A bootstrap based measure robust to the choice of normalization methods for detecting rhythmic features in high dimensional data

Yolanda Larriba[1], Cristina Rueda[1], Miguel A. Fernández[1], and Shyamal D. Peddada [2]

[1]Departamento de Estadística e Investigación Operativa, Universidad de Valladolid, Valladolid, Spain

[2]Biostatistics and Computational Biology Branch, National Institute of Environmental Health Sciences (NIEHS), RTP, NC, USA. *Moved to Department of Biostatistics, University of Pittsburgh, PA 15261, USA*

Abstract

Motivation: Gene-expression data obtained from high throughput technologies are subject to various sources of noise and accordingly the raw data are pre-processed before formally analyzed. Normalization of the data is a key pre-processing step, since it removes systematic variations across arrays. There are numerous normalization methods available in the literature. Based on our experience, in the context of oscillatory systems, such as cell-cycle, circadian clock, etc., the choice of the normalization method may substantially impact the determination of a gene to be rhythmic. Thus rhythmicity of a gene can purely be an artifact of how the data were normalized. Since the determination of rhythmic genes is an important component of modern toxicological and pharmacological studies, it is important to determine truly rhythmic genes that are robust to the choice of a normalization method.
Results: In this paper we introduce a rhythmicity measure and a bootstrap methodology to detect rhythmic genes in an oscillatory system. Although the proposed methodology can be used for any high throughput gene expression data, in this paper we illustrate the proposed methodology using a publicly available circadian clock microarray gene-expression data. We demonstrate that the choice of normalization method has very little effect on the proposed methodology. Specifically, for any pair of normalization methods considered in this paper, the resulting values of the rhythmicity measure are highly correlated. Thus it suggests that the proposed measure is robust to the choice of a normalization method. Consequently, the rhythmicity of a gene is potentially not a mere artifact of the normalization method used. Lastly, as demonstrated in the paper, the proposed bootstrap methodology can also be used for simulating data for genes participating in an oscillatory system using a reference dataset.
Availability: A user friendly code implemented in R language can be downloaded from `http://www.niehs.nih.gov/\research/atniehs/labs/bb/staff/peddada/index.cfm`.
Contact: `sdp47@pitt.edu`

Keywords: Rhythmicity, microarray technology, normalization, oscillatory systems, circadian genes

1 Introduction

One of the major difficulties dealing with high-throughput gene-expression experiments is the noisy nature of the data Tu *et al.* (2002); Klebanov and Yakovlev (2007) that is intrinsic to each array. Thus an important component of gene expression analysis is pre-processing the data to remove (or reduce) sources of variation of non-biological origin among arrays Bolstad *et al.* (2003); Irizarry *et al.* (2003). A variety of pre-processing methods are available in literature, such as the Model-based Expression Index (MBEI) Li and Wong (2001), MAS 5.0 Hubbell *et al.* (2002); Liu *et al.* (2003) and Robust Multi-array Average (RMA) Irizarry *et al.* (2003). They usually involve three distinct steps, namely, Background correction, Normalization, and Summarization (Wu (2009)). Normalization is an important component of pre-processing Bolstad *et al.* (2003); Cheng *et al.* (2016), since it removes technical (i.e. non-biological) variations from the expression data. There are numerous methods available in the literature to normalize gene expression data and in this paper we consider the following popular normalization methods: *Quantile Bolstad et al. (2003), (Cyclic) Loess Bolstad et al. (2003), Contrast Åstrand (2003), Constant Bolstad et al. (2003), Invariant Set Li and Wong (2001), Qspline Workman et al. (2002) and Variance Stabilization Normalization (VSN) Huber et al. (2002).* Each normalization strategy is based on certain model and assumptions. Consequently, the resulting normalized expression data, and the downstream analyses, are expected to depend upon the normalization method used. It is well-known that many biological processes, such as metabolic cycle Slavov *et al.* (2012), cell-cycle Oliva *et al.* (2005); Peng *et al.* (2005); Rustici *et al.* (2004); Barragán *et al.* (2015) or the circadian clock Hughes *et al.* (2009) are governed by oscillatory systems consisting of numerous components that exhibit rhythmic or periodic patterns over time. There are several algorithms available in the literature to determine whether a gene is rhythmic or not. Some recent examples include JTK_Cycle (from now on JTK) Hughes *et al.* (2010), RAIN Thaben and Westermark (2014) and ORIOS Larriba *et al.* (2016). The performance of such algorithms potentially depends upon, among other factors, the normalization methods used. For example, Oliva *et al.* (2005); Peng *et al.* (2005); Rustici *et al.* (2004) conducted long-series time-course cell-cycle microarray study on *S. pombe* to identify rhythmic genes. The number of such genes identified by the three studies vary. Oliva *et al.* (2005) discovered 750 genes to be rhythmic, Peng *et al.* (2005) found about 747 rhythmic genes, whereas Rustici *et al.* (2004) discovered only 407 rhythmic genes. What is more interesting is that only 150 genes were identified to be periodic by all three studies. For more details, one may refer to Caretta-Cartozo *et al.* (2007).

There has not been a systematic evaluation of the impact of normalization methods on identifying rhythmic genes in studies involving oscillatory systems. Yet, researchers are interested in identifying rhythmic genes. A goal of this paper is to introduce a bootstrap based rhythmicity measure that is highly correlated across various normalization methods. As a consequence, a gene declared to be rhythmic under one normalization scheme is likely to be rhythmic under a different one. A by-product of our methodology is that the bootstrap procedure introduced in this paper can be used for simulating potentially realistic time-course circadian gene-expression data. Although several authors have developed algorithms for simulating time-course gene-expression data (cf. Freudenberg *et al.* (2004); Nykter *et al.* (2006); Parrish *et al.* (2009); Dembélé (2013)), each of them was specific to the experiment discussed in the paper and not broadly applicable. However, our proposed algorithm is very generic. It not only helps to identify rhythmic genes, but it also provides a tool to simulate potentially realistic circadian gene-expression data.

2 Methods

We begin this section by considering time-course data on two genes, namely, *Serpina3k* and *Maml1* from mouse liver tissue (see Hughes *et al.* (2010)) as the motivating examples. We normalized the data using, *Quantile, Constant, (Cyclic) Loess* and *Invariant Set* normalization methods. For illustration purposes, in the top panel of Figure S1 we report the time-course plots of the gene *Serpina3k* using *Quantile* (left panel) and *Constant* (right panel) normalization procedures. In the bottom panel of Figure S1 we provide the time-course plots of the gene *Maml1* using *Loess* (left panel) and *Invariant Set* (right panel) normalization procedures. As one can see, the time-course profiles of these genes are markedly different, depending upon which normalization procedure was used. Furthermore, if rhythmicity detection algorithm ORIOS is used then *Serpina3k* and *Maml1* are rhythmic genes if *Quantile* and *Loess* normalizations are used, respectively. But they cease to be rhythmic genes if *Constant* and *Invariant Set* normalization procedures are used. Similar conclusions are drawn if other rhythmicity detection algorithms, such as JTK and RAIN are used on these data. Such results in a genome-wide analysis can be very confusing and difficult to interpret.

