Published in final edited form as:

Methods Mol Biol. 2014 ; 1158: 45–56. doi:10.1007/978-1-4939-0700-7_3.

Global profiling of the circadian transcriptome using microarrays

Polly Yingshan Hsu¹ and Stacey L. Harmer^{1,*}

¹Department of Plant Biology, University of California, Davis, California, United States of America

i. Summary

Circadian rhythms are biological cycles with a period length of approximately 24 hours that are generated by endogenous clocks. The application of microarrays for high-throughput transcriptome analysis has led to the insight that substantial portions of the transcriptomes of both humans and many model organisms are clock-regulated. In a typical circadian time course microarray experiment, samples are collected from organisms maintained in constant environmental conditions, gene expression at each time point is determined using microarrays, and finally clock-regulated transcripts are identified using statistical algorithms. Here, we describe how to design the experiment, process RNA, determine expression profiles using ATH1 microarrays, and use a non-parametric statistical algorithm named JTK_CYCLE in order to identify circadian-regulated transcripts in Arabidopsis. This basic procedure can be modified to identify clock-regulated transcripts in different organisms or using different expression analysis platforms.

Keywords

Transcriptome; microarray; circadian rhythms; cycling genes; JTK_CYCLE

1. Introduction

Circadian rhythms are physiological and behavior cycles with a period of approximately 24 hours. Numerous circadian microarray experiments have shown that a substantial portion of the transcriptomes of cyanobacteria, Arabidopsis, Drosophila and mammals are controlled by the clock [1-5]. The examination of global transcript abundance over circadian time allows the identification of both components of the circadian oscillator as well as genes and pathways under circadian regulation.

Microarray analysis is a well-established technique that enables the simultaneous, genome-wide monitoring of essentially all transcripts in a sample. It involves the hybridization of biotin-labeled copy RNA (cRNA) (or DNA, depending upon the labeling procedure) molecules to thousands, or even millions of different probe sequences that are immobilized on a solid support to form a microarray. For Affymetrix gene expression arrays, these probes are 25 nucleotide– long oligonucleotides complementary to RNA transcripts. A group of

* To whom correspondence should be addressed. slharmer@ucdavis.edu.

probes complementary to the same transcript is called a probeset; to help control for local differences in hybridization efficiency, probes belonging to a single probeset are usually distributed across the entire array.

After biotin-labelled nucleic acids are hybridized to an array, the arrays are washed, stained with streptavidin-phycoerythrin, and then scanned to determine the patterns of hybridization. This raw image data is then preprocessed to reduce technical sources of variation, such as uneven hybridization efficiency among different microarrays. Here we describe the robust multi-array average (RMA) [6] method for this low-level processing. RMA allows for background correction, quantile normalization and summarization using a robust, multi-array linear model [6]. This low-level analysis requires a .cdf (chip description file) file which provides a “map” indicating the position of each probe on the array and generates expression values for individual probesets and thus transcripts. RMA analysis of each array in a time course therefore results in an average expression value for each probeset on the array for every time at which a sample was harvested.

After obtaining expression values for individual transcripts over such a time course, clock-regulated transcripts are commonly identified as those with expression patterns that fit a cosine wave with a near-circadian period (typically 20-28 hours). Many statistical algorithms have proven useful in these studies [7-9]. We focus here on the use of a non-parametric algorithm, JTK_CYCLE [10], to identify circadian-regulated transcripts. JTK_CYCLE is a computationally efficient algorithm that has been implemented in the statistical language and environment R, making it relatively fast and easy to use.

In this chapter, we will describe the steps involved in identifying circadian-regulated transcripts using microarrays. These steps are: 1) collection of samples over a circadian time course, 2) RNA extraction and quality control, 3) labeling of RNA and hybridization to microarrays, 4) low-level analysis of microarray data, and 5) identification of circadian-controlled genes using JTK_CYCLE. Since step three is usually carried out in core facilities, we will focus our attention on the other steps, which are commonly carried out in individual labs.

