Guided Conjugate Bayesian Clustering for Uncovering Rhythmically Expressed Genes

Paul E. Anderson[1,2], Jim Q. Smith*[1,2], Kieron D. Edwards[3] and Andrew J. Millar[1,3]

[1]Interdisciplinary Program for Cellular Regulation, University of Warwick, Coventry, CV4 7AL, UK
[2]Department of Statistics, University of Warwick, Coventry, CV4 7AL, UK
[3]Institute of Molecular Plant Sciences, University of Edinburgh, EH9 3JH, UK

Email: Paul E. Anderson - p.e.anderson@warwick.ac.uk; Jim Q Smith*- j.q.smith@warwick.ac.uk; Kieron D. Edwards - k.edwards@ed.ac.uk; Andrew J. Millar - andrew.millar@ed.ac.uk;

*Corresponding author

Abstract

Background: An increasing number of microarray experiments produce time series of expression levels for many genes. Some recent clustering algorithms respect the time ordering of the data and are, importantly, extremely fast. The focus of this paper is the development of such an algorithm on a microarray data set consisting of 22,810 genes of the plant *Arabidopsis thaliana* measured at 13 time points over two days. Circadian rhythms control the timing of various physiological and metabolic processes and are regulated by genes acting in feedback loops. The aim is to cluster and classify the expression profiles in order to identify genes potentially involved in, and regulated by, the circadian clock.

Results: A greedy search over time series of expression levels (where series are compared pairwise, the two most similar put in the same cluster and so forth) will get a fast result but will only explore a very limited number of the possible partitions of the profiles. We propose an improved, deterministic method based on a multi-step application of a conjugate Bayesian clustering algorithm. It allows the entire space to be searched more fully and intelligently. The values of the summary statistics are used to not only score clusters of genes, but also to guide the search of the vast partition space. By following this procedure, we are able to cluster genes that are known to be rhythmically expressed with genes of previously unknown function; thus suggesting potentially interesting targets for future experiments.

Conclusions: We show that by employing a guided search of a vast space of potential clusterings, using coefficients of the posterior distributions and genes known to biologically interesting, we can quickly uncover and classify clusters of potentially rhythmically expressed genes. The algorithm can be adapted to cluster many other data sets consisting of large numbers of time series.

Background

Cluster models are well established in the analysis of gene expression data, see for example [1] and references therein. However, many traditional methods are not suited to time course experiments. Any sensible statistical analysis of time series needs to utilise the time order of the data and not be invariant to permutations of the time points [2]. In particular, observations taken in quick succession can be expected to be more related than those taken further apart. The regression model we employ, based on [2], allows us to meet this requirement. Further, since we have a moderately long time course ($T = 13$ compared to $T = 6$ in [2]), with an appropriate choice of basis function, we can also use the estimated coefficients of the basis to guide the search through the partition space and thus obtain a more informative final clustering. The basis functions determine how the time series are clustered: for example, on linear differential changes (via a linear spline basis) or periodicity (via a Fourier basis). A partition is simply a particular grouping of the genes into a distinct set of clusters. Clearly there can be a huge number of different partitions if the number of genes and/or time points is large.

Clustering algorithms are extremely useful for studying profiles of gene expressions for a number of reasons. Firstly, they perform a data reduction by clustering similar profiles together reducing, in our case, over 20,000 separate profiles into one hundred clusters of 5,696 potentially rhythmically expressed genes. This makes comparisons of groups of profiles feasible.

Note that ranking the genes in some way does not confer this property. For example, the COSOPT algorithm [3–5] fits cosine waves from a range of periods and phases to each profile and calculates the best fit. The user can then rank the profiles in terms of their phase, amplitude, etc. Whilst this may be useful as a summary for small numbers of genes, it does not allocate genes to distinct groups.

Two recent papers, which give very impressive results, have also considered other ways of detecting periodic genes [6, 7]. The problem of detecting genes with an unknown period is considered in [6]. In our

context, we have a clearly defined period (24 hours) but could investigate other periods by amending the basis functions and the measure of periodicity (currently, the second harmonic ratio or SHR). However, their technique is very useful in cases where the periodicity is unclear at the outset. Spectral analysis is combined with multiple hypothesis testing in [7]. A particular strength of this method is that it works well even when the noise is non-Gaussian. In both [6] and [7], the time series are transformed into a single number, the test statistic, which is used to rank the genes. In this sense, [6] is like a Bayesian version of COSOPT. It's important to realise, however, that their aims are different to ours.

Primarily, we calculate a score for a *partition* of the genes and not the individual genes themselves. This score is what determines the clustering. Also, the focus of this paper is not just on those genes that are periodic, but those that have a similar profile. Clustering provides a framework in which we can group genes with similar expression. We can then compare which genes fall into which clusters, see how close the time series are to the posterior mean of their cluster and examine how this distribution changes between different experiments. Clustering works by comparing genes directly rather than calculating measures on the individual genes and then ranking them. This leads to a second advantage of clustering in general. Genes with similar expression profiles are often hypothesised to be functionally related. A clustering algorithm can thus act as a filter for more complex models.

An advantage specific to both the algorithm presented in [2, 8] and the adaptation described here, is that the clustering is extremely fast. 6,000 genes from the 13 time point series detailed below take only a couple of minutes on a Pentium 4 PC. This allows us to cluster and re-cluster the data many times permitting an efficient exploration of the vast number of potential partitions. It should also be noted that, in our method, all 22,810 genes are used in the clustering at some stage and we do not perform any pre-filtering (such as omitting genes of low differential expression) at the first step. This speed is conferred by a conjugate Bayesian analysis. In essence, each partition is scored by the logarithm of its associated marginal likelihood evaluated at the observed data. When the analysis is conjugate, the scores of partitions are given by a simple algebraic formula and can thus be calculated almost instantaneously [9].

In Bayesian statistics, conjugacy ensures that the prior, likelihood and posterior all stay in the same family of distributions. The posterior then follows from the prior by a straightforward calculation of the new parameters from the old ones, with no need to resort to expensive numerical techniques such as Markov Chain Monte Carlo (MCMC).

A second benefit peculiar to this algorithm but inherent in any Bayesian analysis and exploited in this paper, is that when the basis functions are customised to contextual information, the corresponding

summary statistics of the posterior distribution of the regression coefficients have a clear interpretation. For example, when using a Fourier basis, the regression coefficients are Fourier coefficients. These can be used to characterise each cluster, and also (as we will demonstrate) to guide the search for more useful partitions of the model space in an intelligent manner. Furthermore, since these coefficients are calculated anyway to score associated partitions, their use for this purpose incurs virtually no additional cost. Essentially, we use the algorithm in [2] many times on various partitions of the genes with a Fourier basis to draw out rhythmically expressed genes. We show that this multi-step, customised approach has many advantages. In particular, it searches the enormous space of partitions by concentrating on regions of interest. This is in contrast to the single greedy search algorithm advanced in [2] which does not use contextual information to guide the search.

The data set clustered in this paper is taken from a recent microarray experiment on the plant model organism *Arabidopsis thaliana*. It was designed to detect genes whose expression levels, and hence functionality, may be connected with circadian rhythms. Plant samples were harvested every four hours during a 48 hour interval of constant light. RNA prepared from these samples was analysed using Affymetrix ATH1 microarrays. A full analysis and exposition of this data, together with a discussion of its biological significance is given in [10]. The methods section provides more detail on the experimental setup. The circadian clock is thought to comprise a relatively small number of genes (of order 10). Their protein products are connected in interlocking feedback loops [11]. Circadian rhythms of approximately 24 hour period are an emergent property of this gene network. These cycles control the regulation of various processes in the plant. Circadian clocks can be found in most eukaryotes. Despite differing in composition and construction, they are all self-sustaining, entrainable and can maintain timing across a range of temperatures.

The genes understood to be in the clock to date were identified genetically: certain mutations affect the timing of all circadian rhythms in the plant. The circadian clock circuit drives the rhythmic expression of a much larger number of downstream, clock-regulated genes (of order 1,000). These are the genes we aim to identify in this work. Rhythmic output from the circadian clock to clock-regulated genes is presumably mediated by a small number of regulators that are expressed at discrete circadian phases and which, directly or indirectly, control the clock regulated genes. Few of the rhythmic output regulators have so far been identified.

In summary, we present a method for intelligently and efficiently searching a vast space of time series clusterings by using a set of criteria that define measures of periodicity. We also briefly discuss how it can

be adapted to clustering other time series.

Results and Discussion

The objectives, nature and consequences of clustering

In this application it is hypothesised that sets of genes involved in the same regulatory pathway will have similar expression profiles over time. The objectives of our clustering algorithm are therefore to:

1. Identify the clusters containing genes that are co-expressed.
2. Find those clusters which appear to be *potentially circadian* (PC). (We define this precisely later.)
3. (a) Earmark a set of genes G from these clusters for future individual biological analysis in terms of their possible regulatory function.