Given a normalization method n and a rhythmicity detection algorithm a, the identification of rhythmic genes is based on the Benjamini-Hochberg adjusted p-values (p-value$^g(n,a)$, for $g = 1, \ldots, G$). For each gene $g = 1, \ldots, G$ we define the standard measure of gene rhythmicity associated to gene g, as follows:

$$M^g(n,a) = 1 - \text{p-value}^g(n,a). \tag{1}$$

In a vector notation we write $\mathbf{M}(n,a) = (M^1(n,a), \ldots, M^G(n,a))$, whose components take values between 0 and 1. Closer 0 indicates potentially non-rhythmic gene and closer 1 indicates potentially rhythmic gene.

For the plots in Figure S1 we have $M^{Serpina3k}(Quantile, ORIOS) = 0.996$, $M^{Serpina3k}\ (Constant, ORIOS) = 0.639$, $M^{Maml1}(Loess, ORIOS) = 0.992$, and $M^{Maml1}(InvariantSet,\ ORIOS) = 0.668$. Thus implying *Serpina3k* is potentially rhythmic under *Quantile* normalization but not under *Constant* and similarly, *Maml1* potentially rhythmic under *Loess* normalization but not likely under *Invariant Set.* This observation that normalization method n may impact the rhythmicity of a gene is not limited to the above genes but is rather a common feature of long-series time-course data as noted in Table 1. Of course, as

Table 1: Number of genes detected as rhythmic by ORIOS, JTK and RAIN according to the different normalization strategies and $M^g(n,a) \geq 0.99$ for $g = 1, \ldots, 45101$

Normalization Strategy	ORIOS	JTK	RAIN
0 Unnormalized	6432	923	4196
1 Quantile	9259	4998	12381
2 Loess	8812	3932	10571
3 Contrast	8435	4181	10273
4 Constant	6657	2726	9357
5 Invariant set	9604	5062	13385
6 Qspline	9163	4546	11828
7 VSN	8397	3608	10700

seen in Table 1, the rhythmicity algorithm a may also impact on determining if a gene is rhythmic or not. In modern pharmacological and toxicological studies, Zhang *et al.* (2014), there is a need for objective determination of rhythmic genes using high throughput time-course gene expression data. Motivated by this, we now introduce $\mathbf{M}_{Robust}(n,a)$, a modification of $\mathbf{M}(n,a)$ which is more robust with respect to n, the normalization method, than $\mathbf{M}(n,a)$ is. The proposed bootstrap methodology also provides us a tool to simulate time-course expression data for genes participating in oscillatory systems such as the circadian clock using a reference dataset.

2.1 Bootstrap methodology

Let $\mathbf{R}$ denote the tri-dimensional array of raw intensities obtained from a reference high-throughput data of an oscillatory system, such as the circadian clock. Data in $\mathbf{R}$ are expressed at probe level, where R^g_{pt} states the raw intensity value for gene g on probe p at time point ($array$) t, where $g = 1, \ldots, G$, $p = 1, \ldots, P$ and $t = 1, \ldots, T$. Let $\mathbf{X}$ be the tri-dimensional array derived from $\mathbf{R}$ after background correction. After normalization and summarization steps, a matrix of gene-expression values is finally obtained as the output of the pre-processing, see Figure S1 in the supplementary materials for details. The bootstrap approach proposed in this work is based on a linear model from corrected intensities $\mathbf{X}$ of a reference dataset as follows. Let $b = 1, \ldots, B$, denote bootstrap replications. Simulated gene expression datasets $\mathbf{X}^{(b)*}$ are generated according to parametric bootstrap, see Efron and Tibshirani (1994), as:

$$\log_2(X^{(b)g*}_{pt}) = \hat{\alpha}^g_p + \hat{\beta}^g_t + \epsilon^{(b)g*}_{pt}, \tag{2}$$

where $g = 1, \ldots, G$, $p = 1, \ldots, P$, $t = 1, \ldots, T$, $b = 1, \ldots, B$ and $\{\hat{\alpha}^g_p\}^G_{g=1}$ and $\{\hat{\beta}^g_t\}^G_{g=1}$ denote original estimates of probe and array effects obtained from corrected (and unnormalized) intensities $\mathbf{X}$. Following the methodology described in Irizarry *et al.* (2003), the median polish algorithm is used to estimate model parameters (Emerson and Hoaglin (1983)). This algorithm is similar to a two-way ANOVA based estimation procedure except that it employs medians instead of means to ensure robustness to outliers. Additionally as explained in Irizarry *et al.* (2003), it allows taking into account probe and array effects. $\epsilon^{(b)g*}_{pt}$ are identically and independently distributed according to a normal distribution $N(0, \hat{\sigma}^{2g})$, where $\hat{\sigma}^{2g}$ is the usual MSE under the original two-way model. If the expression data are count data, as in the case of RNA-seq, the observed counts may be transformed using a suitable variance stabilization transformation before appealing to the above model. Thus, for example, if the data are Poisson count data, then one may use the square transformation.

Using the time-course gene-expression data on *Copg* and *Bgee* (top and bottom left panels in Figure S2 respectively), we demonstrate how well our bootstrap based simulated data (the two right panels in Figure S2) resembles the pattern of expression of the real data. Thus it suggests that, in addition to detecting rhythmic genes robustly, our bootstrap methodology may also be useful for simulating reasonably realistic time-course expression patterns.

2.2 Robust measure of gene rhythmicity

For a rhythmicity detection algorithm a and a normalization strategy n, and a random realization of data, consider the rhythmicity statistic $\mathbf{M}(n,a)$. Let $\boldsymbol{\theta}(n,a) = \mathbb{E}(\mathbf{M}(n,a))$ be the parameter of interest and $\hat{\boldsymbol{\theta}}(n,a) = \mathbf{M}(n,a)$ be its estimator. For the b^{th} bootstrap sample using (2), $b = 1, 2, \ldots, B$, let

$\hat{\boldsymbol{\theta}}^{(b)*}(n,a) = (\hat{\theta}^{1(b)*}(n,a), \ldots, \hat{\theta}^{G(b)*}(n,a))$, denote the bootstrap estimate of $\boldsymbol{\theta}(n,a)$. Let $\widehat{\mathbb{E}}(\hat{\boldsymbol{\theta}}(n,a)) = \frac{1}{B}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)*}(n,a))$ and $\widehat{RMSE}(\hat{\boldsymbol{\theta}}(n,a)) = \sqrt{\frac{1}{B-1}\sum_{b=1}^{B}(\hat{\boldsymbol{\theta}}^{(b)*}(n,a) - \hat{\boldsymbol{\theta}}(n,a))^2}$. Then we define

$$\mathbf{M}_{\text{Robust}}(n,a) = \widehat{\mathbb{E}}(\hat{\boldsymbol{\theta}}(n,a)) - \widehat{RMSE}(\hat{\boldsymbol{\theta}}(n,a)), \tag{3}$$

as **measure of gene rhythmicity**. We call it a "robust" measure of gene rhythmicity because, as demonstrated later in this paper, by correcting for sample to sample variation in the rhythmicity measure (i.e. $RMSE$), it reduces the effect of the normalization method used.

