Journal of Bioinformatics and Computational Biology
Vol. 10, No. 6 (2012) 1271002 (19 pages)
© Imperial College Press
DOI: 10.1142/S0219720012710023

Imperial College Press
www.icpress.co.uk

DETECTION AND DECOMPOSITION: TREATMENT-INDUCED CYCLIC GENE EXPRESSION DISRUPTION IN HIGH-THROUGHPUT TIME-SERIES DATASETS

YUHUA JIAO[*,¶,¶¶], BRUCE A ROSA[†,‖,¶¶], SOOKYUNG OH[*,**], BERONDA L MONTGOMERY[*,‡,††], WENSHENG QIN[†,‡‡,‖‖] and JIN CHEN[*,§,§§,‖‖]

[*]*MSU-DOE Plant Research Laboratory*
Michigan State University
East Lansing, MI 48824, USA

[†]*Biorefining Research Initiative and Department of Biology*
Lakehead University, Thunder Bay, ON P7B 5E1, Canada

[‡]*Department of Biochemistry and Molecular Biology*
Michigan State University
East Lansing, MI 48824, USA

[§]*Department of Computer Sciences and Engineering*
Michigan State University
East Lansing, MI 48824, USA
[¶]*yuhjiao@msu.edu*
[‖]*barosa@lakeheadu.ca*
[**]*ohsookyu@msu.edu*
[††]*montg133@msu.edu*
[‡‡]*wqin@lakeheadu.ca*
[§§]*jinchen@msu.edu*

Received 30 July 2012
Revised 6 August 2012
Accepted 7 August 2012
Published 12 October 2012

Higher organisms possess many genes which cycle under normal conditions, to allow the organism to adapt to expected environmental conditions throughout the course of a day. However, treatment-induced disruption of regular cyclic gene expression patterns presents a significant challenge in novel gene discovery experiments because these disruptions can induce strong differential regulation events for genes that are not involved in an adaptive response to the treatment. To address this cycle disruption problem, we reviewed the state-of-art periodic pattern detection algorithms and a pattern decomposition algorithm (PRIISM), which is a knowledge-based Fourier analysis algorithm designed to distinguish the cyclic patterns from the rest gene expression patterns, and discussed potential future improvements.

Keywords: Cyclic gene expression; time-series; pattern detection; pattern decomposition.

[¶¶] The authors are equal contributors to this paper.

[‖‖] Corresponding authors.

1. Disruption of Cyclic Gene Expression Patterns

Cyclic gene expression patterns have been observed in almost every organism ranging from prokaryotes to eukaryotes,[1] and have been found to provide a competitive advantage in fitness and survival.[2–4] Whole-genome differential expression analyses have enabled scientists to investigate how maintaining or disrupting a rhythmic mechanism creates an adaptive advantage for an organism. One of the findings from such studies is that treatment-induced disruption of the core set of cyclic genes (which control many downstream pathways) occurs in almost all kinds of organisms, including mamals, plants and prokayotes (Fig. 1).[3–8]

One very important problem which is not typically considered in gene expression studies which utilize a biotic or abiotic treatment (or analyze organisms under different environmental conditions) is that strong positive and negative fold-change values can result from small disruptions in cyclic gene expression patterns, where a fold-change value is calculated as the treatment expression level divided by the control expression level at the same timepoint. In an illustrative example in Figs. 2(a)–2(f), a 2-hours phase-shift in cyclic expression results in strong positive and negative differential regulation cycles; a 33% increase in the minimum amplitude (relative to the maximum amplitude) results in significant peaks at several timepoints; and a 10% reduction in cycle frequency leads to varying differential regulation at various timepoints. Figures 2(g)–2(j) show real examples of the gene expression patterns and fold change curves resulting from the disruptions of two core circadian clock genes due to cold treatment in *Arabidopsis*.[9]

When a stress treatment is applied, stress-response genes are expected to be differentially regulated, while influences from the disrupted circadian clock will cause significant fold changes in gene expression because genes can be (I) differentially regulated due to direct stress responses, (II) indirectly differentially regulated through disruption of clock pathways induced by the stress or (III) differentially regulated through a combination of both (as shown in Fig. 3). Additional complications in regulation patterns arise from the complexity of transcription factor pathways, in which targets may be regulated by clock components directly or through interactions with their transcription factors. These disruptions in core clock components complicate the identification of "true" treatment-response genes (i.e. genes which physiologically participate in the biological acclimation to treatment). Correcting for differential genes expression patterns induced by the circadian clock disruption is required for the discovery of "true" treatment-response gene.

Up to 10% of mammalian genes[11] are thought to be under the regulatory control of molecular circadian rhythm pathways. The large number of physiological and pathological consequences of disrupting these circadian pathways through sleep deprivation or unnatural light regimes have been studied in detail on whole organisms,[12] and large changes in the oscillation patterns of all the tested core clock genes have been observed for mice under different feeding regimes.[11] Mammalian gene expression studies are more typically performed on cell cultures, which have been