2. Materials

2.1 Circadian time course sample collection

1. Growth media: 1X Murashige and Skoog, 0.7% agar, 3% sucrose, pH 5.7.
2. Petri dishes: 100 × 100 × 15mm disposable petri dishes with grids (Fisher Scientific, cat. number 0875711A)
3. Growth chamber: Conviron TC 16 growth chamber
4. 1.5mL screw-cap tubes
5. 2mm zirconia beads (BioSpec Products, Cat. number 11079124zx)
6. Liquid nitrogen
7. Forceps for collecting plant samples

2.2 RNA extraction and quality control

1. Liquid nitrogen
2. Mini-beadbeater-96 (BioSpec Products, Cat. number 1001) and 2.0ml vial rack (BioSpec Products, cat. number 702VH45)
3. Ice bucket with ice
4. Trizol (Invitrogen)
5. Chloroform
6. 4°C microfuge
7. Isopropanol
8. 70% Ethanol
9. 95% Ethanol
10. DNase I (Qiagen): resuspend the DNase I power in nuclease-free water and aliquot according to the manufacturer's manual. Store at -20°C.
11. RNeasy MinElute Cleanup Kit (Qiagen)
12. NanoDrop ND 1000 (NanoDrop Technologies)
13. DNase-, RNase-free pipet tips and eppendorf tubes

2.3 Low-level analysis of microarray data

1. R: R is a free software environment for statistical computing and graphics. It compiles and runs on a wide variety of operating systems, including UNIX, Windows and MacOS platforms. You can download R here: <http://www.r-project.org/>
2. Practice microarray data set: GSE8365 data can be downloaded from here: <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE8365> (In the bottom of the page, click (http) under Download, unzip the file, place all of the .CEL files (raw data for each arrays) in your working folder)

2.4 Detection of circadian-controlled genes using JTK_CYCLE

1. JTK_CYCLE: a nonparametric test procedure that detects cycling components from a large data set, developed by Hughes et al. [10]. You can download JTK_CYCLE here: http://openwetware.org/wiki/HughesLab:JTK_Cycle (click "Image:JTK Cycle (2-19-13).zip" to download, unzip the file)
2. Microsoft Excel

3. Methods

3.1 Collection of samples to generate a circadian time course

1. Plant materials: Arabidopsis seeds are sterilized and plated on growth media (~30 seeds for each sample) (see Note 1) and stratified at 4 °C in the dark for two days before release to growth chamber.
2. Circadian time course: For a circadian time course, after stratification, the seeds are germinated and plants are entrained in 12 hr light /12 hr dark cycles (light: we use ~55μEi, cool white fluorescent bulbs) at 22°C for a week and released to free-run (constant light: 55μEi white fluorescent light and constant temperature: 22°C), and samples are collected starting after 24 hours in free-run (24 hours after the last light to dark transition) at a four-hour interval for two days (see Figure 1). It has been reported that shorter time intervals and/or a longer time course increase both the resolution of the time course and the statistical power for identifying circadian-regulated genes [1,11]. However, we have found our usual experimental design allows for the identification of thousands of clock-regulated transcripts.
3. Sample collection: Prepare and label screw-cap tubes each containing 3-4 beads. At each time point, quickly harvest plants using forceps, place plants in a screw-cap tube, freeze samples immediately in liquid nitrogen, and store the samples at -80°C.

3.2 RNA extraction and quality control

*** If not otherwise specified, RNA samples should be kept on ice at all times.

*** All centrifuge steps should use a microfuge cooled to 4°C.

*** Consult your core facility first to determine the quantity and quality of RNA required for your microarray analysis.

1. Pre-cool the microfuge to 4°C. Take samples from -80°C and place them in liquid nitrogen. Slowly pour liquid nitrogen on the vial rack of the beadbeater to cool it down. Place the samples in the cooled vial rack (put tubes in the center if possible so the samples can keep cold), and shake for 2 minutes. Quickly place the samples back in the liquid nitrogen container (see Note 2).
2. Prepare ice bucket with ice in the fume hood. In the fume hood, move 1-2 tubes from liquid nitrogen and place in the ice bucket. Open the screw caps and add 0.7ml- 1ml Trizol (see Note 3) per samples. Quickly place the cap back on and vortex the samples until the powdered tissue is completely resuspended (see Note 4). Leave these processed samples on ice while working on the rest of the samples.