   (b) Identify as accurately as possible the set G^* of those genes that are associated in some way or another to circadian functionality, classifying these genes as closely as possible in terms of their various expression profiles.

A specification of G and G^* for our data set is given in appendix A. The search over partitions needs to be customised in the light of the prior information we have about how the algorithm will be used.
Firstly, from the third objective listed above, two partitions sharing the same partition over PC genes but differing on their partitions of the remaining genes, will have identical scores however the underlying scoring function is devised. We therefore require a high score on the predictive distributions of potentially circadian clusters and are indifferent to their behaviour elsewhere. A score, or utility, is a measure of the validity of the clustering. For example, the score employed here is the log marginal likelihood of the clustering. This allows us to determine which of the enormous number of possible partitions is the most appropriate. Obviously any search of the partition space should therefore focus on correctly classifying rhythmically expressed genes.
Secondly, suppose that a partition ϕ' is a refinement of the partition ϕ so that each of the clusters $c_i \in \phi$, $1 \leq i \leq n(\phi)$, is partitioned into the clusters $\{c'_{i,1}, \ldots, c'_{i,n_i}\} \in \phi'$ where $n_i \geq 1$. Under mild regularity conditions, the following then holds for the partitions ϕ and ϕ' for both G and G^*: two expected utilities associated with either ϕ or ϕ' will be close in absolute distance whenever all the expression profiles of $\{c'_{i,1}, \ldots, c'_{i,n_i}\} \in \phi'$ are close to those in c_i, for $1 \leq i \leq n(\phi)$. Consequently, two partitions with very different numbers of clusters, and perhaps with quite different log marginal likelihood scores, can lead to almost equivalent optimal decisions provided that their profiles correspond in the sense described above.

A misleading clustering can be produced in two circumstances. The first is when PC genes with profiles that would be interpreted quite differently by the scientist are clustered together erroneously. Fortunately, our particular data set seemed to not suffer from this problem. The second is when PC genes are placed in non-circadian clusters or, less importantly, when non-clock regulated genes appear in, and therefore distort the profile of, a circadian cluster: a real problem that must be overcome in our application as outlined in the description of our guided clustering method.

Bayesian techniques are famed for incorporating prior beliefs and information from practitioners in the field under investigation. This expert judgement, concerning which expression profiles might be linked to circadian regulation, comes from two main sources. The first is the result of a mathematical consideration of the postulated physical processes. As well as the obvious necessity for the existence of an approximately 24 hour periodicity (here expressed through a relatively strong second harmonic in the posterior mean of the profiles), there are at least three features that single out a cluster of particular interest.

1. A cluster containing genes involved in regulation may often express at certain times. Technically, clusters responsive to this primary regulatory cluster will exhibit a group delay, see for example [12, 13].

2. As well as exhibiting a group delay, responsive clusters will act in a similar way to a low frequency band pass filter of the primary regulatory clusters [13]. This means in particular that responsive clusters — loosely clusters more associated with effects rather than causes — will be more sinusoidal with a 24 hour period. Regulatory *genes*, on the other hand, might be expected to have a single spike over about 24 hours or, alternatively, asymmetric cycles with a sharp rise to a peak and a slower decline. Some of these clusters are identified in appendix A.

3. From the results in [10], we expect clusters rich in transcription factors to have higher amplitudes. Thus clusters with larger amplitude may contain regulatory genes.

As will be seen later, the measure of the strength of periodicity (the SHR) is able to detect clusters which exhibit properties 1 and 2 (types I to V in the classification in appendix A), and the clustering algorithm itself (rather than the SHR) distinguishes between clusters of different amplitudes.

A second source of expert judgement comes from a small number of genes (currently 39 in our case) that are known or thought to be associated with some aspect of circadian regulation from biological experiments. For simplicity, we use this prior information sparingly although, interestingly, we will see that

these genes often lie in clusters whose profiles have the three features outlined above.

Outline of the Bayesian Agglomerative Hierarchical Clustering (AHC) algorithm

We now describe the model pertinent to our study. There are now many examples of Bayesian clustering algorithms in the literature [14, 15]. Indeed, several model-based clustering algorithms have recently been proposed for gene expression time series [16–19]. However, the approach expounded in [2], as discussed there, has numerous advantages over these methods.

Our algorithm is an extension of the technique used in [2]. Essentially, this groups together those curves that appear to have been drawn from a joint distribution with parameters $\boldsymbol{\beta}$ and $\boldsymbol{\epsilon}$ under the linear model $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Notice that this is a linear model with nonlinear basis functions.

Here, $\boldsymbol{Y}$ is the vector of the logged gene expression profiles and $\boldsymbol{B}$ is the design or basis function matrix that encodes the type of basis used for the clustering (linear spline, Fourier, wavelets, etc). It holds p basis functions. For example, in [2], $\boldsymbol{B}$ is a linear spline basis and it clusters well on linear changes in expression level. The parameter $\boldsymbol{\beta}$ is the vector of coefficients of the basis used to characterise the expression profiles and is estimated by the algorithm. $\boldsymbol{\epsilon}$ is the standard Gaussian noise component, so that $\boldsymbol{\epsilon} \sim \mathrm{N}(0, \sigma^2)$. An important feature of this linear model structure is its ability to account for time dependence.

For the application advanced here, as explained later, a more appropriate choice is the Fourier basis. Time series in the same cluster are assumed to have originated from a common Gaussian mixture process (this will become evident from the marginal likelihood: note that the predictive distributions are assumed to be t-distributed, not Gaussian [9]) and will, within a given cluster c, have the same $\boldsymbol{\beta}_c$ and σ_c^2. Notice that $\boldsymbol{\beta}_c$, σ_c^2 and N_c (the number of genes in the cluster c) will be different for each cluster, and that the dimensions of $\boldsymbol{B}$ and $\boldsymbol{Y}$ will need to be adjusted accordingly to enable the calculations below to be carried out. (Specifically, $\boldsymbol{B}_c$ will be an $N_cT \times p$ matrix and $\boldsymbol{Y}_c$ an $N_cT \times 1$ vector, assuming we have T time points.) The parameters $\boldsymbol{\beta}_c$ and σ_c^2 are assumed a priori independent across clusters.

As with any conjugate Bayesian analysis, various priors must be specified so that the prior-to-posterior analysis remains in the same family of distributions. Details, along with a description of single-step AHC, are given in the methods section. Note that a key attraction of this model is the calculation of the most probable number of clusters: unlike many algorithms, the number of clusters is found a posteriori and is not fixed in advance. Also, although these properties are not required in this paper, each gene can be measured at different time points or even a different number of time points [2].

Problems with single-step AHC

This conjugate algorithm is quite stable in practice and very fast, taking only a few minutes to cluster 6,000 genes. However, certain aspects of the corresponding predictive distributions arising from a conjugate analysis do not quite match with expert judgements. This mismatch, particularly when used with a coarse search technique such as AHC, can lead to inappropriate "optimal" (that is, highest scoring) partitions. Further problems with AHC are discussed and illustrated in appendix B. Fortunately, the simplicity of the posterior distributions allows us to determine when this mismatch might occur. We can then customise the search algorithms to minimise this difficulty and produce simple diagnostics to pick up any remaining problems.

For example, in this application, the joint distribution of the vectors of the Fourier coefficients is a product of multivariate Student t-distributions over each of the clusters. It is well known that t-distributions have heavy tails which means they are prone to outliers. It has been argued [20] that this is a good property for a cluster model because it tends to prevent the appearance of many singleton clusters with idiosyncratic profiles. Such singletons are treated as outliers of large, noisy clusters and are quickly absorbed into them. These clusters can then be routinely checked for such outliers.

However, the AHC *begins* with many singleton clusters. (It's simply a greedy search where time series are compared two at a time, the two most alike put in the same cluster and so forth.) As a consequence, the heavy tails can cause two very different time series to be placed in the same cluster early on. Such so-called *junk clusters* subsequently draw in many other genes, including PC genes with moderate variation in expression over time. This failure is analogous to getting stuck at a local minimum in the energy landscape of a system in statistical physics, such as a spin glass, even though a lower global minimum exists. The particular way in which the AHC combines clusters in a greedy manner causes this discrepancy. Splitting and re-clustering the time series in different combinations so that the PC genes have a chance to be pulled out of the junk clusters overcomes the problem.

Generalising to multi-step can overcome the deficiencies of single-step AHC

In the process of scoring various partitions, the clustering algorithm gives a signature to each cluster via the joint distribution of the regression coefficients and the sample variance. In the light of the comments above it is clear that, at minimal cost, the regression coefficients and sample variance can be used iteratively to improve the search for a near-optimal partition. The summaries of the clusters contained in this partition can also be used to categorise the genes in ways helpful to the scientist.