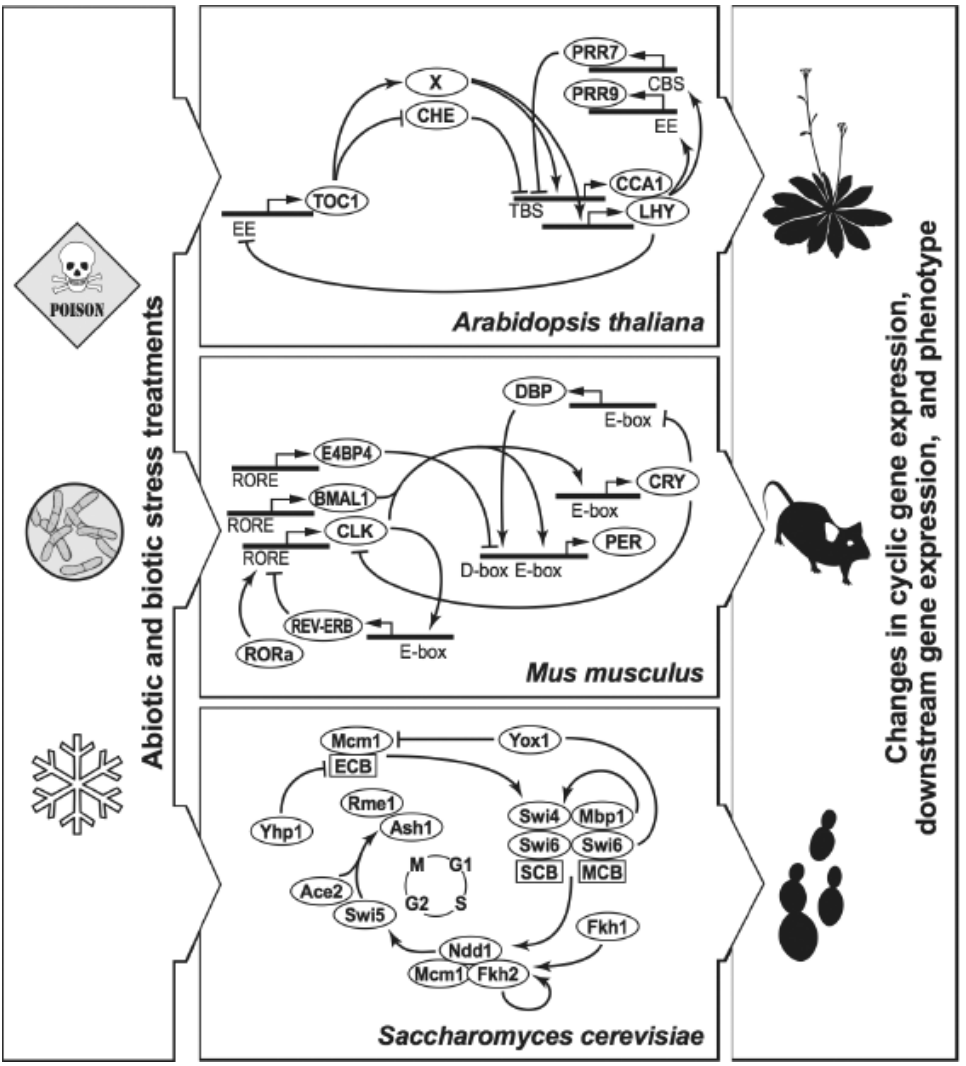

Fig. 1. Detailed structures of the clock mechanisms for *Mus musculus*, *Arabidopsis thaliana* and *Saccharomyces cerevisiae* have been generated through genetic studies. While the components are not orthologous, the schematic model diagrams of circadian-clock and cell-cycle oscillators of the three model organism are similar to each other, and the network architecture of positive and negative feedback loops is conserved.[4]

shown to have molecular circadian rhythm patterns similar to whole organisms.[13] Recently, a protein kinase called c-Jun NH2-terminal Kinase (JNK) was shown to be critical for proper circadian rhythm function and cell proliferation [through Map Kinase Kinase 7 (MKK7) activation] and is activated by a wide range of external stresses, including osmolarity changes, heat shock and UV irradiation.[5] This suggests a mechanism through which treatment-induced circadian-pathway disruptions may

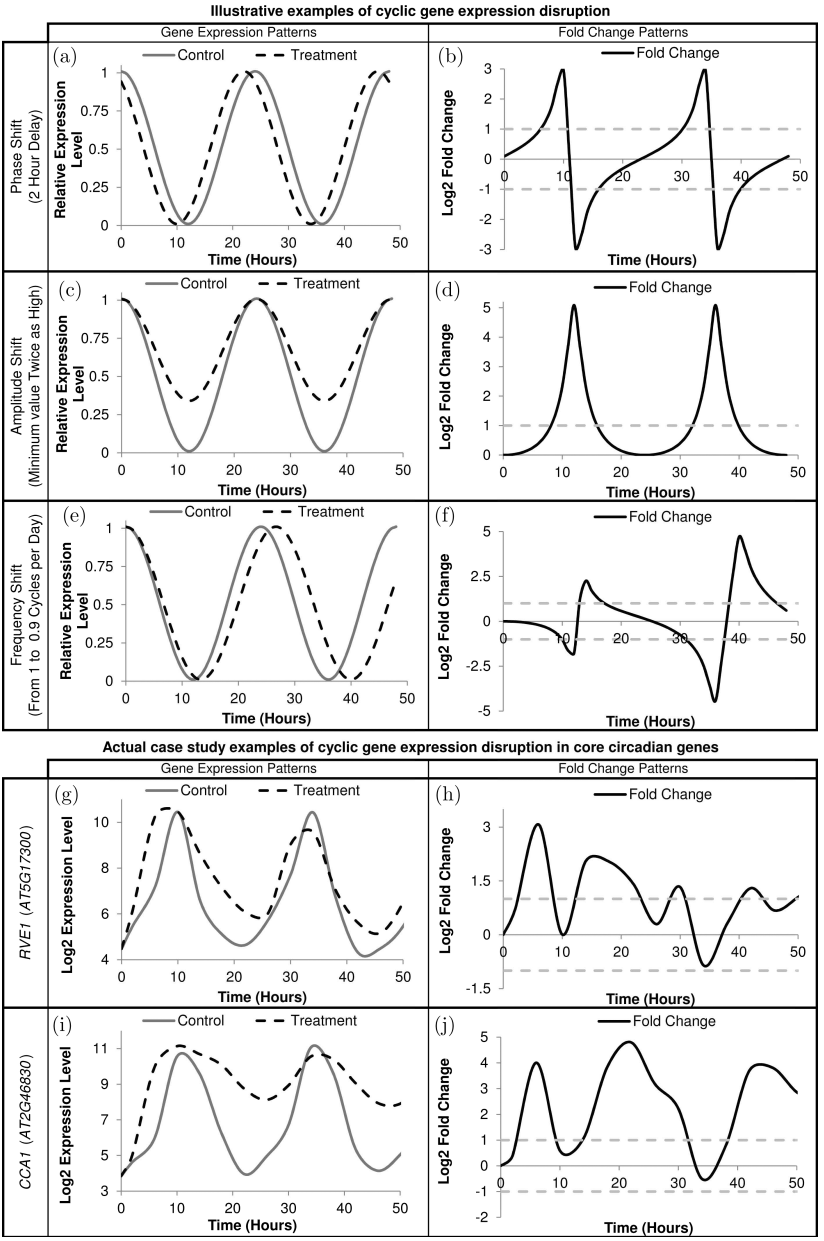

Fig. 2. Strong fold change values can result from small changes in cyclic expression patterns. (a) A 2 hours phase-shift (8.3% of the total cycle time) in cyclic expression results in (b) strong positive and negative differential regulation cycles. (c) Increasing the minimum amplitude by 33% (relative to the maximum amplitude) results in (d) significant peaks at several timepoints. (e) A 10% reduction in cycle frequency leads to (f) strong positive and negative differential regulation at various timepoints. (g) A phase-shift and small increase in the minimum amplitude of *RVE1* due to cold treatment results in (h) significant upregulation fold change values at several timepoints. (i) An increase in the minimum amplitude of *CCA1* results in (j) significant upregulation fold change values at several timepoints.[9]

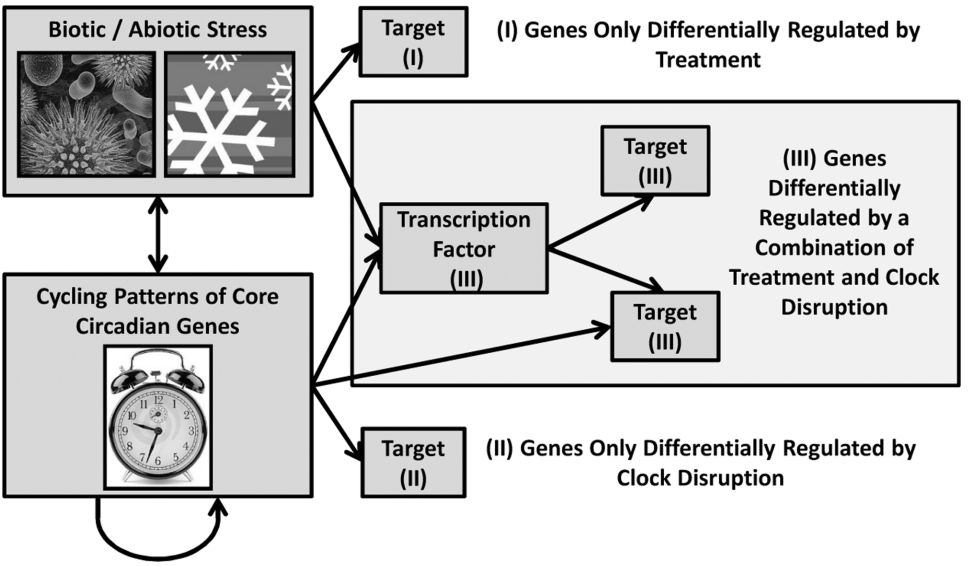

Fig. 3. Biotic and abiotic stresses both directly and indirectly influence target gene expression patterns. Genes found to be differentially expressed may be influenced by (I) only direct treatment influences, (II) only indirect circadian-clock disruption influences, or (III) both direct treatment response and indirect clock influences.[10]

occur, causing differential expression measurements for a large number of genes in treatment-response studies in mammalian cell cultures.