¹For 1 week old Arabidopsis seedlings, follow the Trizol extraction described here, which typically yields about 20-30 μg total RNA per sample. Consult your core facility first about the quantity and quality of RNA that you'll need for your microarray experiment.

²Make sure the samples are still frozen. The samples should consist of a pale green powder when removed from the beadbeater.

³For 1 week old Arabidopsis seedlings, 700 μl Trizol is sufficient for 30 plants.

⁴When mixing the tissue with Trizol, some tissue may be stuck on the screw cap; therefore, shake the tubes up-and-down as well as vortex to make sure they are mixed well.

3. Add 1/5 volume of chloroform (ex. if you used 700µl Trizol, add 140µl chloroform) and invert the tube several times until the solution is thoroughly mixed. Centrifuge the samples in a microfuge at 13000 rpm for 15 minutes.

4. Use 200µl pipet tips (pipetman is set to 150µl), carefully transfer the aqueous layer to a new 1.5ml eppendorf tube (avoid disturbing the aqueous/organic interface; it is okay to leave a small volume behind); it typically takes 3 times to transfer the aqueous layer per sample. Add 1 volume of isopropanol, usually about 450µl. Invert the tube several times until the solution is thoroughly mixed. Centrifuge the samples in a microfuge at 13000 rpm for 15 minutes.

5. Remove the supernatant using 1ml pipet tips. Wash the pellet with 700µl 70% ethanol. Invert the tube several times until the pellet is resuspended. Centrifuge the samples in a microfuge at 13000 rpm for 5 minutes.

6. Remove the supernatant using 1ml pipet tips. Wash the pellet with 700µl 95% ethanol. Invert the tube several times until the pellet is resuspended. Centrifuge the samples in a microfuge at 13000 rpm for 5 minutes.

7. Remove the supernatant using 1ml pipet tips. Briefly spin again. Use 20µl tips to remove as much as liquid as possible. Air-dry the pellet by setting the tube upside-down on a piece of paper towel for 10-15 minutes at room temperature (see Note 5.)

8. Add 87.5µl nuclease-free water and allow the sample to sit on the bench at room temperature for 5 minutes. While waiting, combine DNase I solution and RDD buffer following the manufacturer's protocol.

9. Add 12.5 µl DNase I/RDD solution and pipet up-and-down to mix. Allow it to sit on the bench at room temperature for 10 minutes.

10. Clean RNA samples using Qiagen mini-elute clean up kit following the manufacturer's protocol (see Note 6).

11. Examine the quality and quantity of RNA using Nano-Drop. Select "Nucleic Acids" application module; select "RNA-40" sample type. Use 2µl of nuclease-free water to blank the machine and 2µl of each sample for measurement. The 260/280 and 260/230 ratios (ratios of sample absorbance at 260 and 280nm, or 260 and 230nm) are measures of nucleic acid purity. A ratio of ~2.0 for 260/280 is generally accepted as "pure" for RNA (a lower ratio may indicate contamination with protein or phenol). The ratio for 260/230 for "pure" nucleic acid is often higher than its respective 260/280 ratio (around 2.2) [12]. Sample concentration is presented in ng/µl. (see Note 7)

⁵Watch the pellets to make sure they are dry enough—they should look transparent rather than whitish. Any residual ethanol in the sample will cause a lower 260nm: 230nm ratio and may interfere with downstream enzymatic application, like cDNA synthesis.

⁶The Qiagen mini-elute clean up kit procedure is done at room temperature. It is best to split the samples into several batches (like 6 or 8 samples per batch) to reduce the time that the RNA sits at room temperature and to ensure high-quality RNA.

⁷We typically obtain 260/280 ratios of ~2.1 and 260/230 ratios of ~2.3 using the Trizol extraction method described here. Your core facility may perform additional quality control procedures, for example, using a Bioanalyzer (Agilent) machine to determine the size distribution and degradation level of the RNA samples.