3 Results

We re-analyzed a publicly available mouse liver data (GSE11923) (`http://www.ncbi.nlm.nih.gov/geo/`) of Hughes *et al.* (2010). The data consisted of 45101 probe sets (genes) at 48 time points representing two periods. Taking $M^g(n,a) \geq 0.99$ as the criterion to declare a gene to be rhythmic (the choice of this criterion is motivated by the findings of Larriba *et al.* (2016)), in Table 1 we summarize the results of three rhythmicity detection algorithms, namely ORIOS, JTK and RAIN using unnormalized data and seven normalization methods (*0.-Unnormalized, 1.-Quantile, 2.-(Cyclic) Loess, 3.-Contrast, 4.-Constant, 5.-Invariant Set, 6.-Qspline, 7.-VSN*). The number of rhythmic genes identified varies vastly among the normalization methods within each rhythmicity detection algorithm (Table 1). Thus it suggests that normalization methods have a large influence on whether a gene is classified as rhythmic or not.

To better illustrate this fact, a multiple correspondence analysis (MCA) was performed (see Figure S3). Multiple correspondence analysis (MCA) is an extension of correspondence analysis (CA) which allows one to analyze the pattern of relationships among several categorical variables Benzécri (1979); Greenacre (1984). Since we consider 3 rhythmicity identification algorithms and 8 normalization strategies consisting of unnormalized data and 7 normalization methods, each probe set can be described by 24 binary variables consisting of $1^{'}$s and $0^{'}$s depending on whether an algorithm a and a normalization strategy n declare a gene to be rhythmic or not. Thus resulting in a matrix of 45101 rows and 24 columns.

MCA is a dimension reduction procedure that can be used to represent distances among high dimensional vectors in a low-dimensional space, such as 2-dimensional plane. Using the MCA plots, one typically tries to interpret what each axis represents and evaluates relationships among the categories of different variables based on the distance among their representations on the graph.

The MCA plot based on the first two dimensions, which explain $\sim 64\%$ of the total variation in the data, is provided in Figure S3. Elements of the plot are as follows. For a rhythmicity algorithm a, a normalization method n and a rhythmicity category r ($r = 1$ if genes are declared as rhythmic and $r = 0$ if genes are declared as non-rhythmic), we plotted $3 \times 8 \times 2$ categories denoted by a_n_r. Then, we averaged the expression values of those genes that are declared as rhythmic (or non-rhythmic) under all normalizations strategies, i.e. those with $r = 1$ (or $r = 0$) for all strategies under a given algorithm, and overlaid these averaged profiles on the plot. For algorithm a, the averaged profile of rhythmic genes is denoted by a_Av_1 and the averaged profile of non-rhythmic one is denoted by a_Av_0. Furthermore, we also overlaid on this plot six figures $G_1, G_2, \ldots, G_6$ (as defined in inset table in Figure S3) representing patterns of those probe sets that are unanimously declared as either rhythmic or non-rhythmic by all normalization methods within a given algorithm. For example, G_1 (*Cyclic*) is a pattern of all probe sets that are declared as rhythmic by all normalization methods and all three algorithms. On the other hand, G_2 (*Quasi Cyclic*) is a pattern of all probe sets that are declared as rhythmic by all normalization methods using ORIOS but not rhythmic under all normalizations methods when using JTK or RAIN. Since genes declared as rhythmic by JTK algorithm are also declared as rhythmic by RAIN algorithm, therefore we are describing only 6 patterns $G_1, G_2, \ldots, G_6$ and not 8 patterns as one might expect.

In Figure S3, we interpret horizontal axis (Dim1) as the axis describing rhythmicity because all a_n_1 appear on the right hand side and almost all a_n_0 appear on the left hand side of the graph. Using $G_1, G_2, \ldots, G_6$, we see that Dim1 separates rhythmicity (Cyclic-shaped patterns) against non-rhythmicity (Flat-shaped patterns). Furthermore, it is interesting to note that rhythmic-shaped patterns (Cyclic, Quasi Cyclic and Asymmetric) identified by ORIOS are located in the upper portion of the first quadrant of the MCA plot and the third quadrant exclusively consists of non-rhythmic patterns identified by ORIOS. Thus the first and the third quadrants of MCA plot appear to distinguish ORIOS from the others. The vertical axis (Dim2) may be interpreted as the axis drawing distinctions between ORIOS and RAIN algorithms. Lastly, it is clear from the MCA plot that ORIOS normalization methods are less separated than JTK or RAIN, i.e. rhythmic (and non-rhythmic) groups are more compact when using ORIOS, which is one more reason, in addition to the results provided in Larriba *et al.* (2016) to prefer ORIOS as the algorithm for detecting rhythmic genes.

To show that our proposed rhythmicity measure $\mathbf{M}_{\text{Robust}}(n,a)$ is generally robust to the normalization methods, we computed the Spearman and Pearson correlation coefficients between $\mathbf{M}_{\text{Robust}}(n_i,a)$ and $\mathbf{M}_{\text{Robust}}(n_j,a)$, for all pairs of normalization methods n_i, n_j, $i \neq j$ and compared the correlations with those corresponding to the standard measure $\mathbf{M}(n,a)$. In addition to Spearman and Pearson correlation coefficient, we also computed the percent of concordance of rhythmic and non-rhythmic genes across all normalization methods using standard measure $\mathbf{M}(n,a)$ and the proposed robust measure $\mathbf{M}_{Robust}(n,a)$. Due to space reasons, in the main paper we only present the results for ORIOS, i.e. $a = ORIOS$, but the results corresponding to JTK and RAIN are provided in the supplementary materials.

In our correlation and concordance analyses reported in Figure S4, S5 and S6, we limited to only those probe sets that were considered to be rhythmic by the criterion $M^g(n, ORIOS) \geq 0.99$ for at least one normalization method n. Thus we limited to 15369 probe sets out of 45101. The left hand panels of Figure S4 S5 and S6 correspond to $\mathbf{M}(n, ORIOS)$, whereas the right hand panels correspond to $\mathbf{M}_{Robust}(n, ORIOS)$. From these figures it is clear that both correlation and the concordance increase substantially for every pair of normalization methods from the left panel to the right panel. To illustrate this fact, observe that the Spearman correlation between $\mathbf{M}(Qspline, ORIOS)$ and $\mathbf{M}(VSN, ORIOS)$ is 0.65 (left panel of Figure S4). However, the Spearman correlation between $\mathbf{M}_{Robust}(Qspline, ORIOS)$ and $\mathbf{M}_{Robust}(VSN, ORIOS)$ is 0.95 (right panel of Figure S4), which is a substantial increase. The increase is even more dramatic if one were to consider the Pearson correlation coefficient which increases from 0.31 to 0.91 (Figure S5). Even the the percentage of concordant genes between these normalization procedures increases dramatically by more than 27%, from 69.85% to 97.48% (Figure S6). For each normalization method n, these increases are further illustrated using scatter plots of the pairs $(M^g(n, ORIOS), M^g(Qspline, ORIOS))$ (left panel) and $(M^g_{Robust}(n, ORIOS), M^g_{Robust}(Qspline, ORIOS))$ (right panel) in Figure S7. The scatter plots on the left generally display highly non-elliptic scatter of points with no clear correlation. However, the scatter plots on the right panel which correspond to our robust method, appear to be very elliptic and in some cases with very small minor axis. As a by-product, these scatter plots together with Figure S4, S5 and S6, imply that among the 7 normalization methods, the *Constant* and *Invariant Set* normalization methods may be the least preferred normalization methods as the robust measure corresponding to these methods seem to be least correlated with others.