Plants also possess molecular circadian clock pathways which influence gene expression to modify physiology and metabolism in preparation for predictable changes in light and temperature conditions in the environment.[14] Plants with circadian clocks that are properly synchronized to their environments have been found to fix more carbon, grow larger and survive better than plants with clocks that are out of phase with their environments.[2] Several studies have shown that between 6% and 31% of the *Arabidopsis* genome is influenced by circadian clock genetic components,[6,15,16] while another study suggests that there are significant baseline circadian oscillations for nearly 100% of the genome.[17] Many different biotic and abiotic stress treatments in *Arabidopsis* studies have been shown to disrupt rhythmic circadian-clock expression patterns through amplitude changes or phase and amplitude shifts.[6,9,18–21]

In addition to circadian rhythm pathways, the cyclic expression of cell-cycle genes in synchronized cultures may be disrupted by treatment responses. In one study, 13% of genes in cell-cycle synchronized yeast cultures were found to have significant cyclic patterns,[22] while another study found that at least 27% of yeast genes are linearly correlated with growth rates.[23] Two separate studies have found that many genes previously characterized as being stress responsive[7] are likely actually differentially regulated in response to a reduction in growth rate secondary to stress,[23,24]

suggesting a similar complication in cell-cycle regulated genes as described above for circadian genes.

2. Biological Solutions to Address the Cycle Disruption Problem

In principle, there are two approaches to address the cycle disruption problem in high-throughput gene expression experiments. First, the experimental design may be changed in order to attempt to reduce the influences of cyclically expressed genes on global gene expression. To remove the influences of circadian-pathway genes, constant light treatments are sometimes used in an attempt to avoid the influences of cyclic expression patterns, but circadian rhythms continue under constant light, treatment responses often depend of the phase of the clock at which they are applied, and the unnatural conditions reduce the applicability of the results.[9,25] Another approach to reducing clock-related influences is to use clock-gene knockout strains, but this also complicates the results due to the unnatural conditions and specific effects resulting from changes in the expression of clock components.[9,26]

Even though these experimental-design approaches may remove the cyclic patterns from circadian genes, the unnatural conditions required to accomplish this will always reduce the applicability of the results. For this reason, time-series experiments are typically run on synchronized samples by entraining circadian rhythms using light and/or temperature cycle[5,9] and leave the problem of cycle disruption identification and removal to post-experimental data analysis.

3. Computational Solutions to Address the Cycle Disruption Problem

The general computational approach for addressing the cycle disruption problem is to run the experiment under natural conditions and then process the datasets in order to distinguish cyclic patterns from other gene expression patterns in the datasets. In general, the process has two stages: periodic gene expression pattern detection (see Sec. 3.1) and pattern decompostion (see Sec. 3.2). Both of these stages use time-series gene expression data as an input.

From gene expression data, cyclic patterns can be characterized by periodic pattern detection methods. Then, cyclic patterns can be decomposed from original expression data resulting in the removal of disruption influences. The framework of pattern decomposition is shown in Fig. 4, where one gene expression profile is decomposed into three distinct gene expression patterns: (1) the treatment-frequency gene expression pattern, which has much of the complicating circadian influences removed, and consequently can be used to more accurately identify differentially regulated genes which are involved in direct treatment response; (2) the clock-frequency gene expression pattern, representing rhythmic patterns with a period of approximately one cycle per day; and (3) the noise-frequency gene expression pattern.

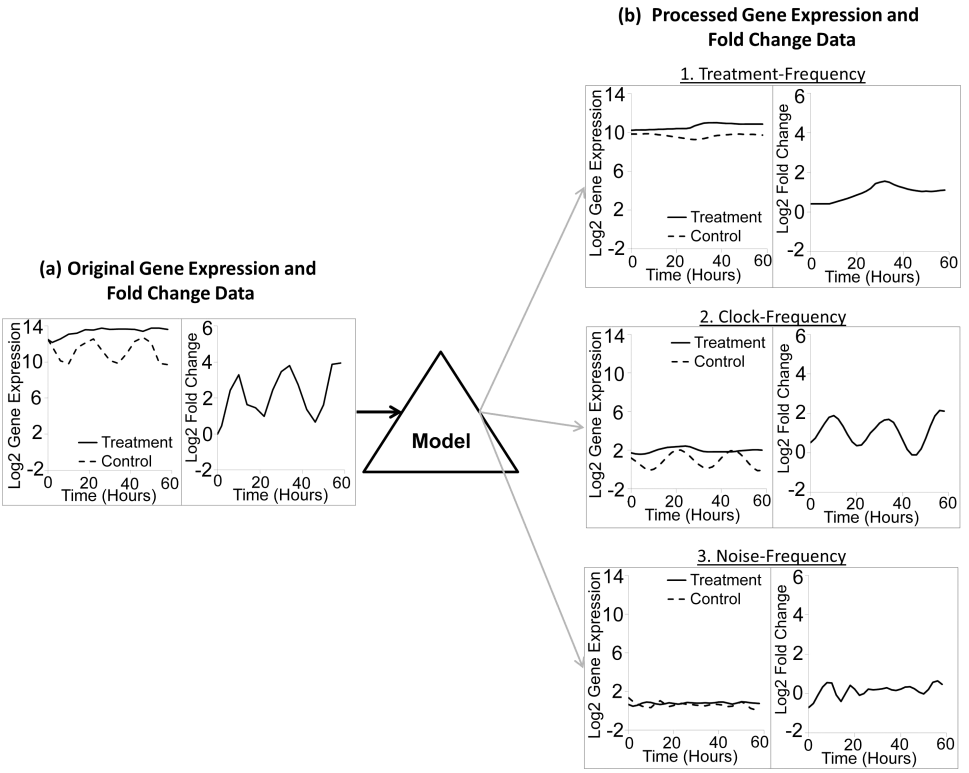

Fig. 4. An abstract framework to decompose gene expression data into three independent gene expression patterns. (a) The original gene expression data under control and treatment conditions (used to calculate the fold-change pattern). (b) Treatment-frequency, clock-frequency and noise-frequency gene expression patterns.[10]

Note that, in addition to cyclic pattern identification and decomposition, the distortion of circadian patterns has be studied with statistical hypothesis testing methods such as analysis of variance (ANOVA),[27,28] in that it is capable of estimating the overall probability of difference in expression. However, ANOVA cannot measure the circadian distortion at a given timepoint and cannot study how the distortion changes over time. Therefore, we will not discuss the use of ANOVA in the rest of this paper.