12. Aliquot and dilute RNA samples as recommended by your core facility to avoid repeated freeze-thaw cycles. (see Note 8)

13. RNA should be stored in -80°C until ready for the next steps.

3.3 Labeling and microarray hybridization

These steps are commonly performed by a core facility affiliated with a university or research institute and are not discussed here.

3.4 Low-level analysis of microarray data

This and the next data analysis step assume some basic familiarity with the statistical language and software environment R, in that we expect you are able to import data to R and carry out basic data manipulations. Although to those not familiar with the command line this may seem daunting, a small investment of time will give you enough experience with R to carry out the analytical steps described below. Working through the early chapters of either Dalgaard's *Introductory Statistics with R* [13] or Adler's *R in a Nutshell* [14] should provide sufficient background for these analysis steps.

We will use a published dataset generated by Covington *et al.* [15] using Affymetrix ATH1 arrays (GSE8365) to demonstrate how to analyze microarray data. In this dataset, Arabidopsis seedlings were grown in light/dark cycles for 7 days and then moved to constant light and temperature. After 24 hours in this free-run condition, plants were harvested at four-hour intervals over two days. These samples were labels ZT24 – ZT68, corresponding to time of harvesting (ZT0 = the last dark to light transition). RNA was extracted, labeled, and then hybridized to ATH1 arrays in a core facility. Arrays were scanned to generate the raw image data (.CEL files) that were processed as described below.

*** Follow Method .2 to download the .CEL files from GSE8365.

*** The machine used in this demonstration is Mac OS $\times$ 1.4 with processor: 1.8GHz Intel Core i5, and memory: 4GB 1600 MHz DDR3. The software is R 2.15.2 and JTK_CYCLE updated on 2013/02/19.

*** Commands typed in R are presented in courier font.

1. In R, change the working directory to where the microarray raw data (.CEL files) are stored by clicking Misc -> Change Working Directory, and selecting the correct folder.
2. Install the "affy" package:

```
source("http://bioconductor.org/biocLite.R")
```

```
biocLite("affy")
```

Load the "affy" package:

```
library(affy)
```

⁸The concentration and volume of RNA needed for RNA labeling/hybridization vary depending on the procedures used and should be discussed with your core facility first.

3. Read all of the microarray raw data in the current working directory into memory; the CDF file corresponding to the ATH1 microarray platform will be automatically loaded (see Note 9)

```
rawData=ReadAffy()
```

Check information about the raw data and the microarray platform:

```
rawData
```

You should see the following information:

```
AffyBatch object
```

```
size of arrays=712x712 features (21 kb)
```

```
cdf=ATH1-121501 (22810 affyids)
```

```
number of samples=12
```

```
number of genes=22810
```

```
annotation=ath1121501
```

```
notes=
```

4. Evaluate hybridization of each individual array:

```
image(rawData[,1])
```

This produces an image for the first array (takes about 30 seconds). For arrays with good hybridization results, you should see a fairly even signal across the entire chip.

Change the value of $\times$ in `rawData[,x]` to different numbers (1-12 for these 12 arrays) to examine different arrays.

5. Rename the sample names in `rawData`:

```
sampleNames(rawData) = c(
  "ZT24", "ZT28", "ZT32", "ZT36", "ZT40", "ZT44", "ZT48", "ZT52", "ZT56",
  "ZT60", "ZT64", "ZT68")
```

6. To determine whether some arrays have overall higher or lower hybridization values than the others, examine the distribution of the signal intensities across all of the arrays in the experiment:

```
boxplot(rawData, main="Before RMA", ylab = "log2 intensity", las=2)
```

`main = "Before RMA"` generates a title for the plot, `las = 2` indicates that text in x-axis is vertical (use `las = 1` to specify horizontal text). It is common for the distribution of signal intensities to vary between individual arrays (Figure 2a), suggesting better hybridization on some arrays than others. To correct for this, it is

⁹If you are using custom Affymetrix microarrays, the corresponding .cdf file will not be loaded automatically. You will have to obtain the .cdf from the manufacturer and set the cdf environment using the bioconductor packages "altcdfenvs" [17] and "makecdfenv" [18] or follow the recommendations of your manufacturer.

necessary to normalize the data by adjusting the distribution of the signal intensity among all of the arrays in the experiment. This allows for the comparison of expression values for the same transcript over different arrays (see Note 10).

