$ is rejected then we do not test $H_0$ against $H_1$ (or more precisely $H_1 - H_0$) and declare the gene to be quasi cyclical. The reason for this classification is as follows. Since the cyclical shape is rejected by the present step, so this result taken together with the result of the filtering step, a logical conclusion one arrives at is that the gene has a rhythmic but not cyclical shape. Among the patterns considered in this paper, this pattern corresponds to the quasi cyclical pattern.
2 Supporting Figures

Figures in this Section are numbered according to the order they are referenced in the main manuscript. Figures S1, S2 and S3 correspond to the discussion in MATERIALS AND METHODS Section; Figures S4, S5 and S6 correspond to RESULTS and Figure S7 corresponds to DISCUSSION Section.

Plots in Figure S1 illustrates the wide range of genes with periodic shaped signal. In particular, they provide examples of cyclical and quasi cyclical patterned genes. Plots in Figure S2 represent examples of cyclical but not perfectly sinusoidal genes. Figure S3 contains genes for each of four shapes of signals described in the cartoon plots in Figure 1.

Figures S4, S5 and S6 contain a sample of genes, obtained from mouse pituitary, NIH3T3 and U2OS, that are declared as rhythmic by ORIOS but not by JTK and RAIN.

Finally, Figure S7 provides further examples of disagreements in classification among ORIOS, JTK and RAIN.
Figure S1: Rhythmic genes from mouse liver with circadian times from CT18 to CT65. Panels (a), (b), (c) and (d) are (sinusoidal) cyclical genes. Panels (e), (f), (g) and (h) quasi cyclical genes.
(a) Temporal profile of gene Ifrd1 (probe set 1416067_at)

(b) Temporal profile of gene Cnot7 (probe set 1423641_s_at)

(c) Temporal profile of gene Copg2os2 (probe set 1427320_at)

(d) Temporal profile of gene Stx5a (probe set 1449679_s_at)

Figure S2: Cyclical but not perfectly sinusoidal genes from mouse liver with circadian times from CT18 to CT65
(a) Temporal profile of Ptp4a1 (1419024_at): Cyclical profile

(b) Temporal profile of Hspa1b (1427126_at): Quasi cyclical profile

(c) Temporal profile of probe set 1444527_at: Non-flat and non-periodic profile

(d) Temporal profile of Tnnt3 (1450118_at): Flat profile

Figure S3: Gene examples from mouse liver for each of the four signal classes

Figure S4: Some genes in pituitary declared as rhythmic by ORIOS but declared to be non-rhythmic by JTK and RAIN with $\alpha = 0.01$
Figure S5: Some genes in NIH3T3 cell lines declared as rhythmic by ORIOS but declared to be non-rhythmic by JTK and RAIN with $\alpha = 0.01$.

Figure S6: Some genes in U2OS cell lines declared as rhythmic by ORIOS but declared to be non-rhythmic by JTK and RAIN with $\alpha = 0.01$. 
Figure S7: Different gene detections in mouse liver among ORIOS, JTK and RAIN. (a) and (b) are declared as rhythmic by ORIOS while JTK and RAIN declare them as non-rhythmic. (c) is declared as non-rhythmic by ORIOS while JTK and RAIN declare it as rhythmic. Moreover ORIOS classifies (a) as cyclical (indeed asymmetric), (b) as quasi cyclical and (c) as non-flat and non-periodic. (d) is declared as non-rhythmic by ORIOS and JTK while RAIN declares it as rhythmic.
3 Supporting Tables

Tables in this Section are numbered according to the order they are referenced in the Discussion Section of the main manuscript.

Table S1 shows the CPU execution time for ORIOS, JTK and RAIN under different numbers of probe sets, using simulated data according to the original simulation design proposed in the main document. Tables S2 and S3 illustrate the performance of ORIOS, JTK and RAIN with $\alpha = 0.05$ for sparse time course data where time points are available every 2 hours over 2 days (denoted as 2/2 design) and every 4 hours over 2 days (denoted as 4/2 design), respectively.