3.1. *Detecting periodic gene expression patterns*

The first stage in addressing cycle disruption is to characterize cyclic patterns in time-series gene expression datasets. The existing approaches for detecting periodic expression patterns are generally categorized into two groups: model-based pattern identification in time domain and spectral analyses in frequency domain.[29]

Many time-series approaches detect periodic gene expression patterns by matching expression profiles to periodic models.[16,29–32] One such approach,

CORRCOS[16] presented in 2000, computed cross-correlations between gene expression data and synthetic cosine functions. A CORRCOS-based study of more than 8000 *Arabidopsis* genes suggested that around 6% of them are controlled by the clock, many of which have been previously reported to be under circadian control.[16] As an extension of CORRCOS, COSOPT[30] (which provides a reliable estimation of period, phase and oscillatory amplitude[33]) was first applied to mammalian systems,[30] and then extended to the analyses of circadian patterns in microarray data from other genomes, including *Drosophila*[34] and *Arabidopsis*.[15] In both algorithms, sine and cosine curves are used as the ideal model for periodic gene expressions.[35] However, as reported in Lin *et al.*,[36] a significant number of expression profiles are actually not sinusoidal. To address this issue, Luan and Li[31] developed a shape-invariant model in 2004 with a cubic B-spline based periodic function for characterizing cyclic patterns in gene expression data. The proposed model successfully identified 86% of the known periodically expressed genes in a yeast cell cycle dataset[37] at a false discovery rate of 0.5%. Based on these previous works, HAYSTACK[32] (which includes multiple cyclic patterns including asymmetric, rigid, spike, cosine and/or box-like patterns) was developed by Mockler *et al.* in 2007. By computing the Pearson correlation coefficients between gene expression profiles and user-defined models with a wide range of types of cyclic patterns, HAYSTACK identifies periodically expressed genes and their corresponding best-fitting model with the phase and the period information.[32] The above methods are relatively simple and computationally efficient but are limited because the models must be pre-defined and may not match any true biological patterns. To overcome this limitation, in 2009, Chudova *et al.* applied a Bayesian procedure to detect nonsinusoidal periodic patterns in circadian expression data by estimating the contribution of a periodic component in observed profiles.[29]

The second category of approaches for detecting periodic expression patterns is spectral analyses in frequency domain. In frequency-domain analyses, the process of periodic pattern detection has been simplified to the identification of the significant peaks in the amplitude spectrum. To obtain the spectrum of the expression profile, frequency-domain approaches perform transformations on the time-series expression profile of each probe. A Fourier transform−based study, which decomposes periodic signals into the sum of a set of simple oscillating functions (complex exponentials), was first presented by Spellman *et al.*[37] to analyze time-series gene expression data, resulting in the identification of 800 cell-cycle genes in synchronized *S. Cerevisiae* cultures. It was later extended to successfully detect cyclic genes in human cell cultures,[38] *Plasmodium falciparum*[39] and fission yeast.[40] However, there are several limitations to the application of the Fourier transform. First, it requires evenly spaced data, which are not usually available for high-throughput gene expression experiments. A more critical limitation of the application of the Fourier transform is the issue of the frequency resolution (determined by the sampling interval in a gene expression experiment). Existing time-series datasets are usually short, and consequently their frequency resolutions are not high enough to distinguish periodic

pattens of interest.[41,42] In short, Fourier transform−based frequency domain methods are model independent, but they are limited because they require evenly and densely sampled time-series datasets that are not commonly available cases for high-throughput gene expression profiles. To address the first limitation, the Lomb−Scargle periodogram[43] and Laplace periodogram[44] were proposed to treat unevenly spaced timepoints. Realizing the second limitations of the Fourier transform, Wichert *et al.*[45] developed a new graphical device called the "average periodogram," a nonparametric method of obtaining frequency estimators, which is suitable to detect periodic pattern in very short time-series expression data.

The aforementioned approaches are able to characterize cyclic patterns in gene expression data but they are not designed to distinguish cycle disruption from treatment-induced gene expression patterns. Decomposing these influences to investigate only treatment-induced changes is a much more complex problem than identifying cyclic genes, and as discussed in Sec. 1, is very important for understanding biological responses to treatment.[4]

3.2. *Distinguishing cyclic patterns from other gene expression patterns*

For distinguishing cycle disruption patterns from treatment-induced gene expression patterns, we have previously presented a frequency-based algorithm called PRIISM.[10] PRIISM has three steps. In step one, it takes advantage of the Fourier transform to characterize the periodic patterns in expression profiles. In step two, a clock vector is derived based on the Fourier spectra of core circadian genes. In step three, the input expression profile is decomposed into three components by using a set of filters which are defined according to the clock vector. The workflow of PRIISM is outlined in Fig. 5.

PRIISM was evaluated on a relatively high-resolution time-series dataset[9] of *Arabidopsis* under cold treatment. Results of this study showed that the treatment-related fold change data produced by PRIISM constantly outperforms the original data, and the 26 hours timepoint in our dataset was the best statistic for identifying the most known cold-response genes. In addition, six novel cold-response genes were discovered. PRIISM also provides a gene expression pattern which represents only circadian clock influences, and may be useful for circadian clock analysis studies.

3.2.1. *PRIISM Step 1: Fourier transform*

The treatment-induced disruption of clock patterns may change over time as the organism adapts to or recovers from the treatment applied. So rather than simply applying Fourier transform on the whole time course data, a coarse sliding-window approach is utilized in PRIISM to capture the time variant frequencies, i.e. the whole time course is divided into overlapped timeframes such that each timeframe covers roughly a one day and one night cycle.

As discussed in the previous section, the Fourier transform requires evenly sampled input. To meet this need, the input data in each timeframe are interpolated to