7. Normalize the arrays using RMA:

```
CovingtonRMA = rma(rawData)
```

This generates an object (CovingtonRMA) that has undergone RMA processing. RMA carries out background correction, normalization and calculation of expression values for each probeset (transcript).

8. Extract the expression values after RMA

```
e = exprs(CovingtonRMA)
```

9. Examine the expression values by checking the first six lines of the data

```
head(e)
```

10. Evaluate normalized expression after RMA

```
boxplot(e, main = "After RMA", ylab = "log2 intensity", las=2)
```

Compare this plot with that generated before RMA normalization (Figure 2a). After RMA processing, the distribution of signal intensities is very similar among the samples (Figure 2b).

11. Output the normalized expression values

```
write.csv(e, file = "Covington_RMA.csv")
```

You have now completed low level analysis and are ready to identify clock-regulated transcripts using JTK_CYCLE.

3.5 Detection of circadian-controlled genes using JTK_CYCLE

1. Prepare the data formats:

- Open the data file (Covington_RMsv) above using Excel and saved as a text delimited file, i.e. Covington_RMA.txt file.
- Copy the annotation column and paste to a new Excel sheet, giving it a header line such as “probeset”, and then save as a text delimited file, . Covington_RMA_annot.txt file.
- Place the data file (Covington_RMxt) and the annotation file (Covington_RMA_annot.txt) in the JTK_CYCLE folder. (see Note 11)

1. Open the file “run_JTK_CYCLE (Example1).R” using R and modify the following:

¹⁰There are three main assumptions underlying this normalization procedure: (1) the majority of genes are not differential expressed, (2) mis-regulated genes are equally likely to be up- or down-regulated, and (3) the differential expression signal is independent of the average gene expression levels [19].

¹¹If you collect samples in replicate, make sure the replicate samples are in adjacent columns when preparing the data file for JTK_CYCLE (see JTK_User's Guide “Example2_data.txt”).

- Line 4: change the project name to something that makes sense to you
- Line 7: change the file name to "Covington_RMA_annoxt"
- Line 8: change the file name to "Covington_RMxt"
- Line 12: indicate the number of time points and replicates (no change needed for Covington's data; see Note 12)
- Line 15: specify the spacing of the time points; for Covington's dataset, the sampling interval is 4 hours, change to "jtnit(periods, 4)"
- Line 14: specify the period length you are interested in. The numbers in Line 14 are based on the number of time points per cycle. In this case, the time points are spaced 4 hours apart. Change lines 14 to: "periods <- 5:7"; this will identify transcripts with periods between 20 and 28 hours.
- Save this modified script file with a different name, for example, "run_JTK_CYCLE_Covington.R"

2. In R, change working directory to the JTK_CYCLE folder with the data and annotation files prepared above by clicking Misc -> Change Working Directory, and selecting the correct folder.
3. After modifying the 'Run_JTK_CYCLE_Covington.R' script, select all and paste into R's console. The first half of script will run quickly, but the other half takes a while, dependent on your computer and the size of your data (takes about 1 minute).
4. JTK_CYCLE outputs two files: (1) a .txt with all the important cycling statistics, and (2) an .rda file. The .txt includes BH.Q (Benjamini-Hochberg q -value), ADJ.P (adjusted p -value), PER (period), LAG (phase), and AMP (amplitude), along with the raw data. The .rda file contains R objects and will not typically be needed by most users.
5. The BH.Q and ADJ.P values include a correction for multiple testing. ADJ.P < 0.05 is often used as a cut-off to identify clock-regulated transcripts [16], but you may want to compare several ADJ.P values (or BH.Q for more stringent criteria) and empirically determine if they are reasonable cut-offs for your data by examining gene expression patterns.

5. Acknowledgement

We thank Hsin-Yen Wu for reading and testing the commands in this manuscript, and Michael Hughes for helpful discussions about JTK_CYCLE. This work was supported by the National Institutes of Health (NIGMS) (<http://www.nigms.nih.gov/>) [GM069418] and the Taiwan Merit Scholarship (<http://web1.nsc.gov.tw/>) [NSC-095-SAF-I-564-014-TMS].