Similar dramatic increases for also seen for JTK and RAIN as described in the figures in the online supplementary text (Figures S2, S3, S4, S5, S6, S7, S8 and S9).

4 Discussion

Determination of circadian clock genes is an important problem in various fields, especially clinical pharmacology Zhang *et al.* (2014); Chen and Yang (2015) where they play an important role in drug delivery and medicine. However, identification of such rhythmic genes in genome wide studies involving oscillatory systems has been a long standing problem. While it is well acknowledged in the literature that normalization methods play an important role in determining differentially expressed genes in a pair of conditions, as demonstrated in this paper, they play a bigger role in determining rhythmic genes in long-series time course experiments. For example, as observed in Figure S1 and as seen from Spearman and Pearson correlations reported in Figure S4 and in Figure S5, the rhythmicity of a gene can be dramatically affected by the normalization method used. This is the first paper we know that studies this problem for long-series time-course experiments and provides a simple bootstrap based methodology that correlates well across various normalization methods. The pairwise correlations among the normalization methods improve dramatically by using our proposed methodology. For example the Pearson correlation coefficient between *Qspline* and *VSN* nearly triples from 0.31 to 0.91 after applying our robust methodology. All statistical decision rules require a user-supplied threshold when making inferences and the proposed methodology is no exception. The threshold of 0.99 used in our criterion for rhythmicity corresponds to 1% level of significance and is largely motivated by the specificity and sensitivity findings of Larriba *et al.* (2016).

Since the Spearman correlation coefficient is based on the ranks, we therefore make a crucial observation from Figure S4 that rank of rhythmicity of a gene is correlated across all normalization methods considered here when our bootstrap based methodology is applied. Thus, if a gene has a high rank of rhythmicity under one normalization method, then it is also expected to have a similarly high rank of rhythmicity under other normalization methods. Conversely, if a gene has a very low rhythmicity rank under one normalization method then it will likely to have low rank under a different normalization method. To illustrate this point, consider the two genes described in the motivating figure of this paper (Figure S1). As noted earlier, under the standard criterion $M^g(n, ORIOS) \geq 0.99$, the rhythmicity calls on these two genes highly depended upon the normalization method n. However, under the criterion $M^g_{Robust}(n, ORIOS) \geq 0.99$, neither of these genes are considered to be rhythmic. Specifically, using the normalization methods used earlier for Figure S1, we obtained the following robust rhythmicity measures $M^{Serpina3k}_{Robust}(Quantile, ORIOS) = 0.127$, $M^{Serpina3k}_{Robust}(Constant, ORIOS) = 0.367$, $M^{Maml1}_{Robust}(Loess, ORIOS) = 0.675$, and $M^{Maml1}_{Robust}(InvariantSet, ($ 0.614. None of these numbers exceed 0.99.

Observe that, unlike Figures S8 and S9 in the Supplementary text for JTK and RAIN algorithms, none of the scatter points in the right panel of Figure S7 for ORIOS take negative values, except for one, thus indicating that $M_{Robust}(n, ORIOS)$ almost always takes positive values for all normalization methods n. However, $M_{Robust}(n, JTK)$ and $M_{Robust}(n, RAIN)$ take negative values. Since $M_{Robust}(n, a) = \hat{E}(\hat{\theta}(n, a)) - \widehat{RMSE}(\hat{\theta}(n, a)$, therefore the variability in p-values for tests for rhythmicity using JTK and RAIN methods is larger than the corresponding estimated p-values. Thus the JTK and RAIN methods produce p-values that are subject to higher variation and uncertainty than the expected p-values. This is in sharp contrast to ORIOS which almost always produced p-values subject to smaller variability than the expected p-values. This is one more reason, in addition to the results provided in Larriba *et al.* (2016), to prefer ORIOS as the method for detecting rhythmic genes.

The bootstrap methodology introduced in this paper is computationally efficient. For the data analyzed in this paper, the method required ~ 70 minutes of CPU time to generate and process 45101 probe sets on Windows 7 Professional 3.60 GHz dual processors computer with disk space using 100 bootstrap samples.

From our investigation of real data and the bootstrap simulated data, we find that our bootstrap procedure provides a simple and a convenient way to simulate oscillatory signals that potentially resemble realistic patterns of expression. Thus, as a secondary contribution, in this paper we introduced a bootstrap methodology that not only provides methodology to detect rhythmic genes but it also allows researchers to conduct simulation studies to generate realistic rhythmic patterns. Notice also that, although for illustration and clarity purposes, in this paper we focused on gene expression studies (such as microarray and RNA-seq), the methodology described here is applicable to any modern high throughput technology involving oscillatory systems. For example, it can potentially be used for analyzing continuous time microbiome data, such as those obtained in Caporaso *et al.* (2011).

Conflict of Interest Statement

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

CR, MAF and SDP conceived aims, conceptual design, data analysis, interpretation of the results, wrote and approved manuscript. YL processed original data, generated simulations, analyzed data, interpreted results, wrote and approved manuscript.

Funding

This work was supported by Spanish Ministerio de Ciencia e Innovación grant [MTM2015-71217-R to C.R. and M.A.F.]; Spanish Ministerio de Educación, Cultura y Deporte [FPU14/04534 to Y.L.] and the Intramural Research Program of the National Institute of Environmental Health Sciences [Z01 ES101744-04 to S.D.P.].

Acknowledgements

The authors thank Drs. Keith Shockley (NIH) and Oscar M. Rueda (University of Cambridge) for their valuable comments on an earlier draft.

References

Åstrand, M. (2003) Contrast normalization of oligonucleotide arrays. *J. Comput. Biol.* 10, 95-102.

Barragán, S., Rueda, C., Fernández, M.A. and Peddada, S.D. (2015) Determination of temporal order among the components of an oscillatory system. *PLoS ONE.* 10:e0124842. doi:10.1371/journal.pone.0124842.

Benzécri, J.P. (1979) Sur le calcul des taux d'inertie dans l'analyse d'un questionnaire. *Cahiers de l'Analyse des Données.* 4, 377-378.

Bolstad, B.M., Irizarry, R.A., Åstrand, M. and Speed, T.P. (2003) A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics.* 19, 185-193.

Caporaso, J.G., Lauber, C.L., Costello, E.K., Berg-Lyons, D., Gonzalez, A., Stombaugh, J., Knights, D., Gajer, P., Ravel, J., Fierer, N., Gordon, J.I., Knight, R. (2011) Moving pictures of the human microbiome. *Genome Biol.*. 12, R50. doi:10.1186/gb-2011-12-5-r50.

Caretta-Cartozo, C., De Los Rios, P., Piazza, F. and Liò, P. (2007) Bottleneck genes and community structure in the cell cycle network of S. pombe. *PLoS Comput. Biol.* 3:e103. doi:10.1371/journal.pcbi.0030103.