In the problem considered in [2], a linear spline function basis was used and the search objective was very simple: to look for ramp changes in expression level. The usefulness of cluster signatures was therefore small in this semi-parametric setting. However, the context of the circadian application we are interested in here is quite different and the cluster summaries are invaluable.

Explicitly, as we shall see, the posterior means of the second harmonic give much information on whether a given cluster is PC or a junk cluster. By applying the AHC to different collections of the time series, we can thus produce more stable results dependent on the exact composition of the pool of genes. Further, we can ensure that the partition space is only searched over genes that are PC. That is, we direct the search to interesting areas of the space. This improved search is akin to the techniques developed in [21, 22]. Note that the multi-step search is guided through the structure of the scoring function rather than prior beliefs.

Description of a multi-step guided extension of AHC

We first define a PC cluster. Biological expert judgement suggested that a PC cluster was one whose second harmonic was high relative to the third, fourth, fifth and sixth harmonics. (Since there are 13 time points, we can use up to six cosine/sine pairs of harmonics.) The expectation, or posterior mean, of the cluster profiles, $y(t)$, (with Fourier coefficients a_i and b_i, $1 \leq i \leq 6$) over a 48 hour time course is given by

$$E(y(t)) = E(a_0) + \sum_{i=1}^{6} \left(E(a_i) \cos\left(\frac{2\pi i t}{48}\right) + E(b_i) \sin\left(\frac{2\pi i t}{48}\right) \right) \tag{1}$$

This is shown as a blue line in figures 2 and 3. We define the *second harmonic ratio* or SHR as

$$\text{SHR} = \frac{\left(E(a_2)^2 + E(b_2)^2\right)^{\frac{1}{2}}}{\sum_{i=2}^{6} \left(E(a_i)^2 + E(b_i)^2\right)^{\frac{1}{2}}} \tag{2}$$

We can then use this measure to delimit the *potentially circadian* (PC), *potentially junk* (PJ) and *other* (O) clusters with the following criteria:

1. A cluster is called PC if SHR$\geq$0.4.
2. A cluster is called PJ if
   - SHR $\in [0.35, 0.4)$ OR
   - SHR $\in [0.3, 0.35)$ AND the largest expression level in the cluster is $> \ln(1.5)$.
3. A cluster is called O if if does not satisfy conditions 1.) or 2.).

The thresholds were determined by examining the cluster plots with the experimentalists. As can be seen in appendix A, clusters with an SHR bigger than 0.4 could hold circadian regulated genes. This accords well with the allocation of genes that biologists know are rhythmically expressing in the final clustering (see for example clusters 84, 85 and 87 in figure 2). The definition of a PJ cluster is a little more involved. Since we apply a multi-step AHC, it must be broad enough to catch any profiles that may not have yet been compared to a similar profile from a different part of the space, yet it must not carry lots of non-circadian profiles into the next step of the clustering, since these will draw the search into uninteresting partitions. In practice, this means clusters with an SHR between 0.3 and 0.4 (anything with SHR$<$0.3 is unlikely to contain circadian regulated genes). We can in fact go further and, for those with SHR between 0.3 and 0.35, restrict ourselves only to those whose largest expression value is bigger than ln(1.5), since (as we are clustering the log expression level) we are only interested in profiles that change significantly in the time course. (See condition three in the first part of this section.) ln(1.5) thus corresponds to a 50% increase in the expression level at some point in the time series. Further, on occasion, there can be another reason for labelling a cluster as PJ: namely if it contains a gene/s already known to be biologically interesting or relevant. However, this condition is rarely needed (see step seven in the algorithm below).

The ultimate aim is to obtain the best possible PC set by exposing each profile to different parts of the space in a bid to draw out the PC clusters. One can imagine various ways to do this. Furthermore, we require that this be done quickly. Below we outline what we did for our particular data set as it illustrates several ways of moving around the space. Although the algorithm as presented is specific to our data set, the method of re-clustering and sieving is generic and can be applied to any collection of time series. (The first step may seem particularly ad hoc. The 22,810 genes were split into four sets for two reasons. Firstly, 5,702 time series can be clustered extremely quickly which is advantageous since the AHC is applied many times. Secondly, 5,702 genes is a big enough sample to contain plenty of each type of cluster and the AHC can discriminate them reasonably well: this is confirmed by the small number of genes in PJ and O clusters at steps three and six of the algorithm shown in figure 1. If there were more genes, we would suggest breaking them up into groups of about 6,000 genes at the first step.)

As can be seen in figure 1, the AHC usually identifies the PC genes in one run, with a further clustering of the PJ and O sets only drawing out a couple of genes into the PC set. (We tried other ways of combining clusters, such as re-clustering the O sets by themselves but, once delimited by the algorithm, we found that re-clustering the O sets would not split them further into PC and PJ clusters: more O sets were produced. This suggested that it was better to combine the O sets with the PJ sets before re-clustering in a bid to

extract any PC sets.)

1. Place the 22,810 genes randomly into four distinct sets (two with 5,702 genes, two with 5,703 genes). Notice ALL genes are used at this step and there is no pre-filtering.

2. Perform AHC for each set. This produces PC[i], PJ[i] and O[i] for $1 \leq i \leq 4$.

3. For each $1 \leq i \leq 4$, perform AHC on PJ[i]$\cup$O[i]. Call the resulting sets PC$'$[i], PJ$'$[i] and O$'$[i]. (This ensures that we extract any PC clusters we may have missed in step 2.)

4. Form the sets
   TOP=PC[1]$\cup$PC$'$[1]$\cup$PJ$'$[1]$\cup$PC[2]$\cup$PC$'$[2]$\cup$PJ$'$[2] and
   BTM=PC[3]$\cup$PC$'$[3]$\cup$PJ$'$[3]$\cup$PC[4]$\cup$PC$'$[4]$\cup$PJ$'$[4].
   (That is, we discard O$'$[i] from any further analysis. We can think of this as clustering the top and bottom halves of the data set separately, but with the O clusters filtered out and removed. The next two steps are then analogous to steps two, three and four above.)

5. Perform AHC on TOP and BTM. This produces PC[j], PJ[j] and O[j] for $j \in \{TOP, BTM\}$.

6. For each $j \in \{TOP, BTM\}$, perform AHC on PJ[j]$\cup$O[j]. Call the resulting sets PC$'$[j], PJ$'$[j] and O$'$[j].

7. For each $j \in \{TOP, BTM\}$, put any clusters in PJ$'$[j] that contain genes that are known to be rhythmically expressed, into the set PJ*$'$[j]. (This ensures that we do not ignore biological prior information.)

8. Form the set
   $FINAL$=PC[TOP]$\cup$PC$'$[TOP]$\cup$PJ*$'$[TOP]
   $\cup$PC[BTM]$\cup$PC$'$[BTM]$\cup$PJ*$'$[BTM].
   (That is, we discard O$'$[j] and PJ$'$[j]$\backslash$PJ*$'$[j] from any further analysis. This set now contains all the clusters that have made it through as PC, and any that contain genes the biologists were interested in prior to the clustering.)

9. Perform AHC on $FINAL$. This produces PC[$FINAL$] (the set we're most interested in), PJ[$FINAL$] and O[$FINAL$] (both of these are remnants of clusters we're not interested in, but in rare cases may contain some genes of interest, see cluster 9 in appendix A).