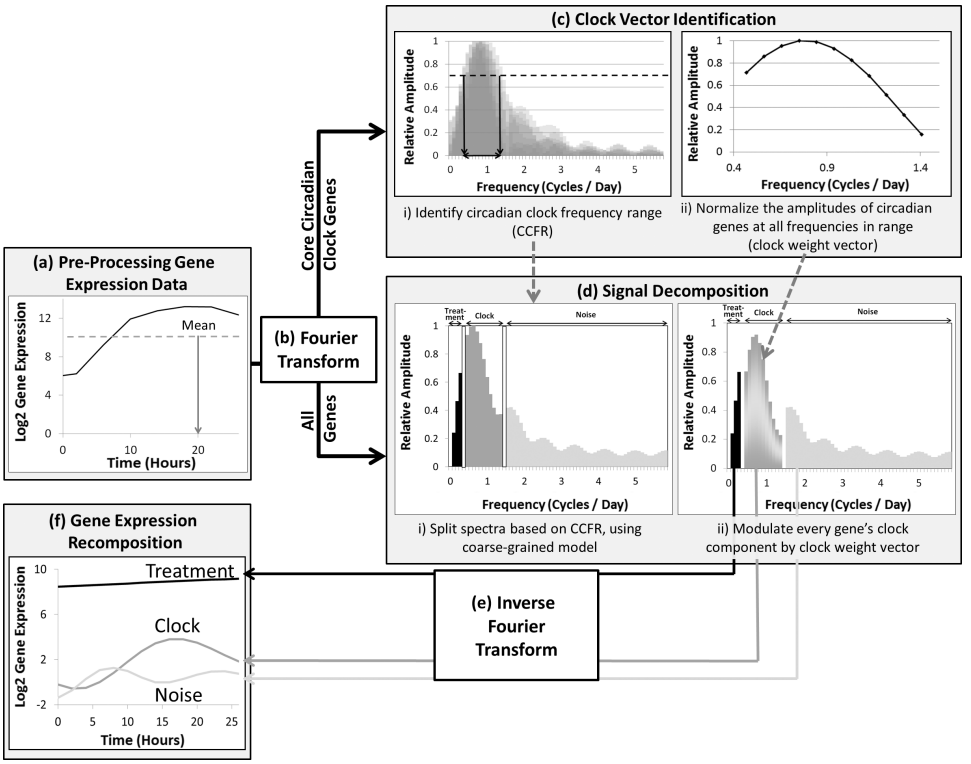

Fig. 5. Workflow of the PRIISM algorithm. PRIISM has three steps. In the first step, gene expression data are pre-processed to fit the requirements of the Fourier transform, after which the Fourier transform is performed to produce an amplitude spectrum for every gene (a, b). In the second step, a clock vector that defines the frequency range and the frequency response of the filters which are used to decompose spectrum. In the third step, every gene's spectrum is decomposed into three components: treatment, clock and noise, after which the inverse Fourier transform is applied to project each spectrum component back to the expression domain, resulting in three independent expression patterns (e, f).[10]

evenly spaced intervals with spline interpolation. This interpolation step also decreases the sampling interval of the data, which enhances the frequency resolution of the Fourier analysis.

The Fourier transform is then performed on all the genes. To avoid the periodicity identification being biased toward zero frequency (due to some constant minimum expression level for most genes), the mean of the time course expression values for each gene is shifted to zero before the Fourier transform. The mean values are later added back proportionally to the reconstructed gene expression values in the reconstruction step.

3.2.2. *PRIISM Step 2: Clock vector identification*

In PRIISM, core cyclic genes are taken as guide genes that specify periodic patterns of expressions under control and treatment. For the PRIISM test experiment on

Arabidopsis, eight genes (*CCA1*, *LHY*, *PRR7*, *PRR9*, *ELF4*, *GI*, *LUX* and *TOC1*) were chosen as core circadian clock genes because they and their downstream gene targets are known to regulate a wide range of downstream pathways, including germination, leaf development, organelle morphology, photosynthesis and cell wall development in plants.[46] For other organisms, similar core clock genes should be selected.

The frequency components of the core clock genes with relative amplitudes greater than 0.7 (corresponding to half of the maximum value in the spectra) are chosen as dominant frequencies. The union of the sets of dominant frequencies are defined as Circadian Clock Frequency Range (CCFR), which gives the frequency range of the circadian patterns under a given condition and bandwidth of the filters that are used in decomposition step. The weight of each frequency component in the CCFR is derived according to the magnitude of the Fourier coefficient of the corresponding frequency component. The vector of the weights (and their corresponding frequencies) forms the "clock vector," which defines the frequency response of a tapering bandpass filter within the CCFR.[10]

3.2.3. *PRIISM Step 3: Decomposition and reconstruction*

The goal of this step is to decompose the whole spectrum into three distinct sections: treatment-frequency, clock-frequency and noise-frequency components and then reconstruct each of them. For decomposing treatment-frequency component, given a relatively narrow frequency band, a low-pass filter with a steep cut-off frequency is used to gain the optimal balance between removing ringing artifacts and approximating the desired frequency responses (see details in Sec. 3.2.4). A tapering bandpass filter is applied to reconstruct the clock-frequency expression pattern that has reduced noise. The reconstructed high-frequency expression pattern is considered to be noise.

For pattern reconstruction, the inverse Fourier transform is performed individually on the treatment-frequency and clock-frequency sections for each gene. Similar to using the clock vector as a tapering band-pass filter to remove noise, PRIISM adds a coarse-graining process to increase the robustness of component selection by making sure there is no overlap between any two frequency bands. The mean of the original gene expression values (which was removed in Step 1), is added back proportionally to each gene expression curve based on the amplitude distribution of each component in the spectrum of original expression. For more details of the decomposition and reconstruction step, please refer to Rosa *et al.*[10]

3.2.4. *Design of the low-pass filter*

The performance of PRIISM is largely determined by the design of the low-pass filter used in the decomposition step. When designing a low-pass filter, the balance between approximating an desired frequency response and reducing ringing artifacts should be considered. Traditionally, to minimize ringing artifacts, tapering low-pass filters (such as a Butterworth filter) are commonly used in signal processing. Note

that the smoothing effect of a tapering filter is dependent on the width of its transition band, i.e. the wider the bandwidth is, the more ringing artifacts are removed (but the farther the designed frequency response is from the desired one). Due to the specificity of the time-series gene expression data, however, an ideal low-pass filter with a strict cutoff frequency is adopted instead of a Butterworth filter for treating treatment-frequency components in PRIISM.

Below, we use *AtGolS3* (*AT1G09350*) as an example and processed its expression data with two different filters (Fig. 6) to explain the choice of the ideal low-pass filter.

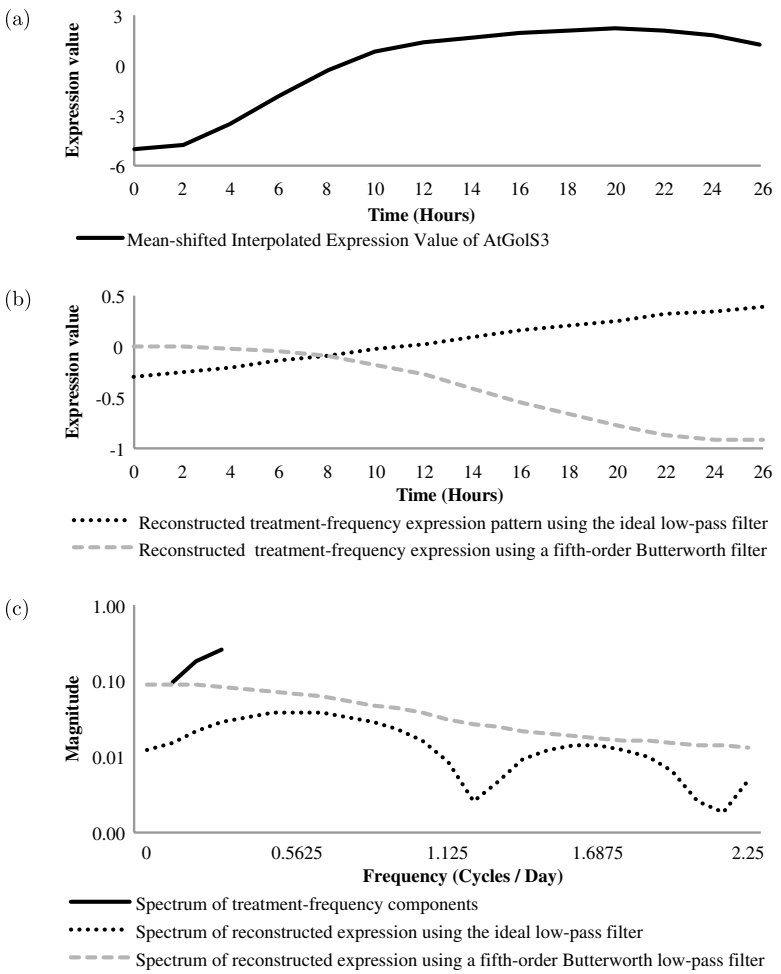

Fig. 6. The mean-shifted and interpolated gene expression values of *AtGolS3* is shown in (a), and the reconstructed gene expression values by the ideal low-pass filter (dotted black line) and a fifth-order Butterworth filter (dash grey line) are shown in (b). Figure (c) shows the comparison of the frequency spectra of *AtGolS3* by using the ideal low-pass filter (dotted black line) or the fifth-order Butterworth low-pass filter (dash grey line). The original treatment-frequency spectrum of *AtGolS3* is also shown in solid black line.