Chen, L. and Yang, G. (2015) Recent advances in circadian rhythms in cardiovascular system. *Front. Pharmacol.* 6:71. doi:10.3389/fphar.2015.00071.

Cheng, L., Lo, L.Y., Tang, N.L.S., Wang, D. and Leung, K.S. (2016) CrossNorm: A novel normalization strategy for microarray data in cancers. *Sci. Rep.* 6:18898. doi:10.1038/srep18898.

Dembélé, D. (2013) A Flexible Microarray Data Simulation Model. *Microarrays.* 44, 115-130.

Efron, B. and Tibshirani, R.J. (1994) *An Introduction to the Bootstrap.* Boca Raton: Chapman & Hall/CRC.

Emerson, J.D. and Hoaglin, D.C. (1983) *Analysis of two-way tables by medians. In Understanding Robust and Exploratory Data Analysis.* New York: John Wiley & Sons.

Freudenberg, J., Boriss, H. and Hasenclever, D. (2004) Comparison of preprocessing procedures for oligo-nucleotide micro-arrays by parametric bootstrap simulation of spike-in experiments. *Method. Inform. Med.* 43, 434-438.

Greenacre, M.J. (1984). *Theory and applications of correspondence analysis.* London: Academic Press.

Hubbell, E., Liu, W.M. and Mei, R. (2002) Robust estimators for expression analysis. *Bioinformatics.* 18, 1585-1592.

Huber, W., Von Heydebreck, A., Sültmann, H., Poustka, A. and Vingron, M. (2002) Variance stabilization applied to microarray data calibration and to the quantification of differential expression. *Bioinformatics.* 18, 96-104.

Hughes, M.E., DiTacchio, L., Hayes, K.R., Vollmers, C., Pulivarthy, S., Baggs, J.E, Panda, S. and Hogenesch, J.B. (2009) Harmonics of circadian gene transcription in mammals. *PLoS Genet.* 5:e1000442. doi:10.1371/journal.pgen.1000442.

Hughes, M.E., Hogenesch, J.B. and Kornacker, K. (2010) JTK-CYCLE: An efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J. Biol. Rhythm.* 25, 372-380.

Irizarry, R.A., Bolstad, B.M., Collin, F., Cope, L.M., Hobbs, B. and Speed, T.P. (2003) Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15.

Irizarry, R.A., Hobbs, B., Collin, F., Beazer-Barclay, Y.D., Antonellis, K.J., Scherf, U. and Speed, T.P. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 4, 249-264.

Klebanov, L. and Yakovlev, A. (2007) How high is the level of technical noise in microarray data?. *Biol. Direct.* 2:9. doi:10.1186/1745-6150-2-9.

Larriba, Y., Rueda, C., Fernández, M.A. and Peddada, S.D. (2016) Order restricted inference for oscillatory systems for detecting rhythmic genes. *Nucleic Acids Res.* 44:e163. doi:10.1093/nar/gkw771.

Li, C. and Wong, W.H. (2001) Model-based analysis of oligonucleotide arrays: Expression index computation and outlier detection. *P. Nati. Acad. Sci. USA.* 98, 31-36.

Liu, G., Loraine, A.E., Shigeta, R., Cline, M., Cheng, J., Valmeekam, V., Sun, S., Kulp, D. and Siani-Rose, M.A. (2003) NetAffix: Affymetrix probesets and annotations. *Nucleic Acids Res.* 31, 82-86.

Nykter, M., Aho, T., Ahdesmäki, M., Ruusuvuori, P., Lehmussola, A. and Yli-Harja, O. (2006) Simulation of microarray data with realistic characteristics. *BMC Bioinformatics.* 7:349. doi:10.1186/1471-2105-7-349.

Oliva, A., Rosebrock, A., Ferrezuelo, F., Pyne, S., Chen, H., Skiena, S., Futcher, B. and Leatherwood, J. (2005) The cell cycle-regulated genes of Schizosaccharomyces pombe. *PLoS Biol.* 3, 1239-1260.

Parrish, R.S, Spencer III, H.J. and Xu, P. (2009) Distribution Modeling and Simulation of Gene Expression Data. *Comput. Stat. Data An.* 53, 1650-1660.

Peng, X., Karuturi, R.K.M., Miller, L.D., Lin, K., Jia, Y., Kondu, P., Wang, L., Wong, L.S., Liu, E.T., Balasubramanian, M.K. and Liu, J. (2005) Identification of cell cycle-regulated genes in fission yeast. *Mol. Biol. Cell.* 16, 1026-1042.

Rustici, G., Mata, J., Kivinen, K., Liò, P., Penkett, C.J., Burns, G., Hayles, J., Brazma, A., Nurse, P. and Bähler, J. (2004) Periodic gene expression program of the fission yeast cell cycle. *Nat. Genet.* 36, 809-817.

Slavov, N., Airoldi, E.M., Van Oudenaarden, A. and Botstein, D. (2012) A conserved cell growth cycle can account for the environmental stress responses of divergent eukaryotes. *Mol. Biol. Cell.* 23, 1986-1997.

Thaben, P.F. and Westermark, P.O. (2014) Detecting rhythms in time series with rain. *J. Biol. Rhythm.* 29, 391-400.

Tu, Y., Stolovitzky, G. and Klein, U. (2002) Quantitative noise analysis for gene-expression microarray experiments. *P. Nati. Acad. Sci. USA.* 99, 14031-14036.

Workman, C., Jensen, L.J., Jarmer, H., Berka, R., Gautier, L., Nielser, H.B., Saxild, H.H., Nielsen, C., Brunak, S. and Knudsen, S. (2002) A new non-linear normalization method for reducing variability in DNA microarray experiments. *Genome Biol.* 3, research0048.1-research0048.16.

Wu, Z. (2009) A review of statistical methods for preprocessing oligonucleotide microarrays. *Stat. Methods Med. Res.* 18, 533-541.

Zhang, R., Lahens, N.F., Ballance, H.I., Hughes, M.E. and Hogenesch, J.B. (2014) A circadian gene expression atlas in mammals: implications for biology and medicine. *P. Natl. Acad. Sci. USA.* 111, 16219-16224.