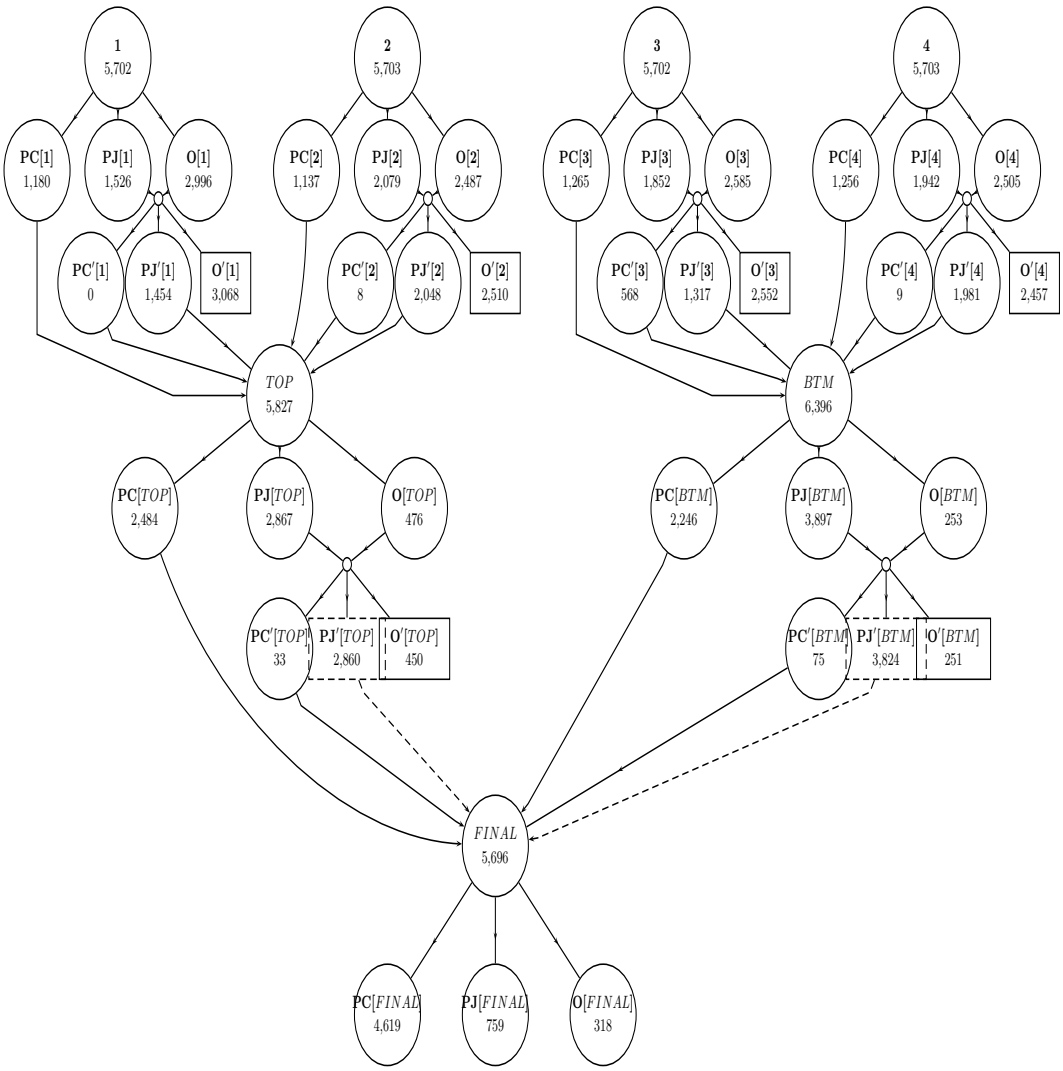

Figure 1: The clustering scheme applied to our data set. The top number in each circle is the name of the set, the bottom number the number of genes in that set. The small circles with no text indicate the conglomeration of PJ and O clusters. The boxes indicate sets of clusters that were dropped at the next step of clustering — where there is a dotted border, one cluster was carried forward to the next step (the cluster containing VIG for PJ*′[TOP] and the cluster containing PKS1-like for PJ*′[BTM]).

Figure 1 shows this process for our data set (where PJ$^{*\prime}[TOP]$ is the cluster that contains the circadian regulated gene VIG and PJ$^{*\prime}[BTM]$ is the cluster that contains the PKS1-like gene potentially involved in the circadian clock, though its involvement has not yet been determined directly experimentally). Here, $FINAL$ contains 5,696 genes in 100 clusters: 4,619 genes are in the 81 PC clusters, 759 are in the 13 PJ clusters and 318 are in the 6 O clusters.

Note that this approach is fully Bayesian in the sense that we are still calculating posterior predictive scores and optimising over them. However, we are searching many more cluster configurations than under greedy search and it is specifically tailored to extracting rhythmically expressed genes via the SHR measure.

The performance of the multi-step AHC

We applied this algorithm to a 13 point time series of 22,810 gene expression profiles taken over a two day period [10]. The log expression profile of some interesting clusters obtained from the set $FINAL$ are given in figure 2. Five of these clusters contain 12 interesting genes with proposed functions in the core of the clock mechanism. GIGANTEA (GI), LUX ARRHYTHMO (LUX), EARLY FLOWERING 4 (ELF4) and TIMING OF CAB EXPRESSION 1 (TOC1) are evening-expressed genes that contribute to the activation of LHY and CCA1 expression (see below) [23–25]. The paralogues of TOC1, PSEUDO RESPONSE REGULATOR genes 3, 5 and 7 (PRR3, PRR5, PRR7), are expressed at intervals throughout the day and also function in the clock mechanism. EARLY PHOTOCHROME RESPONSIVE 1 (EPR1) and COLD AND CIRCADIAN REGULATED 2 (CCR2) are circadian output regulators [26, 27]. PHYTOCHROME INTERACTING FACTOR 4 (PIF4), CRYPTOCHROME 1 (CRY1) and FLAVIN-BINDING KELCH DOMAIN F BOX 1 (FKF1) have varying functions in light perception. Further, CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) and LATE ELONGATED HYPOCOTYL (LHY) are closely-related, morning-expressed genes that function both in the clock mechanism and in regulating output genes [26, 28, 29]. Their profiles are shown in figure 3 of appendix A.

The algorithm places all of these genes in PC clusters. Cluster 73 is an example of a PJ cluster and does not contain any genes earmarked by the biologists. More on the biological implications of the distribution of interesting genes between clusters is discussed in [10]. How the allocation of genes changes between experiments in different conditions is currently under investigation.

All 100 clusters in the *FINAL* are shown in appendix A. In all these clusters, although there is a considerable amount of variation, the different signals can be easily distinguished and is summarised by the Fourier coefficients of the posterior mean.

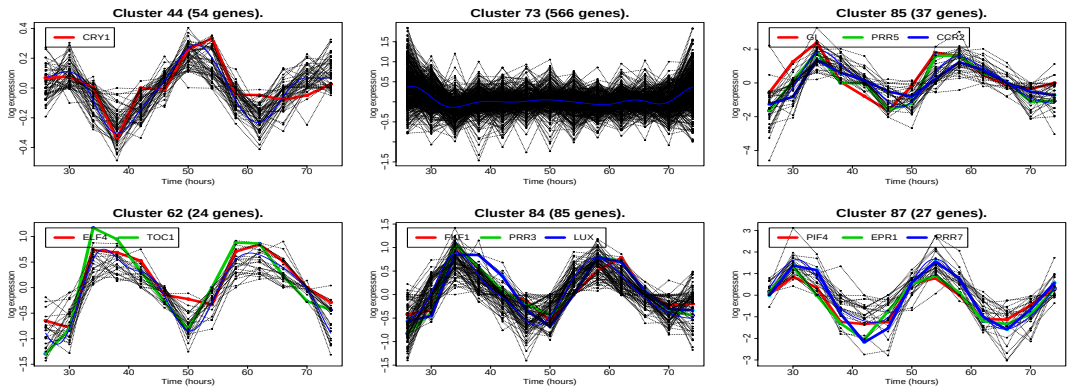

Figure 2: Some example clusters. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in broad lines of red, green and blue. Clusters 44, 62, 84, 85 and 87 are PC and each contain at least one previously studied gene. Cluster 73 is a good example of a PJ cluster.

As noted earlier, the aim is to earmark a set of genes G from the PC clusters that warrant further individual analysis. Detailed biological investigation of single genes can be an expensive and time consuming process. Here, potential candidates may reside in clusters such as 83, 85 and 87 (see figure 2), which each contain three interesting genes already known to the biologists. Once these have been sifted for homologues and genes that are involved in other pathways, there may be an indication of other genes intrinsic to the clock, though it is important to note that this experiment alone will not be definitive proof of whether a gene is in the clock core. The investigation of those genes associated in some way with circadian functionality, G^*, may also be worthwhile. One could specify higher SHR thresholds to do this. For example, to get a more refined shortlist of PC clusters, we could select those clusters with SHR>0.5 as those that should be examined in other microarray or more targeted experiments. Also, it is possible to match the gene clusters with one another via their average profile: classifying groups of clusters as peaking at dawn, peaking at dusk, potentially involved in switching on genes in other clusters, etc. However, the detailed analysis of this classification is beyond the scope of this paper.

The interesting profiles seem to fall into five main shapes: sharply rising then sharply falling and its complement, sharply rising and then drifting back and its complement and an approximately sinusoidal profile. These classifications are explored and explained in appendix A.