¹²If you collect time points with replicates, for example, if you have 13 time points, and 2 replicates per time point, change Line 12 to read: "jtkdist(13,2).

6. References

1. Covington MF, Maloof JN, Straume M, Kay SA, Harmer SL. Global transcriptome analysis reveals circadian regulation of key pathways in plant growth and development. *Genome Biol.* 2008; 9:R130. [PubMed: 18710561]
2. McDonald MJ, Rosbash M. Microarray analysis and organization of circadian gene expression in *Drosophila*. *Cell.* 2001; 107:567–578. [PubMed: 11733057]
3. Kucho K, Okamoto K, Tsuchiya Y, Nomura S, Nango M, et al. Global analysis of circadian expression in the cyanobacterium *Synechocystis* sp. strain PCC 6803. *J Bacteriol.* 2005; 187:2190–2199. [PubMed: 15743968]
4. Akhtar RA, Reddy AB, Maywood ES, Clayton JD, King VM, et al. Circadian cycling of the mouse liver transcriptome, as revealed by cDNA microarray, is driven by the suprachiasmatic nucleus. *Curr Biol.* 2002; 12:540–550. [PubMed: 11937022]
5. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, et al. Coordinated transcription of key pathways in the mouse by the circadian clock. *Cell.* 2002; 109:307–320. [PubMed: 12015981]
6. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, et al. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics.* 2003; 4:249–264. [PubMed: 12925520]
7. Straume M. DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning. *Methods Enzymol.* 2004; 383:149–166. [PubMed: 15063650]
8. Levine JD, Funes P, Dowse HB, Hall JC. Signal analysis of behavioral and molecular cycles. *BMC Neurosci.* 2002; 3:1. [PubMed: 11825337]
9. Wichert S, Fokianos K, Strimmer K. Identifying periodically expressed transcripts in microarray time series data. *Bioinformatics.* 2004; 20:5–20. [PubMed: 14693803]
10. Hughes ME, Hogenesch JB, Kornacker K. JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets. *J Biol Rhythms.* 2010; 25:372–380. [PubMed: 20876817]
11. Hughes ME, DiTacchio L, Hayes KR, Vollmers C, Pulivarthy S, et al. Harmonics of circadian gene transcription in mammals. *PLoS Genet.* 2009; 5:e1000442. [PubMed: 19343201]
12. NanoDrop, Technologies, Inc. ND-1000 Spectrophotometer User's Manual. 2007
13. Dalgaard, P. *Introductory Statistics With R.* Springer-Verlag; New York: 2009.
14. Adler, J. *R in a Nutshell.* O'Reilly Media; 2010.
15. Covington MF, Harmer SL. The circadian clock regulates auxin signaling and responses in *Arabidopsis*. *PLoS Biol.* 2007; 5:e222. [PubMed: 17683202]
16. Hsu PY, Harmer SL. Circadian phase has profound effects on differential expression analysis. *PLoS One.* 2012; 7:e49853. [PubMed: 23185460]
17. Gautier L, Moller M, Friis-Hansen L, Knudsen S. Alternative mapping of probes to genes for Affymetrix chips. *BMC Bioinformatics.* 2004; 5:111. [PubMed: 15310390]
18. Irizarry RA, Gautier L, Huber W, Bolstad B. *makecdfenv: CDF Environment Maker.* R package version 1360. 2006
19. Gohlmann, H.; Talloen, W. *Gene Expression Studies Using Affymetrix Microarrays.* CRC Press; 2010.

Figure 1.

Design of a circadian time course experiment. Arabidopsis seedlings are typically grown under light/dark cycle for a week and then released to free-run (constant light and temperature). Samples are collected after 24 hours in free-run at 4-hour intervals over two consecutive days.

Figure 2.

(a) Distributions of signal intensities in $\log_2$ scale among individual arrays before RMA normalization, (b) distributions of signal intensities in $\log_2$ scale among individual arrays after RMA normalization.