The mean-shifted interpolated gene expression values of *AtGolS3* are shown in Fig. 6(a), and the resulting expression patterns with an ideal low-pass filter or a fifth-order Butterworth filter are shown respectively in Fig. 6(b). To compare the performance of the two reconstructed series, we computed their spectra by performing FFT again. The spectrum of the original data (mean-shifted interpolated) and the spectra of the reconstructed series by the ideal low-pass filter and the Butterworth low-pass filter are all shown in Fig. 6(c). The spectrum of the ideal low-pass filter data has a similar changing trend as the spectrum of the original data in the low-frequency range. Its ringing artifacts (the phenomenon of output oscillating near a sharp transition in the input) appear, but the heights of the peaks at high-frequency range are relatively low because gene expression values usually do not change sharply. Consequently, there is no visible oscillation in the reconstructed pattern in Fig. 6(b). The transition band in the Butterworth filter is narrow because the treatment-frequency band of gene expression data is usually relatively narrow (less than 1/10 of the bandwidth of the original gene expression), which results in the opposite frequency response.[10] As a result, the spectrum obtained by the Butterworth filter is totally different from the spectrum of the original data. Consequently, the reconstructed gene expression pattern goes down over time, which is opposite to the original expression pattern.

In summary, more artifacts were added by using the Butterworth filter compared with the ideal low-pass filter. Therefore, considering the tradeoff between removing ringing artifact and approximating desired frequency response, an ideal low-pass filter with a steep cutoff frequency rather than a Butterworth filter was applied for treating treatment-frequency components in PRIISM. In the future, more advanced filtering methods should be designed to better capture clock disruptions.

4. Future Directions

To the best of our knowledge, PRIISM is the first attempt to identify and distinguish differential expression resulting from cycle disruption from adaptive-response differential expression. By applying the Fourier transform with a ideal low-pass filter and a tapering bandpass filter, time-series gene expression data are decomposed into three constituent components, each of which corresponds to treatment, circadian clock and noise. PRIISM provides an insight into the structure of the treatment-induced cyclic gene expression pattern disruption and achieves much better performance than the other existing methods in detecting cold-responsive genes in the *Arabidopsis* study.[10] However, this solution is based on several assumptions which simplify the biological problem.

First, PRIISM assumes the gene expression values are time invariant in a given period timeframe, so that the Fourier analysis can be applied. Advanced periodograms, such as wavelet or other time-frequency analysis techniques, and the methods discussed in Sec. 3.1, should be used to more accurately detect the time variant periodic patterns.

Second, the current implementation of PRIISM requires the user to pre-define several well-known genes that exhibit circadian regulation. However, core circadian genes have not been identified in some species and may be poorly studied or unverified in others. In these cases, the periodic detection methods outlined in Sec. 3.1 could be applied to automatically define dominant clock frequency components without using core circadian genes.

Third, the reconstructed gene expression patterns [see an example in Fig. 6(b)] have a similar changing trend as the original one but its spectrum magnitude is much lower than that of the spectrum of original data because of Parseval's theorem.[47] Therefore, a better low-pass filter to reconstruct the gene expression pattern at the low-frequency band is desired.

Fourth, measuring cyclic expression in time-series high-throughput gene expression datasets is complicated by high rates of noise, particularly for microarray datasets.[48] In addition, most of these datasets are very sparsely sampled (with 75% of the datasets containing five or fewer timepoints[49]), which may result in missed peaks or valleys in expression, or incorrect peak interpolation due to temporal aggregation effects.[50] RNA-seq technology, which sequences transcripts and results in much more accurate and less noisy quantification of gene expression, is expected to replace microarrays for high-throughput gene expression measurement.[51] As the cost of running RNA-seq experiments continues to fall,[51] higher-resolution datasets with significantly more information and less noise than existing datasets may be used to further improve the ability of researchers to identify and remove cycle disruption influences.

5. Conclusion

The treatment-induced disruption of regular cyclic gene expression patterns is a significant challenge in novel gene discovery experiments because these disruptions can induce strong differential regulation events for genes that are regulated by circadian clocks but are not involved in a response to the treatment. Although there are many advanced existing approaches for detecting clock and cell-cycle patterns in gene expression data, to the best of our knowledge, PRIISM is the first and only approach which tries to distinguish the clock patterns from the data. It can be integrated with any existing analysis approach on gene expression data to decompose circadian-influenced changes in gene expression.

PRIISM has been applied to study circadian rhythm patterns but it may also be used to separate fold-change patterns resulting from the disruption of cell cycle patterns in synchronized cell culture experiments. Circadian rhythms exist due to presence of environmental cues (i.e. light and darkness), but mitotic divisions require chromosomal enlargement, organelles and chromosomes to divide and the cell walls to expand and pinch off into two cells entirely.[52] Regardless of the little overlap between these functions at the genetic-activity level, PRIISM, which is not limited by the predefined models, may be extended to study cell-cycle oscillations. In

addition, PRIISM may be applied on high-throughput protein-expression datasets, or on metabolic flux datasets, which are similarly influenced by treatment-induced circadian clock and cell-cycle pattern disruptions.

In summary, PRIISM can be applied on any high-throughput time-series expression dataset which has (A) a relatively dense sampling rate which covers at least one cycle of the rhythmic patterns expected; (B) regular gene expression oscillations (including circadian rhythm patterns or synchronized cell-cycle patterns) across a large portion of the genome; (C) a list of core cycle genes identified either by performing literature searches, identifying genes with very distinct clock patterns in control conditions (Sec. 3.1), or finding clock gene orthologs based on bioinformatics comparisons to other organisms[53]; (D) control and treatment samples, where the treatment may be a biotic or abiotic stress (which has a frequency of treatment which is distinct from the cycle frequency), a different living or growing environment, or a different strain of the same organism which has altered clock gene patterns.

Clock patterns may vary considerably between organisms in different environmental conditions so it is not advisable to apply PRIISM on multiple pooled datasets or on data from experiments which have not had carefully controlled environmental conditions.

As higher-resolution time-series high-throughput gene expression datasets become available, there are many aspects of PRIISM which may be improved in order to better distinguish clock and treatment−response influences, to produce better datasets for performing novel gene discovery.

Acknowledgments

We thank Dr. Eva Farré for her feedback and helpful advice. This project has been funded by the U.S. Department of Energy (Chemical Sciences, Geosciences and Biosciences Division, grant no. DEFG0291ER20021 to J.C. and B.L.M.), the National Science Foundation (grant no. MCB-0919100 to B.L.M.), the Natural Sciences and Engineering Research Council of Canada (NSERC) through a Post-Graduate Scholarship to B.R. and NSERC Collaborative Research and Development grant to W.Q., and Ontario Research Chair funding to W.Q.