The necessity of our modification of the algorithm for this type of application is clearly demonstrated in figure 3. The penultimate classification of the interesting gene PKS1 like was in a PJ cluster. However, re-clustering this set with genes in PC$[TOP]$, PC$'[TOP]$ and PJ$'[TOP]$ allowed this gene to move into a

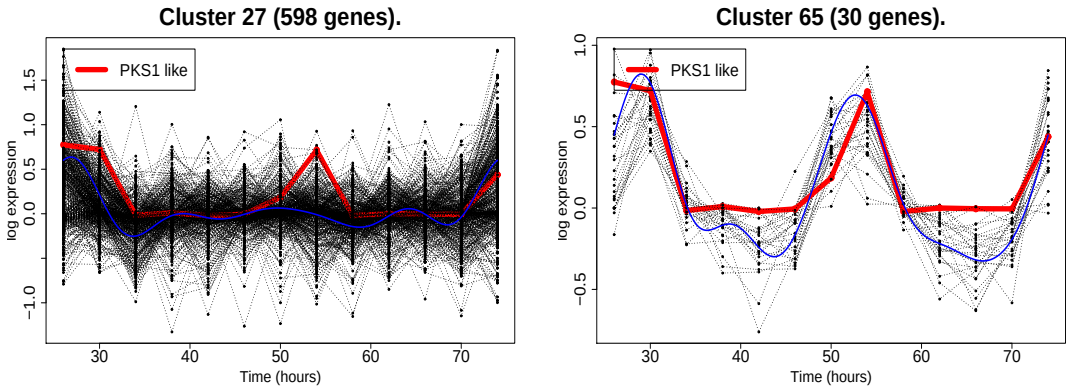

Figure 3: Demonstration of the utility of multi-step clustering. On the left is a PJ cluster from the penultimate step (from the set PJ$^{*\prime}[BTM]$) containing PKS1-like which may be involved in the clock. The cluster on the right shows that PKS1-like ends up in a PC cluster after the final clustering into the 5,696 gene set. This demonstrates vividly that clustering with the "right" genes can draw out PC clusters from PJ clusters.

more suitable and clearly PC cluster. Since it is unguided, the AHC search used in [2] cannot find this better partition whatever the settings of the cluster algorithm's parameters. By exposing sets of genes to different parts of the space at each step, figure 1 shows that PC clusters can be pulled out of members of PJ clusters. This advantage seems to be especially strong when a potentially rhythmic gene is only moderately expressed.

This algorithm can be extended to cluster other time series. The user needs to simply choose an appropriate basis, find a measure that can be used to guide the search, choose a suitable number of sets for step one and tune the hyperparameter v to ensure the log marginal likelihood is maximised (see the methods section for instructions on how to do this).

Conclusions

Guided Bayesian clustering methods like the one described here can enhance the performance of Bayesian clustering algorithms for longitudinal time series whilst respecting the temporal order; something not possible with traditional Euclidean clustering. In particular, our proposed modification explores larger regions of the partition space than greedy search and is more robust to augmentation of the time series (here by inclusion or deletion of subsets of gene profiles). The method we propose may even be enhanced by slowing down the selection procedure and employing more complex moves around the partition space. Also, our method can be used when the model is not conjugate and numerical calculations are required. In both cases however, we then start to lose a major attraction of this method, namely speed.

Perhaps one of the most exciting aspects of a Bayesian clustering algorithm is that it can be used to initialise the in-depth analysis of more structured models. However, for a full and complete Bayesian model of circadian regulatory mechanisms (even under the hypothesis of linear transfer) we obviously need to allow the individual series to be dependent on each other. The cluster models described here assume clusters vary independently and so are not appropriate for this description. Our algorithm is a good first step towards this goal though.

However, the posterior distributions of the clusters in a chosen partition are entirely valid provided that we accept that they are based only on data from the genes they contain. So, as well as providing useful technical information on different sets of PC genes (assuming dependence at the cluster level only), the comparison of the periodogram and group delays are an extremely valuable tool for forming hypotheses of broadly consistent regulatory mechanisms. This sort of interaction between experimental data and statistical algorithms is likely to become an increasingly important part of bioinformatics and systems biology. Furthermore, the posterior distributions obtained for the partition can be used to centre and construct proposal distributions for more elaborate, and therefore necessarily numerical, Bayesian models. The methodology for making this transition from a coarse but easily understood model to a more complex one is analogous to that proposed in [30] in a different context. This could be an interesting and fruitful direction for future research. We have shown that a customised, guided multi-step Bayesian clustering algorithm can be very useful in uncovering and classifying rhythmic time series in large longitudinal data sets.

Methods

Description of the data set

The experimental results concerning the circadian rhythms of the plant *Arabidopsis thaliana* were provided by the co-authors Kieron D. Edwards and Andrew J. Millar. The time series are gene expression levels as measured by Affymetrix microarrays. Essentially, complementary probes for each gene in the organism are placed at random on a glass slide (the microarray or chip) and the sample is washed over the chip. The level of hybridisation between the complementary probe and the gene correspond to the expression level of that gene. Of course, the experimental procedure is more complicated than this and there are real issues in deciphering the signal from the hybridisation level and normalisation. Much more detail is provided in [1]. The advantage of microarrays compared to traditional methods is that many or even all (as is the case for Affymetrix chips) the organism's genes can be measured at once.

The time series were produced from a group of plants grown in the usual diurnal light-dark cycles (12 hours of light, 12 hours of dark) for just over a week. At $t = 0$, at the start of a 12 hour light period, constant white light was shone on the plants. This remained on for the rest of the time course. The first microarray was taken at $t = 26$ hours, with samples every four hours up to $t = 74$ hours. Thus, there are two cycles of data (13 time points) for each of the 22,810 genes available on the *Arabidopsis* microarray chip. Subjective dawn occurs at about $t = 24$ and $t = 48$ hours — this is when the plant has been trained to expect light after 12 hours of darkness.

The data were normalised using the gcRMA routine in GeneSpring v.7.2 and the log expression profiles were clustered. We chose to use the log for three reasons. Firstly, it is most natural to think of the effects as being proportional (i.e. multiplicative) and not additive. Secondly, the variance of highly expressed genes appears much greater than those with lower expression. However, the algorithm implicitly assumes this variance does not depend on the mean. On the log scale, this mismatch is much less pronounced so we can expect to find more appropriate clusters. Finally, gcRMA normalisation is centred around the median which is invariant to increasing transformations like this one. A more detailed exposition of the experiment is given in [10].

Details of the single-step Bayesian AHC algorithm

As discussed earlier, partitions can be scored and compared. The score function here, for each cluster c, is given by the log marginal likelihood:

$$\lambda_{c(\boldsymbol{y})} \equiv \log p_{c(\boldsymbol{y})} = \log \left(\int \int p_c(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2) p_c(\boldsymbol{\beta}|\sigma^2) p_c(\sigma^2) \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \right) \tag{3}$$

Thus the score measures the probability that time series in the same cluster arose from the same underlying stochastic process. As a consequence of the independence assumption, the score λ_ϕ of a particular partition ϕ consisting of n clusters is simply $\sum_{i=1}^{n} \lambda_{c_i(\boldsymbol{y})}$. To proceed in a Bayesian setting we need to specify priors on $\boldsymbol{\beta}$ and σ^2.

Because of the magnitude of the model space and to avoid time consuming numerical methods, the exploratory nature of this methodology is suited to a conjugate analysis as recommended in [2, 8, 21, 22, 31]. Provided such priors are plausible, enormous gains in efficiency can then be made. We therefore use the well established normal inverse-gamma conjugate family à la [2]. Thus

$$\boldsymbol{\beta}|\sigma^2 \sim \mathrm{N}(\boldsymbol{m}, \sigma^2 \boldsymbol{V}) \quad \text{and} \quad \sigma^2 \sim \mathrm{IGamma}\left(\frac{\alpha}{2}, \frac{\gamma}{2}\right) \tag{4}$$

That is,

$$p_c(\boldsymbol{\beta}|\sigma^2) = \frac{1}{(2\pi\sigma^2)^{p/2}|\boldsymbol{V}|^{1/2}} \exp\left(\frac{-(\boldsymbol{\beta}-\boldsymbol{m})'\boldsymbol{V}^{-1}(\boldsymbol{\beta}-\boldsymbol{m})}{2\sigma^2}\right)$$

$$p_c(\sigma^2) = \frac{(\gamma/2)^{\alpha/2}}{\Gamma(\alpha/2)}(\sigma^2)^{-(\frac{\alpha}{2}+1)} \exp(-\gamma/2\sigma^2)$$

$\boldsymbol{V}$ is the prior covariance matrix and $\boldsymbol{m}$ is a vector. Note that this regression model is a Gaussian mixture as the marginal distribution of the error variance, σ^2, is inverse gamma. We experimented with different settings of the hyperparameters. For the purposes of illustration, we chose the covariance matrix $\boldsymbol{V} = v\boldsymbol{I}$ with $v = 0.498$ and all other parameters set to their defaults in the code available at [32]. v is a very important parameter in the algorithm so, as in [2], we explored the log marginal likelihood over a grid of values of v (v=0.1 to 10,000, see appendix B for details).