Table S1: Execution CPU time (in seconds) for ORIOS, JTK and RAIN

| Number of probe sets | ORIOS | JTK  | RAIN  |
|----------------------|-------|------|-------|
| 10000                | 28 sec| 31 sec| 208 sec|
| 20000                | 73 sec| 70 sec| 410 sec|
| 40000                | 128 sec| 127 sec| 760 sec|

Table S2: False positive and negative rates and mean error for the different signals in the 2/2 design for each classification algorithm considered ($\alpha = 0.05$)

|          | ORIOS | JTK | RAIN |
|----------|-------|-----|------|
| Cosine   | 0.003 | 0.000| 0.000|
| Cosine Two| 0.000| 0.000| 0.000|
| Cosine Peak| 0.002| 0.177| 0.001|
| Sine Square | 0.000| 0.000| 0.000|
| Asymmetric | 0.060| 0.987| 0.701|
| Quasi Cyclic | 0.032| 0.999| 0.366|
| Flat     | 0.108 | 0.004| 0.075|
| One Peak | 0.000 | 0.006| 0.060|
| Two Peaks| 0.000 | 0.004| 0.044|
| Cosine Flat| 0.020| 0.226| 0.638|
| Flat Trend| 0.020| 0.046| 0.472|
| MEAN ERROR | 0.030| 0.016| 0.057|
|          |       | 0.360| 0.258|
|          |       | 0.178|

Table S3: False positive and negative rates and mean error for the different signals in the 4/2 design for each classification algorithm considered ($\alpha = 0.05$)

|          | ORIOS | JTK | RAIN |
|----------|-------|-----|------|
| Cosine   | 0.141 | 0.448| 0.004|
| Cosine Peak| 0.145| 0.904| 0.187|
| Sine Square | 0.034| 0.533| 0.003|
| Asymmetric | 0.164| 0.995| 0.737|
| Flat     | 0.065 | 0.003| 0.097|
| One Peak | 0.000 | 0.002| 0.112|
| Two Peaks| 0.000 | 0.000| 0.056|
| Cosine Flat| 0.000| 0.016| 0.264|
| Flat Trend| 0.000| 0.002| 0.196|
| MEAN ERROR | 0.013| 0.121| 0.005|
|          |       | 0.720| 0.145|
|          |       | 0.233|
References

[1] Bartholomew, D.J. (1961) A Test of Homogeneity of Means Under Restricted Alternatives. *J. R. Stat. Soc.*, **23**, 239–281.

[2] Robertson, T. and Wegman, E.J. (1978) Likelihood Ratio Tests for Order Restrictions in Exponential Families. *Ann. Stat.*, **6**, 485–505.

[3] Iverson, G.J. and Harp, S.A. (1987) A conditional likelihood ratio test for order restrictions in exponential families. *Math. Soc. Sci.*, **14**, 141–159.

[4] Menéndez, J.A., Rueda, C. and Salvador, B. (1991) Dominance of Likelihood Ratio Tests under Order Constraints. *Ann. Stat.*, **20**, 2087–2099.

[5] Fernández, M.A., Rueda, C. and Peddada, S.D. (2012) Identification of a core set of signature cell cycle genes whose relative order of time to peak expression is conserved across species. *Nucleic Acids Res.*, **40**, 300–330.

[6] Rueda, C., Fernández, M.A. and Peddada, S.D. (2009) Estimation of Parameters Subject to Order Restrictions on a Circle with Application to Estimation of Phase Angles of Cell-Cycle Genes. *J. Am. Stat. Assoc.*, **104**, 338–347.

[7] Peddada, S., Harris, S., Zajd, J. and Harvey, E. (2005) ORIOGEN: order restricted inference for ordered gene expression data. *Bioinformatics*, **21**, 3933–3934.

[8] Peddada, S.D, Umbach., D.M. and Harris, S. (2012) Statistical analysis of gene expression studies with ordered experimental conditions. In Chakraborty, R., Rao, C.R. and Sen., P.(eds), *Handbook of Statistics: Bioinformatics in Human Health and Heredity*, Elsevier, NY, pp. 39–66.

[9] Robertson, T., Wright, F.T. and Dykstra, R.L. (1988) *Order Restricted Statistical Inference*. John Wiley & Sons, Chichester.

[10] Menéndez, J.A., Rueda, C. and Salvador, B. (1992) Testing Non-Oblique Hypothesis. *Commun. Stat. Theory*, **21**, 471–484.

[11] Hu, X. and Wright, F.T. (1994) Monotonicity properties of the power functions of likelihood ratio tests for normal mean hypotheses constrained by a linear space and a cone. *Ann. Stat.*, **22**, 1547–1554.

[12] Wollan, P.C. and Dykstra, R.L. (1986) Conditional Tests with an Order Restriction as a Null Hypothesis. In Dykstra, R., Robertson, T. and Wright, F.T. (eds), *Advances in Order Restricted Statistical Inference*, Springer, New York, pp. 279–295.