References

1. Cooper S, Shedden K, Microarray analysis of gene expression during the cell cycle, *Cell Chromosome* **2**:1, 2003.
2. Dodd AN, Salathia N, Hall A, Kévei E, Tóth R, Nagy F, Hibberd JM, Millar AJ, Webb AA, Plant circadian clocks increase photosynthesis, growth, survival, and competitive advantage, *Science* **309**(5734):630−633, 2005.
3. Dyczkowski J, Vingron M, Comparative analysis of cell cycle regulated genes in eukaryotes, *Genome Inform* **16**(1):125−131, 2005.
4. Doherty CJ, Kay SA, Circadian control of global gene expression patterns, *Annu Rev Genet* **44**(1):419−444, 2010.

5. Uchida Y, Osaki T, Yamasaki T, Shimomura T, Hata S, Horikawa K, Shibata S, Todo T, Hirayama J, Nishina H, Involvement of stress kinase mitogen-activated protein kinase kinase 7 in regulation of mammalian circadian clock, *J Biol Chem* **287**(11):8318−8326, 2012.
6. Michael TP, Mockler TC, Breton G, McEntee C, Byer A, Trout JD, Hazen SP, Shen R, Priest HD, Sullivan CM, Givan SA, Yanovsky M, Hong F, Kay SA, Chory J, Network discovery pipeline elucidates conserved time-of-day-specific cis-regulatory modules, *PLoS Genet* **4**(2):e14, 2008.
7. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO, Genomic expression programs in the response of yeast cells to environmental changes, *Mol Biol Cell* **11**(12):4241−4257, 2000.
8. Breeden LL, Periodic transcription: A cycle within a cycle, *Curr Biol* **13**(1):R31−R38, 2003.
9. Espinoza C, Degenkolbe T, Caldana C, Zuther E, Leisse A, Willmitzer L, Hincha DK, Hannah MA, Interaction with diurnal and circadian regulation results in dynamic metabolic and transcriptional changes during cold acclimation in *Arabidopsis*, *PLoS One* **5**(11):e14101, 2010.
10. Rosa BA, Jiao Y, Oh S, Montgomery BL, Qin W, Chen J, Frequency-based time-series gene expression recomposition using PRIISM, *BMC Syst Biol* **6**(69), 2012.
11. Hughes ME, DiTacchio L, Hayes KR, Vollmers C, Pulivarthy S, Baggs JE, Panda S, Hogenesch JB, Harmonics of circadian gene transcription in mammals, *PLoS Genet* **5**(4):e1000442, 2009.
12. Takahashi JS, Hong H, Ko CH, McDearmon EL, The genetics of mammalian circadian order and disorder: Implications for physiology and disease, *Nat Rev Genet* **9**(10):764−775, 2008.
13. Izumo M, Sato TR, Straume M, Johnson CH, Quantitative analyses of circadian gene expression in mammalian cell cultures, *PLoS Comput Biol* **2**(10):e136, 2006.
14. Adams S, Carre IA, Downstream of the plant circadian clock: Output pathways for the control of physiology and development, *Essays Biochem* **49**(1):53−69, 2011.
15. Edwards KD, Anderson PE, Hall A, Salathia NS, Locke JCW, Lynn JR, Straume M, Smith JQ, Millar AJ, *FLOWERING LOCUS C* mediates natural variation in the high-temperature response of the *Arabidopsis* circadian clock, *Plant Cell* **18**(3):639−650, 2006.
16. Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, Wang X, Kreps JA, Kay SA, Orchestrated transcription of key pathways in *Arabidopsis* by the circadian clock, *Science* **290**(5499):2110−2113, 2000.
17. Ptitsyn A, Comprehensive analysis of circadian periodic pattern in plant transcriptome, *BMC Bioinformatics* **9**(Suppl 9):S18, 2008.
18. Bilgin DD, Zavala JA, Zhu J, Clough SJ, Ort DR, Delucia EH, Biotic stress globally downregulates photosynthesis genes, *Plant Cell Environ* **33**(10):1597−1613, 2010.
19. Chaves MM, Flexas J, Pinheiro C, Photosynthesis under drought and salt stress: Regulation mechanisms from whole plant to cell, *Ann Bot* **103**(4):551−560, 2009.
20. Bieniawska Z, Espinoza C, Schlereth A, Sulpice R, Hincha DK, Hannah MA, Disruption of the *Arabidopsis* circadian clock is responsible for extensive variation in the cold-responsive transcriptome, *Plant Physiol* **147**(1):263−279, 2008.
21. Nakamichi N, Kusano M, Fukushima A, Kita M, Ito S, Yamashino T, Saito K, Sakakibara H, Mizuno T, Transcript profiling of an *Arabidopsis* pseudo response regulator arrhythmic triple mutant reveals a role for the circadian clock in cold stress response, *Plant Cell Physiol* **50**(3):447−462, 2009.
22. Zhao W, Serpedin E, Dougherty ER, Identifying genes involved in cyclic processes by combining gene expression analysis and prior knowledge, *EURASIP J Bioinform Syst Biol* **2009**(7):683463, 2009.

23. Brauer MJ, Huttenhower C, Airoldi EM, Rosenstein R, Matese JC, Gresham D, Boer VM, Troyanskaya OG, Botstein D, Coordination of growth rate, cell cycle, stress response, and metabolic activity in yeast, *Mol Biol Cell* **19**(1):352−367, 2008.
24. Castrillo J, Zeef L, Hoyle D, Zhang N, Hayes A, Gardner D, Cornell M, Petty J, Hakes L, Wardleworth L, Rash B, Brown M, Dunn W, Broadhurst D, O'Donoghue K, Hester S, Dunkley T, Hart S, Swainston N, Li P, Gaskell S, Paton N, Lilley K, Kell D, Oliver S, Growth control of the eukaryote cell: A systems biology study in yeast, *J Biol* **6**(2):4, 2007.
25. Salome PA, Xie Q, McClung CR, Circadian timekeeping during early *Arabidopsis* development, *Plant Physiol* **147**(3):1110−1125, 2008.
26. Dong MA, Farre EM, Thomashow MF, Circadian clock-associated 1 and late elongated hypocotyl regulate expression of the C-Repeat Binding Factor (CBF) pathway in *Arabidopsis*, *Proc Natl Acad Sci USA* **108**(17):7241−7246, 2011.
27. Kobelková A, Bajgar A, Dolezel D, Functional molecular analysis of a circadian clock gene timeless promoter from the drosophilid fly *Chymomyza costata*, *J Biol Rhythm* **25**(6):399−409, 2010.
28. Yang M-Y, Yang W-C, Lin P-M, Hsu J-F, Hsiao H-H, Liu Y-C, Tsai H-J, Chang C-S, Lin S-F, Altered expression of circadian clock genes in human chronic myeloid leukemia, *J Biol Rhythm* **26**(2):136−148, 2011.
29. Chudova D, Ihler A, Lin KK, Andersen B, Smyth P, Bayesian detection of non-sinusoidal periodic patterns in circadian expression data, *Bioinformatics* **25**(23):3114−3120, 2009.
30. 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* **109**(3):307−320, 2002.
31. Luan Y, Li H, Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data, *Bioinformatics* **20**(3):332−339, 2004.
32. Mockler TC, Michael TP, Priest HD, Shen R, Sullivan CM, Givan SA, McEntee C, Kay SA, Chory J, The Diurnal project: Diurnal and circadian expression profiling, model-based pattern matching, and promoter analysis, *Cold Spring Harb Symp Quant Biol* **72**:353−363, 2007.
33. Straume M, DNA microarray time series analysis: Automated statistical assessment of circadian rhythms in gene expression patterning, *Method Enzymol* **383**:149−166, 2004.
34. Ceriani MF, Hogenesch JB, Yanovsky M, Panda S, Straume M, Kay SA, Genome-wide expression analysis in *Drosophila* reveals genes controlling circadian behavior, *J Neurosci* **22**(21):9305−9319, 2002.
35. Ueda HR, Matsumoto A, Kawamura M, Iino M, Tanimura T, Hashimoto S, Genome-wide transcriptional orchestration of circadian rhythms in *Drosophila*, *J Biol Chem* **277**(16):14048−14052, 2002.
36. Lin KK, Chudova D, Hatfield GW, Smyth P, Andersen B, Identification of hair cycle-associated genes from time-course gene expression profile data by using replicate variance, *Proc Natl Acad Sci USA* **101**(45):15955−15960, 2004.
37. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B, Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol Biol Cell* **9**(12):3273−3297, 1998.
38. Whitfield ML, Sherlock G, Saldanha AJ, Murray JI, Ball CA, Alexander KE, Matese JC, Perou CM, Hurt MM, Brown PO, Botstein D, Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Mol Biol Cell* **13**(6):1977−2000, 2002.
39. Bozdech Z, Llinás M, Pulliam BL, Wong ED, Zhu J, DeRisi JL, The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*, *PLoS Biol* **1**(1):85−100, 2003.