Note that the Bayesian clustering method described below is based on the relative marginal likelihood of different models. Proper priors must be used so that the Bayes factors give good model selection characteristics [9]. Here, this means that $\alpha, \gamma > 0$. Throughout we set $\boldsymbol{m} = 0$ so that the Fourier coefficients are shrunk towards zero. Here, with scant information this simply shrinks the expression profile to one which varies less.

Standard results [9] state that the four parameters $\boldsymbol{m}$, $\boldsymbol{V}$, α and γ are updated in the prior-to-posterior analysis in the following way:

$$\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2 \sim \mathrm{N}(\boldsymbol{m}^*, \sigma^2\boldsymbol{V}^*) \tag{5}$$

$$\sigma^2|\boldsymbol{y} \sim \mathrm{IGamma}\left(\frac{NT+\alpha}{2}, \frac{d+\gamma}{2}\right) \tag{6}$$

with

$$\boldsymbol{V}^* = \left(\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{V}^{-1}\right)^{-1}, \quad \boldsymbol{m}^* = \boldsymbol{V}^*\boldsymbol{B}'\boldsymbol{y} \quad \text{and}$$

$$d = \boldsymbol{y}'\left(\boldsymbol{I} - \boldsymbol{B}\boldsymbol{V}^*\boldsymbol{B}'\right)\boldsymbol{y}$$

where $'$ denotes the transpose.

With this choice of priors, $p_{c(\boldsymbol{y})}$ can be calculated explicitly as a multivariate t-distribution [9]:

$$p_{c(\boldsymbol{y})} = \left(\frac{1}{\pi}\right)^{NT/2} \frac{\gamma^{\alpha/2}\Gamma\left(\frac{NT+\alpha}{2}\right)}{\Gamma\left(\frac{\alpha}{2}\right)} \frac{|\boldsymbol{V}^*|^{1/2}}{|\boldsymbol{V}|^{1/2}} \frac{1}{(d+\gamma)^{(NT+\alpha)/2}} \tag{7}$$

$$= g(NT, \alpha, \gamma)\,|\boldsymbol{V}|^{-1/2} \frac{1}{\left|\boldsymbol{B}'\boldsymbol{B} + \boldsymbol{V}^{-1}\right|^{1/2} (d+\gamma)^{(NT+\alpha)/2}} \tag{8}$$

Let $C(\phi)$ denote the collection of clusters associated with the partition ϕ. Then, assuming the parameters of different clusters are independent, the score $\Sigma(\phi)$ for any partition ϕ with clusters $c \in C(\phi)$ is given by

$$\Sigma(\phi) = \sum_{c \in C(\phi)} \log p_{c(\boldsymbol{y})} + \log \pi(\phi) \tag{9}$$

where $\log \pi(\phi)$ is a known function of the prior distribution over partitions, typically chosen to depend only on the number of clusters in ϕ. Examples of good choices of $\log \pi(\phi)$ are given in [2].

Useful partitions of the profiles will have high scores. One pleasant feature of this class of models is that the difference between two partitions identical outside a given set c, will depend only on their relative scores over this set of profiles. In particular, suppose they differ only on a set $c = c_1 \cup c_2$ with c_1 and c_2 disjoint. If ϕ^+ and ϕ^- are such that ϕ^+ contains the same clusters as ϕ^- except that cluster c in ϕ^- is separated into two clusters c_1 and c_2 in ϕ^+, then ϕ^+ and ϕ^- are said to be *adjacent*. Thus,

$$\Sigma(\phi^+) - \Sigma(\phi^-) = \log p_{c_1(\boldsymbol{y}_1)} + \log p_{c_2(\boldsymbol{y}_2)} - \log p_{c(\boldsymbol{y})} + \lambda(\phi^+, \phi^-) \tag{10}$$

where $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2)$ and $\lambda(\phi^+, \phi^-) = \pi(\phi^+) - \pi(\phi^-)$ is often set a priori to a fixed constant not depending on ϕ^+ or ϕ^-. Since for moderate sized clusters the functions in the expression above can be calculated almost instantaneously, it is trivial to check whether merging or splitting two clusters to form a new partition is more or less well supported by the data.

The popular AHC is employed to search the space of partitions in [2]. Because this only modifies a partition to one adjacent to it in the sense above, even with large numbers of genes it is possible to find partitions that cluster together genes with similar longitudinal profiles quickly. The AHC starts with each of the N gene profiles in N separate clusters with fixed values of the hyperparameters $\boldsymbol{V}$, α and $\boldsymbol{\beta}$. We then form a sequence of new partitions by sequentially merging two clusters, thus decreasing the number of clusters by one. The clusters that increase the score by the most (or reduce the score by the least) are combined. We repeat until we reach the last partition which has one cluster with all N genes. Since the marginal likelihood has been calculated for each of the partitions, containing 1 to N clusters, we can simply choose the one with the highest score. As mentioned previously, this iterative optimisation of a marginal likelihood criterion is fast, but clearly leaves most of the partition space unsearched.

List of abbreviations

AHC Agglomerative Hierarchical Clustering

O Other

PC Potentially Circadian

PJ Potentially Junk

SHR Second Harmonic Ratio

Authors' contributions

KDE and AJM designed and carried out the experiment, acted as consultants in the biological interpretation of the clustering results and contributed to the biological explanations in the paper. PEA proposed and coded the multi-step clustering method and generated the figures. JQS proposed the change in basis function, which was implemented by PEA by adapting code available at [32], and produced the theoretical motivation and framework for the multi-step clustering. The text was written by PEA and JQS. All authors read and approved the final manuscript.

Acknowledgements

We thank Nick Heard and Chris Holmes for useful discussions and David Rand for reading an earlier draft of the manuscript. KDE's experimental work was supported by BBSRC grant G19886 to AJM. PEA, JQS and AJM are part of the Interdisciplinary Program for Cellular Regulation (IPCR) based at the University of Warwick and thank EPSRC, BBSRC and BioSim for support.

References

1. Draghici S: *Data Analysis Tools for DNA Microarrays*. Chapman and Hall 2003.
2. Heard NA, Holmes CC, Stephens DA: **A Quantitative Study of Gene Regulation Involved in the Immune Response of Anopheline Mosquitoes: An Applictaion of Bayesian Hierarchical Clustering of Curves**. *J. Amer. Statist. Assoc.* 2006, **101**(473):18–29.
3. 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* 2002, **22**:9305–9319.
4. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesch JB: **Coordinated transcription of key pathways in the mouse by the circadian clock**. *Cell* 2002, **109**:307–320.
5. Straume M: **DNA microarray time series analysis: automated statistical assessment of circadian rhythms in gene expression patterning**. *Methods Enzymol* 2004, **383**:149–166.
6. Andersson CR, Isaksson A, Gustafsson MG: **Bayesian detection of periodic mRNA time profiles without use of training examples**. *BMC Bioinformatics* 2006, **7**(63). [Published online on 9th Feb 2006. doi:10.1186/1471-2105-7-63].
7. Andesmäki M, Lähdesmäki H, Pearson R, Huttunen H, Yli-Harja O: **Robust detection of periodic time series measured from biological systems**. *BMC Bioinformatics* 2005, **6**(117).
8. Heard NA, Holmes CC, Stephens DA, Hand DJ, Dimopoulos G: **Bayesian co-clustering of Anopheles gene expression time series: a study of immune defense responses to multiple experimental challenges**. *Proc. Nat. Acad. Sci.* 2005. [Doi:10.1073/pnas.0408393102].

9. Denison DGT, Holmes CC, Mallick BK, Smith AFM: *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics, John Wiley and Sons 2002.

10. 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**. *The Plant Cell (in press)* 2006. [Published online on 10th Feb 2006. doi:10.1104/tpc.105.038315].

11. Locke JC, Southern MM, Kozma-Bognar L, Hibberd V, Brown PE, Turner MS, Millar AJ: **Extension of a genetic network model by iterative experimentation and mathematical analysis**. *Mol Sys Biol 1* 2005. [Doi:10.1038/msb4100018].

12. Brockwell PJ, Davis RA: *Introduction to Time Series and Forecasting*. Springer 1996.

13. Hannan EJ: *Multiple Time Series*. John Wiley and Sons 1970.

14. Banfield JD, Raftery AE: **Model-based gaussian and non-gaussian clustering**. *Biometrics* 1993, **49**:803–821.

15. Fraley C, Raftery AE: **Model-based clustering, discriminant analysis, and density estimation**. *J. Amer. Statist. Assoc.* 2002, **97**:611–631.

16. Luan Y, Li H: **Clustering of time-course gene expression data using a mixed-effects model with B-splines**. *Bioinformatics* 2003, **19**:474–482.

17. Ramoni M, Sebastiani P, Kohane PR: **Cluster analysis of gene expression dynamics**. *Proc. Nat. Acad. Sci.* 2002, **99**:9121–9126.