Figure captions

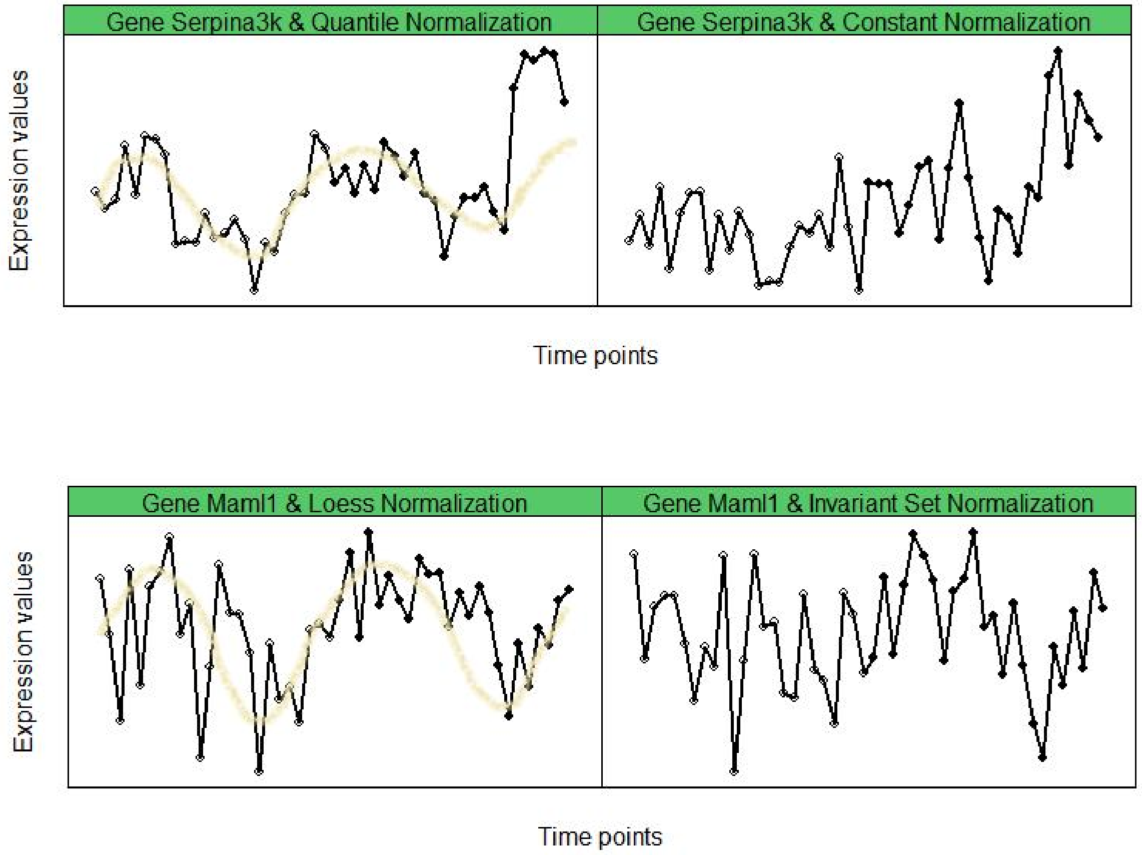


Figure S1: Time-course gene-expression for genes *Serpina3k* (top) and *Maml1* (bottom). *Serpina3k* and *Maml1* are identified as rhythmic by ORIOS according to *Quantile* and *Loess* normalizations, respectively. But they are identify as non-rhythmic by ORIOS for *Constant* and *Invariant Set* normalizations, respectively.

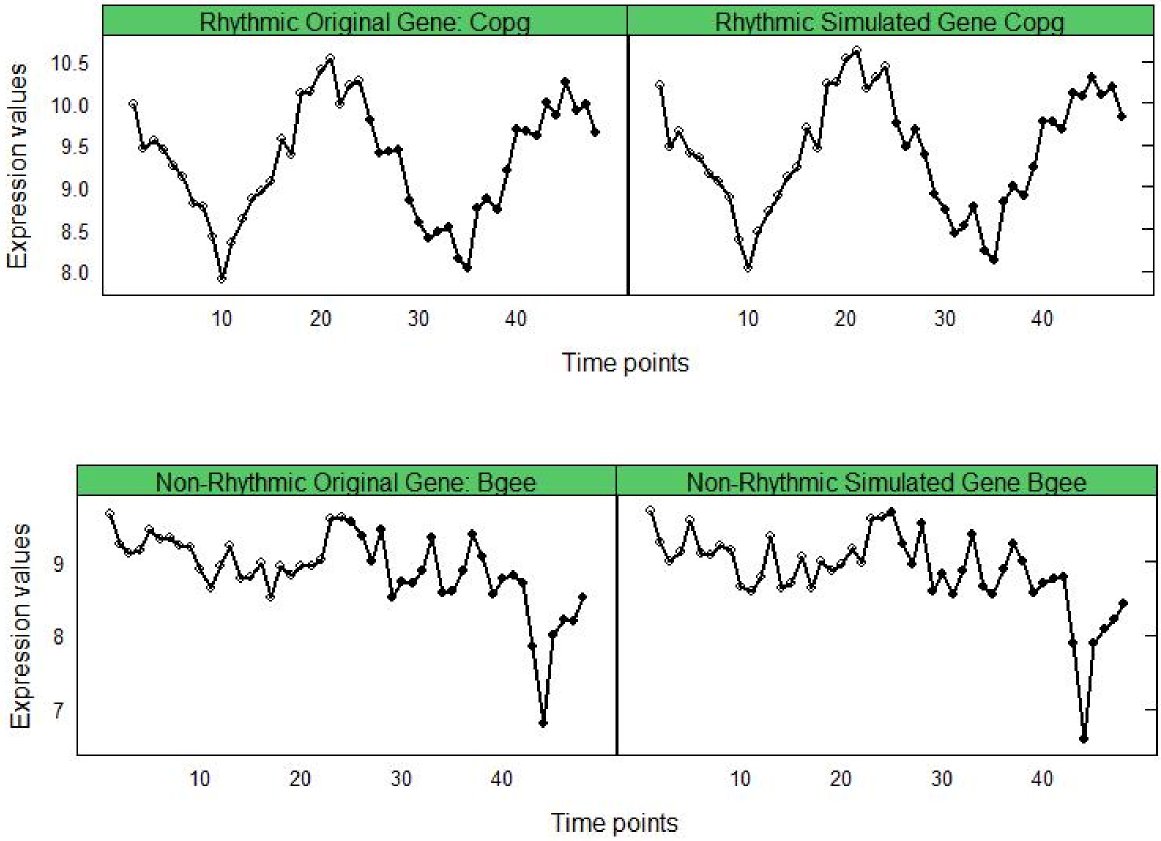
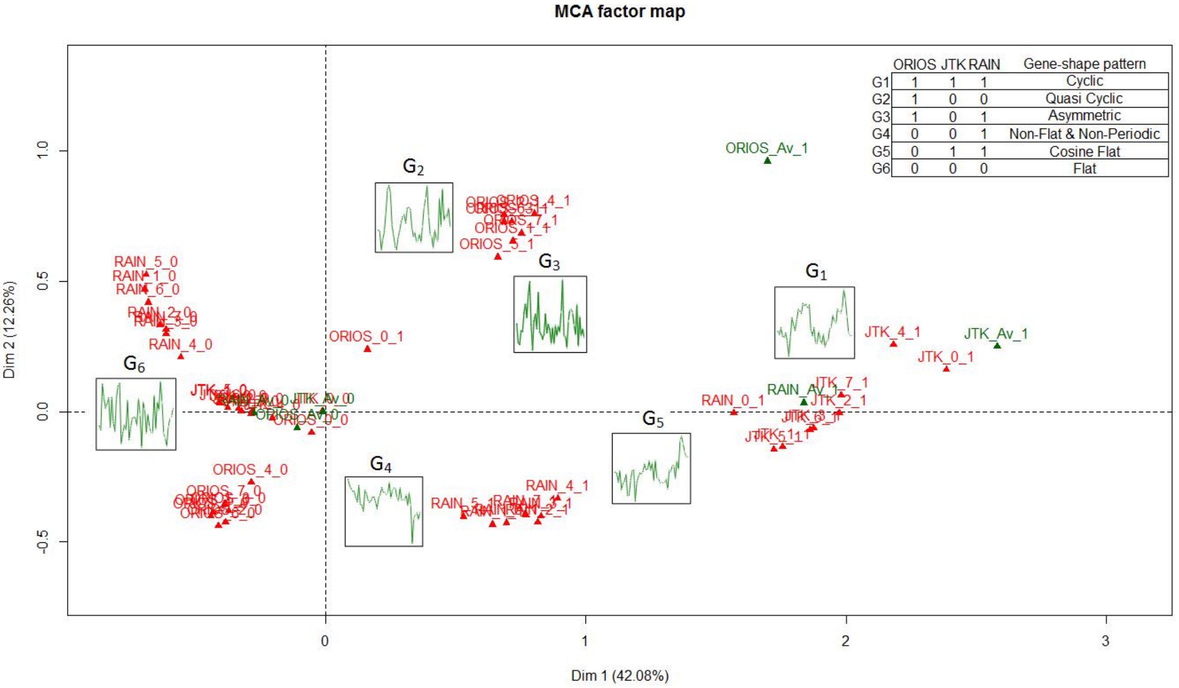
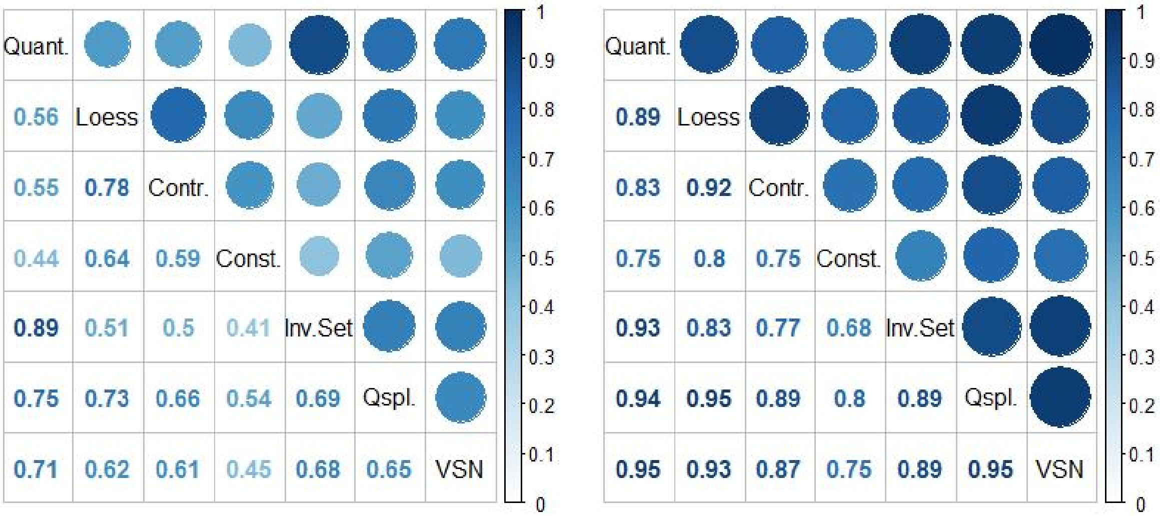