40. Rustici G, Mata J, Kivinen J, Lió P, Penkett CJ, Burns G, Hayles J, Brazma A, Nurse P, Bähler J, Periodic gene expression program of the fission yeast cell cycle, *Nat Genet* **36**:809−817, 2004.
41. Langmead CJ, Yan AK, McClung CR, Donald BR, Phase-independent rhythmic analysis of genome-wide expression patterns, in *Proc Sixth Annu Int Conf Comput Biol*, RECOMB '02, pp. 205−215, New York, NY, USA, 2002.
42. Tai YC, Speed T, On the gene ranking of replicated microarray time course data, Technical report, University of California, Berkeley, 2004.
43. Glynn EF, Chen J, Mushegian AR, Detecting periodic patterns in unevenly spaced gene expression time series using Lomb−Scargle periodograms, *Bioinformatics* **22**(3):310−316, 2006.
44. Liang K, Wang X, Li T, Robust discovery of periodically expressed genes using the Laplace periodogram, *BMC Bioinformatics* **10**(1):15, 2009.
45. Wichert S, Fokianos K, Strimmer K, Identifying periodically expressed transcripts in microarray time series data, *Bioinformatics* **20**(1):5−20, 2004.
46. Mas P, Circadian clock function in *Arabidopsis thaliana*: Time beyond transcription, *Trends Cell Biol* **18**(6):273−281, 2008.
47. Orfanidis S, *Introduction to Signal Processing* (Prentice Hall, New Jersey, USA, 1995).
48. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y, RNA-seq: An assessment of technical reproducibility and comparison with gene expression arrays, *Genome Res* **18**(9):1509−1517, 2008.
49. Edgar R, Domrachev M, Lash AE, Gene expression omnibus: NCBI gene expression and hybridization array data repository, *Nucleic Acids Res* **30**(1):207−210, 2002.
50. Bay SD, Chrisman L, Pohorille A, Shrager J, Temporal aggregation bias and inference of causal regulatory networks, *J Comput Biol* **11**(5):971−985, 2004.
51. Wang Z, Gerstein M, Snyder M, RNA-seq: A revolutionary tool for transcriptomics, *Nat Rev Genet* **10**(1):57−63, 01, 2009.
52. Rieder CL, Khodjakov A, Mitosis through the microscope: Advances in seeing inside live dividing cells, *Science* **300**:91−96, 2003.
53. Bae K, Lee C, Sidote D, Chuang K, Edery I, Circadian regulation of a *Drosophila* homolog of the mammalian *Clock* gene: PER and TIM function as positive regulators, *Mol Cell Biol* **18**(10):6142−6151, 1998.

Yuhua Jiao received her Ph.D. in Computer Science from Harbin Institute of Technology (China) in 2009. She has an extensive research background in data analysis with machine learning, statistical modeling and signal processing. She is a postdoctoral research associate of Dr. Jin Chen at MSU-DOE Plant Research Lab, Michigan State University. Her current studies focus on microarray data analysis for better understanding the energy conversion systems.

Bruce Rosa received his Masters degree in Biology in 2007, and his Ph.D. in Biotechnology from Dr. Wensheng Qin Lab at Lakehead University (Canada) in 2012. He was a visiting scholar from 2010 to 2012 at Dr. Jin Chen's Lab at Michigan State University. He has an extensive research background in microbiology but has focused

his Ph.D. research on bioinformatics, primarily in the study of high-throughput gene expression datasets and time-series data. He is currently an NIH-funded postdoctoral research associate in the Genomics Institute at the Washington University at St. Louis Medical School, performing bioinformatics experiments on parasitic nematodes using both sequence-based and gene expression datasets.

Sookyung Oh received her Ph.D. degree from Dr. Steve van Nocker's lab at Michigan State University with a major in Plant Breeding and Genetics in 2006. She is interested in microarray/RNA sequencing data analysis and mining approaches. Currently she is a research associate at Dr. Beronda Montgomery's Lab at Michigan State University. She is currently engaged in phytochrome-mediated light signaling project with respect to plant development and photosynthesis in *Arabidopsis*.

Beronda Montgomery's research interest centers on elucidating the mechanisms utilized by photosynthetic organisms for adapting to changes in their photo-environment. Her group's current studies focus on the synthesis of photosensory biliproteins and investigations into their physiological roles during selected aspects of photomorphogenesis in the model plant *Arabidopsis* and *cyanobacteria*. Dr. Montgomery received her Ph.D. from the University of California, Davis, in 2001. She then completed a National Science Foundation-funded postdoctoral fellowship in Microbial Biology at Indiana University. In August 2004, she was appointed as a faculty member at Michigan State University in the Department of Energy — Plant Research Laboratory and the Department of Biochemistry and Molecular Biology, where she is now an Associate Professor.

Wensheng Qin obtained his Bachelor's and Master's degrees in China. He then received his Ph.D. degree in Molecular Biology and Biotechnology at Queen's University in Canada in 2005. Afterward, he conducted postdoctoral research at Stanford University from 2005 to 2008. He was an Assistant Professor and then Associate Professor in Biotechnology at Lakehead University and Ontario Research Chair in Biorefining Research since 2008.

Jin Chen is an Assistant Professor in MSU-DOE Plant Research Laboratory, Computer Science and Engineering Department at Michigan State University. He is interested in constructing the plant bioenergy network with machine-learning methods to better understand the energy conversion systems. He obtained his B.E. in Computer Science from Southeast University in 1997 and Ph.D. degree in Computer Science from National University of Singapore in 2007.