18. Wakefield J, Zhou C, Self S: **Modelling gene expression over time: curve clustering with informative prior distributions**. In *Bayesian Statistics 7*. Edited by Bernardo JM, Bayarri MJ, Berger JO, Dawid AP, Heckerman D, Smith AFM, West M, Oxford University Press 2003.

19. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL: **Model-based clustering and data transformations for gene expression data**. *Bioinformatics* 2001, **17**:977–987.

20. Boucheron S, Bousquet O, Lugosi G: **Theory of Classification: A Survey of Recent Advances**. *ESAIM: Probability and Statistics* 2005, **9**:323–375.

21. Chipman H, George E, McCullough R: **Bayesian CART Model Search**. *J. Amer. Statist. Assoc.* 1998, **93**:935–960.

22. Chipman HA, George EI, McCulloch RE: **Bayesian treed models**. *Machine Learning* 2002, **48**(1–3):299–320.

23. Doyle MR, Davis SJ, Bastow RM, McWatters HG, Kozma-Bognar L, Nagy F, Millar AJ, Amasino RM: **The ELF4 gene controls circadian rhythms and flowering time in Arabidopsis thaliana**. *Nature* 2002, **419**:74–77.

24. Alabadi D, Oyama T, Yanovsky MJ, Harmon FG, Mas P, Kay SA: **Reciprocal regulation between TOC1 and LHY/CCA1 within the Arabidopsis circadian clock**. *Science* 2001, **293**:880–883.

25. Hazen SP, Schultz TF, Pruneda-Paz JL, Borevitz JO, Ecker JR, Kay SA: **LUX ARRHYTHMO encodes a Myb domain protein essential for circadian rhythms**. *Proc. Nat. Acad. Sci.* 2005, **102**:10387–10392.

26. Kuno N, Moller SG, Shinomura T, Xu X, Chua NH, Furuya M: **The Novel MYB Protein EARLY-PHYTOCHROME-RESPONSIVE1 is a Component of a Slave Circadian Oscillator in Arabidopsis**. *The Plant Cell* 2003, **15**:2476–2488.

27. Heintzen C, Nater M, Apel K, Staiger D: **AtGRP7, a nuclear RNA-binding protein as a component of a circadian-regulated negative feedback loop in Arabidopsis thaliana**. *Proc. Nat. Acad. Sci.* 1997, **94**:8515–8520.

28. Alabadi D, Yanovsky MJ, Mas P, Harmer SL, Kay SA: **Critical role for CCA1 and LHY in maintaining circadian rhythmicity in Arabidopsis**. *Curr. Biol.* 2002, **12**:757–761.

29. Mizoguchi T, Wheatley K, Hanzawa Y, Wright L, Mizoguchi M, Song HR, Carre IA, Coupland G: **LHY and CCA1 are partially redundant genes required to maintain circadian rhythms in Arabidopsis**. *Dev. Cell* 2002, **2**:629–641.

30. Aguilar O, West M: **Bayesian dynamic factor models and portfolio allocation**. *Journal of Business and Economic Statistics* 2000, **18**:338–357.

31. George EI, Clyde M: **Model uncertainty**. *Statistical Science* 2004, **19**:81–94.

32. Heard NA: **Code for the algorithm described in Heard *et al*, J. Amer. Statist. Assoc., 2006**. *http://stats.ma.imperial.ac.uk/naheard/public_html* 2006.

Appendix A: Plots and classification of the FINAL set clustering

As mentioned in the results and discussion section of the main paper, we can classify the posterior mean profiles into various shapes that will be helpful in eventually examining the biological implications in more detail. The first five classifications cover the clusters identified as circadian over both 24-hour periods; the last five those that aren't. Types I, VI and VII are delineated by objective criteria whilst the remaining types are classified by eye. This is nevertheless useful as a guideline to the broad classes of behaviour that are displayed.

I) Sinusoidal: those clusters with SHR>0.65 and more than 11 genes.

II) Sharply rising then sharply falling.

III) Sharply falling then sharply rising.

IV) Sharply rising then drifting back to zero.

V) Sharply falling then drifting back to zero.

VI) Potential junk: those classified as PJ by the algorithm.

VII) Other: those classified as O by the algorithm.

VIII) Potentially circadian, but not repeated: clusters with a peak or trough in one 24 hour period, but not in the other.

IX) Outliers: clusters containing less than 11 genes.

X) Not interesting: clusters with expressions close to zero and non-circadian profiles.

The profiles of each of the 100 clusters identified from the 5,696 gene $FINAL$ set are shown in figures 4 to 19 classified in order according to the ten types above. Within each type, the clusters are sorted by their phase by maximum (the maximum value of the posterior mean in the first 24 hours). The second harmonic ratio (SHR) and phase by maximum (PBM) are given on each plot.