Figure S2: Original vs Simulated gene-expression for genes *Copg* (top) and *Bgee* (bottom). Left: corrected (and unnormalized) gene-expression from the reference dataset (*mouse liver tissue*). Right: simulated gene-expression attained after bootstrapping.



	ORIOS	JTK	RAIN	Gene-shape pattern
G1	1	1	1	Cyclic
G2	1	0	0	Quasi Cyclic
G3	1	0	1	Asymmetric
G4	0	0	1	Non-Flat & Non-Periodic
G5	0	1	1	Cosine Flat
G6	0	0	0	Flat

Figure S3: Multiple Correspondence Analysis factor map for the different gene profiles under all normalizations and algorithms considered, together with the averaged rhythmic and non-rhythmic profiles for each algorithm and the six gene patterns defined in the table.



Quant.						
0.56	Loess					
0.55	0.78	Contr.				
0.44	0.64	0.59	Const.			
0.89	0.51	0.5	0.41	Inv.Set		
0.75	0.73	0.66	0.54	0.69	Qspl.	
0.71	0.62	0.61	0.45	0.68	0.65	VSN

Quant.						
0.89	Loess					
0.83	0.92	Contr.				
0.75	0.8	0.75	Const.			
0.93	0.83	0.77	0.68	Inv.Set		
0.94	0.95	0.89	0.8	0.89	Qspl.	
0.95	0.93	0.87	0.75	0.89	0.95	VSN

Figure S4: Spearman rank correlation coefficients between $((M^g(n_i, ORIOS), M^g(n_j, ORIOS))$ (left) and between $((M^g_{Robust}(n_i, ORIOS), M^g_{Robust}(n_j, ORIOS))$ (right) for all pairs of normalization procedures n_i and n_j using the 15369 probe sets.

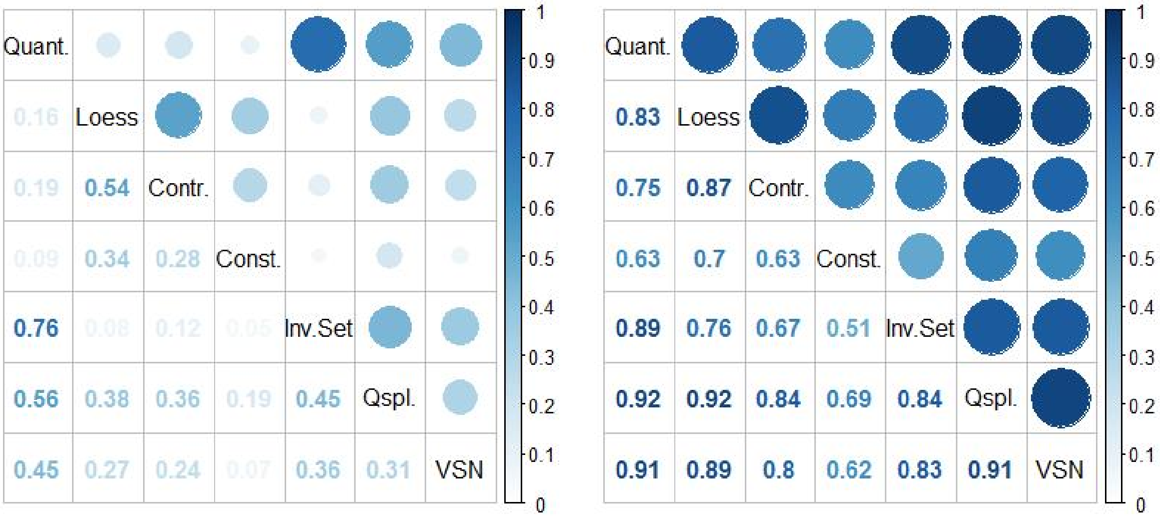


Figure S5: Pearson correlation coefficients between $((M^g(n_i, ORIOS), M^g(n_j, ORIOS))$ (left) and between $((M^g_{Robust}(n_i, ORIOS), M^g_{Robust}(n_j, ORIOS))$ (right) for all pairs of normalization procedures n_i and n_j using the 15369 probe sets.

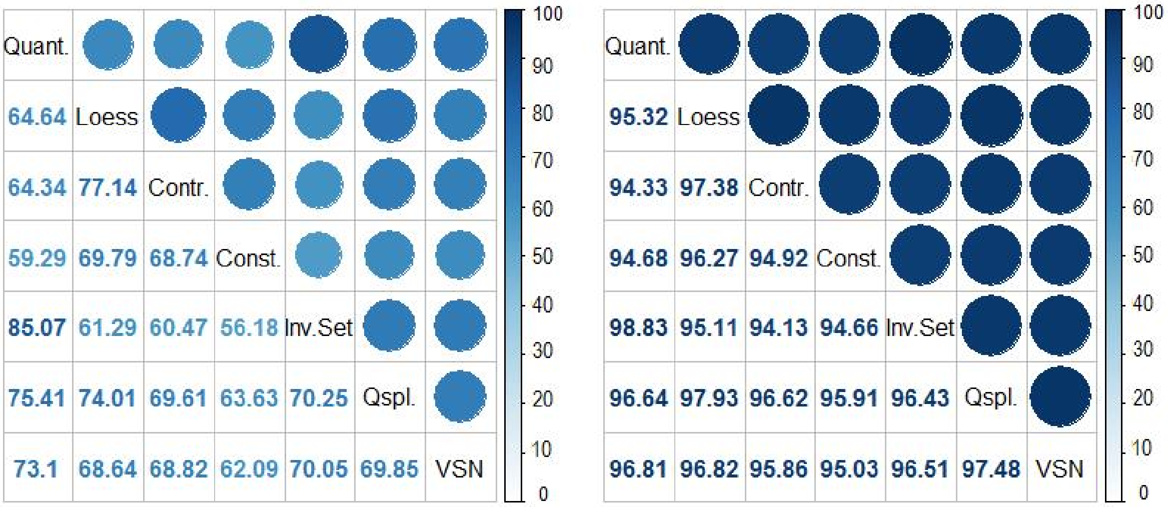
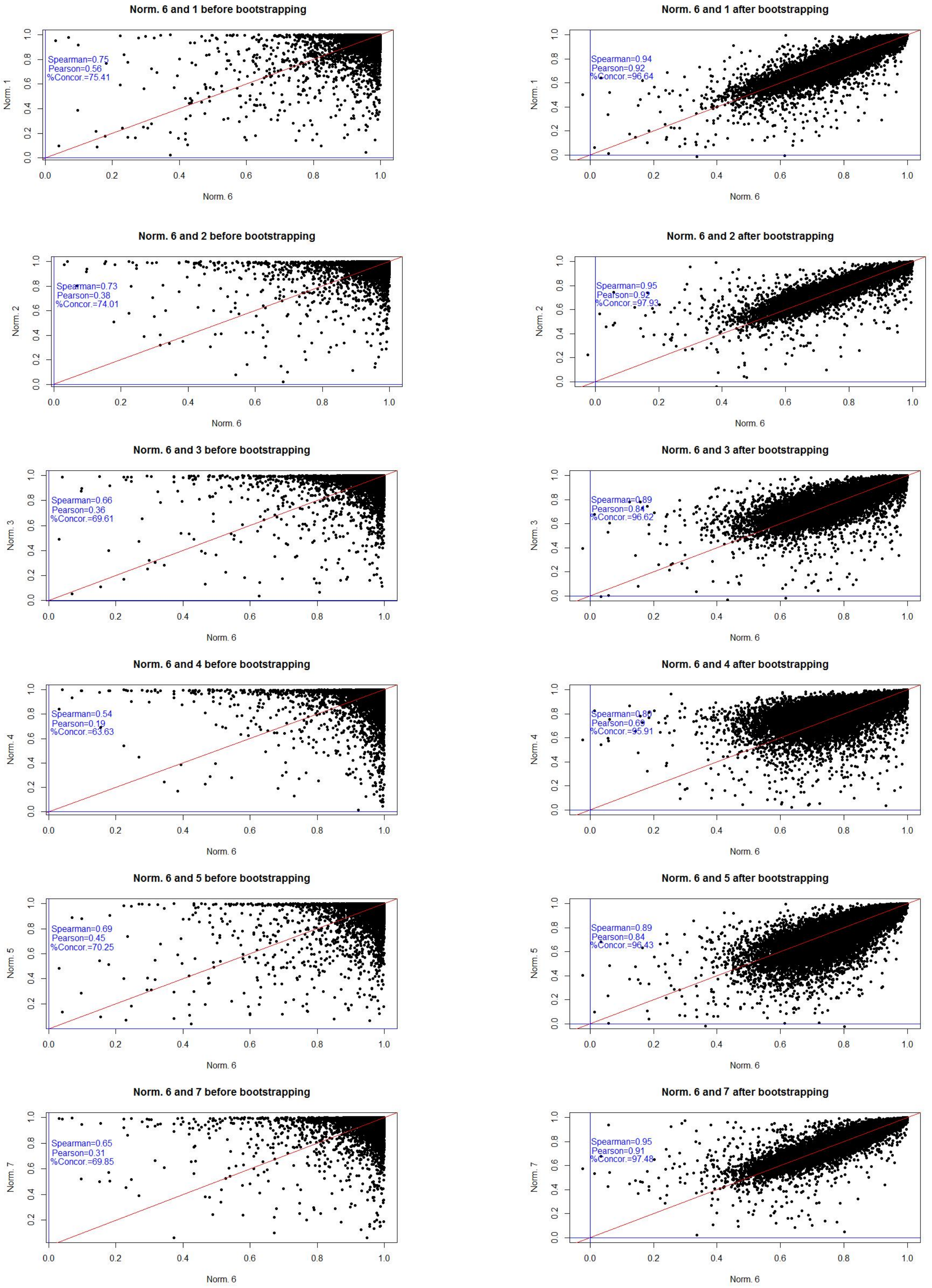


Figure S6: Percentage of (rhythmic and non-rhythmic) concordant probe sets before (left) and after (right) bootstrapping for all pairs of normalization procedures n_i and n_j using the 15369 probe sets.


Norm. 6 and 1 before bootstrapping

Norm. 6 and 1 after bootstrapping

Norm. 6 and 2 before bootstrapping

Norm. 6 and 2 after bootstrapping

Norm. 6 and 3 before bootstrapping

Norm. 6 and 3 after bootstrapping

Norm. 6 and 4 before bootstrapping

Norm. 6 and 4 after bootstrapping

Norm. 6 and 5 before bootstrapping

Norm. 6 and 5 after bootstrapping

Norm. 6 and 7 before bootstrapping

Norm. 6 and 7 after bootstrapping

Figure S7: For each normalization method n, the left panels represent the pairwise scatter plots of $(M^g(n, ORIOS), M^g(Qspline, ORIOS))$ and the right panels represent the pairwise scatter plots of $(M^g_{Robust}(n, ORIOS), M^g_{Robust}(Qspline, ORIOS))$. Red line is the 45^o diagonal and the blue lines are the Cartesian axes.