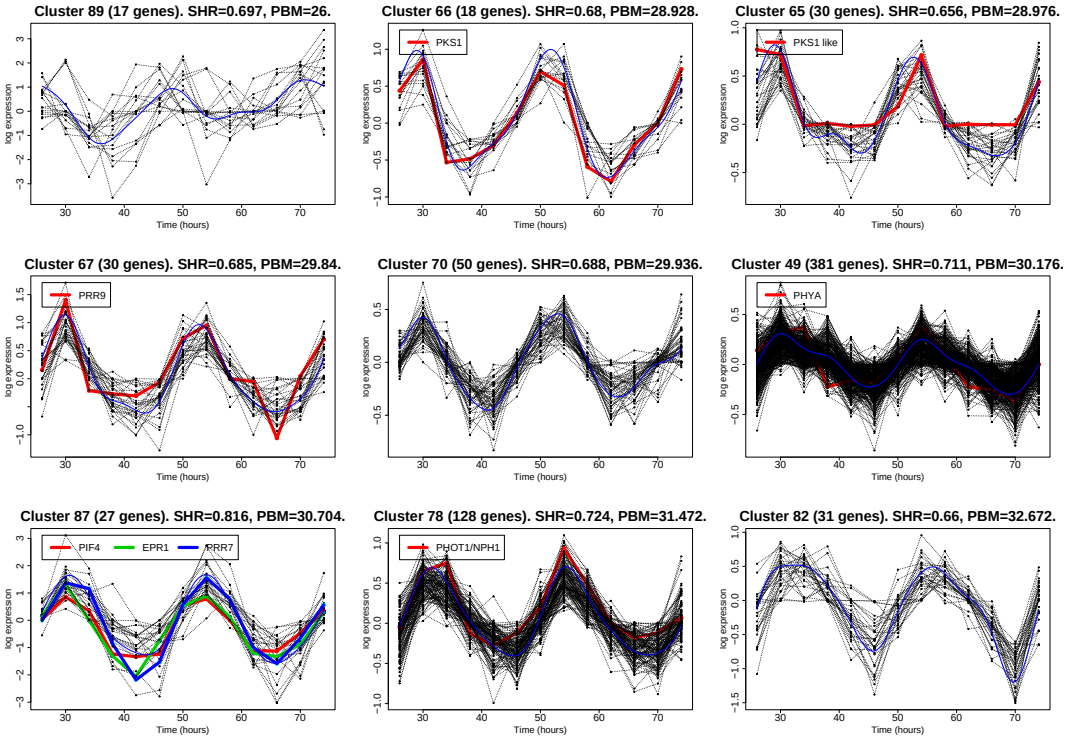

Figure 4: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

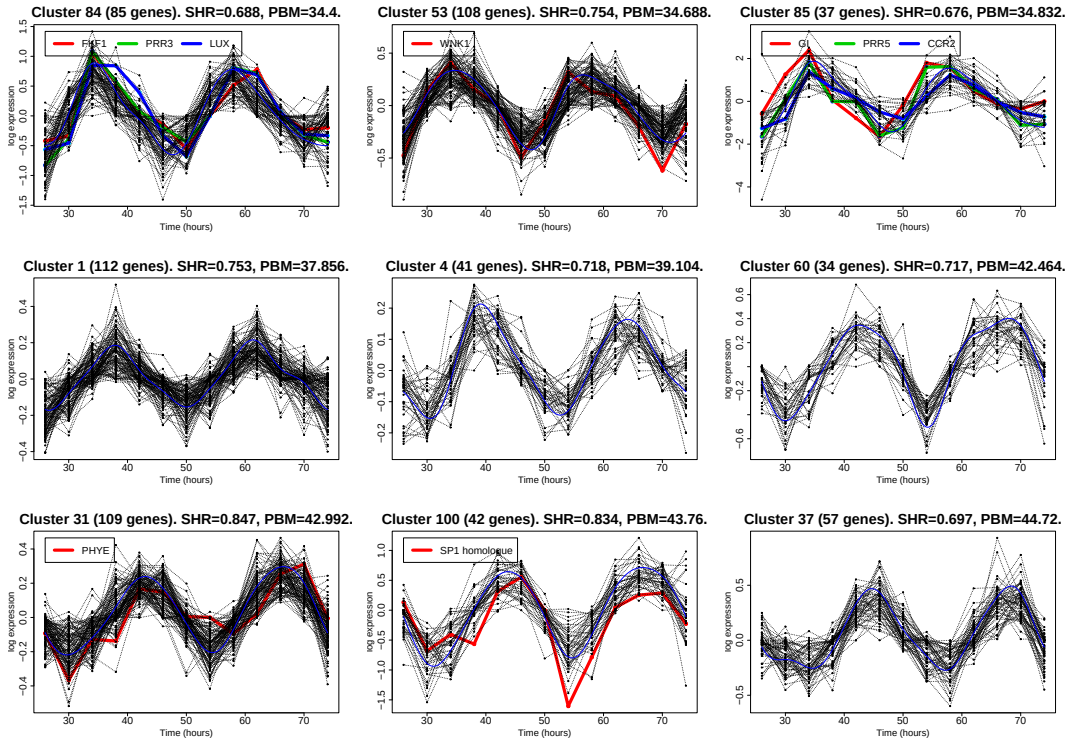

Figure 5: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

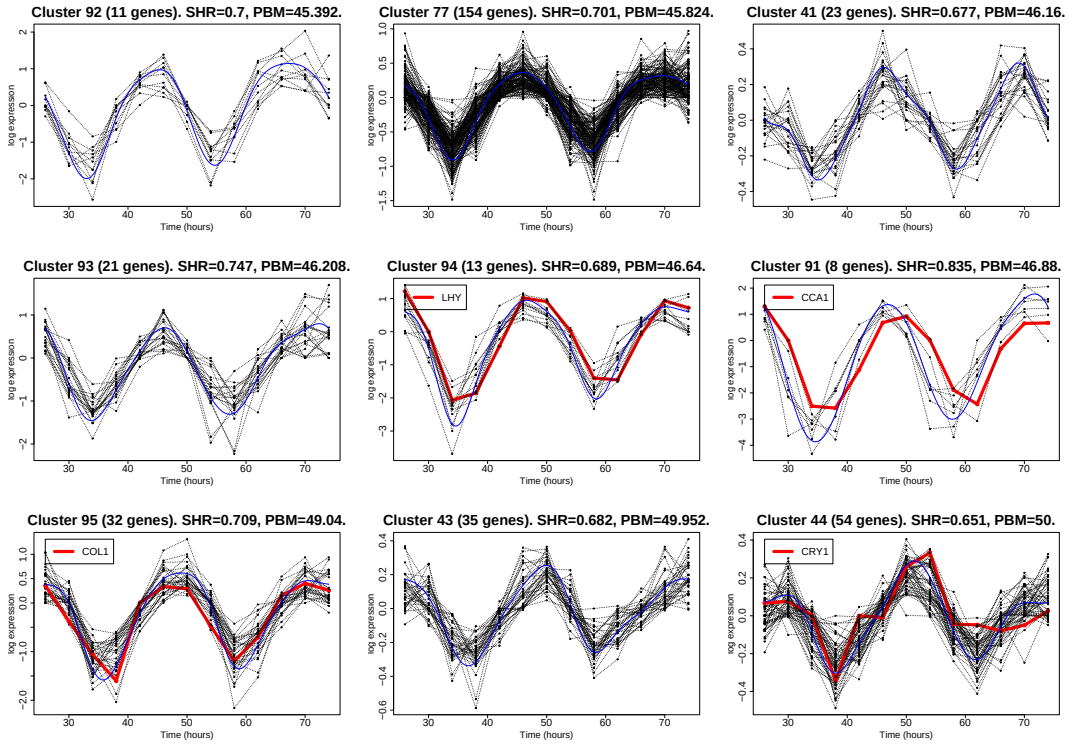

Figure 6: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

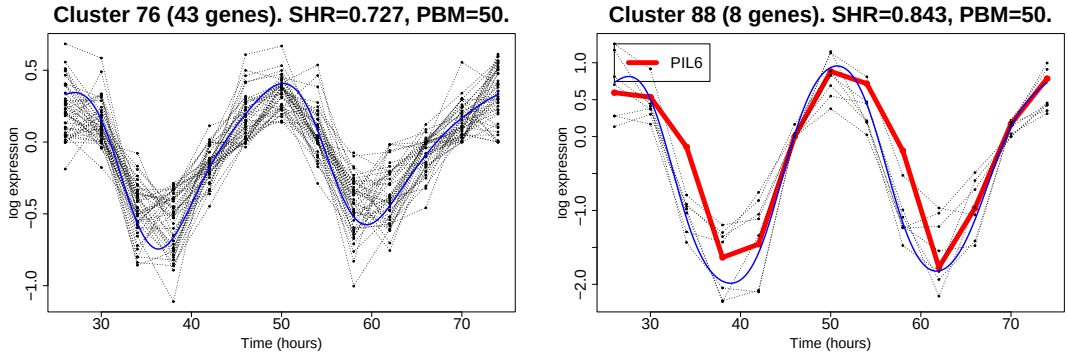

Figure 7: Type I clusters: sinusoidal. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

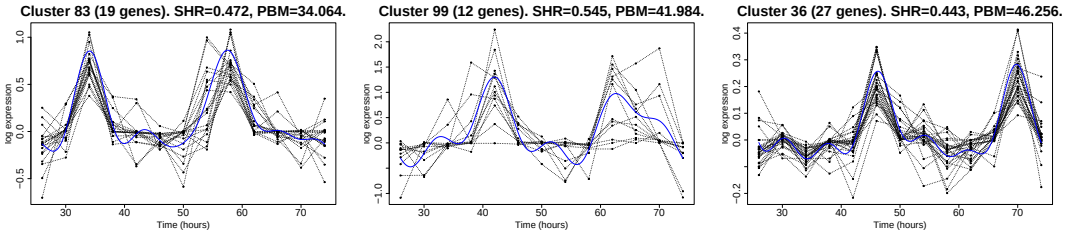

Figure 8: Type II clusters: sharply rising then sharply falling. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

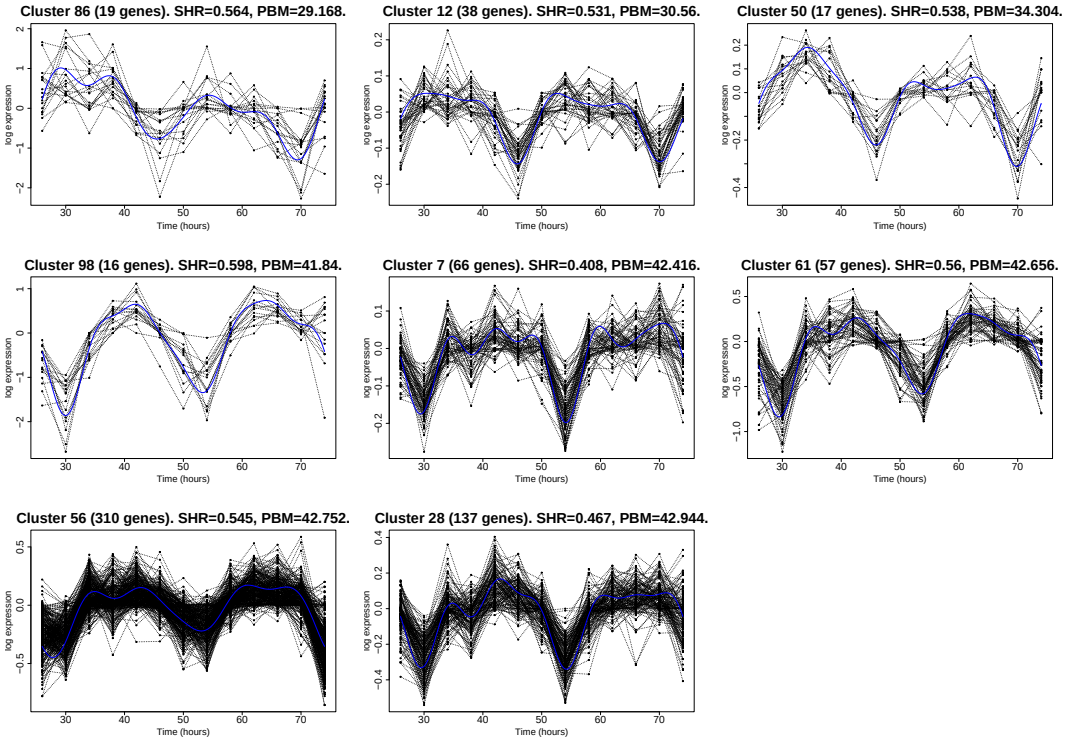

Figure 9: Type III clusters: sharply falling then sharply rising. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

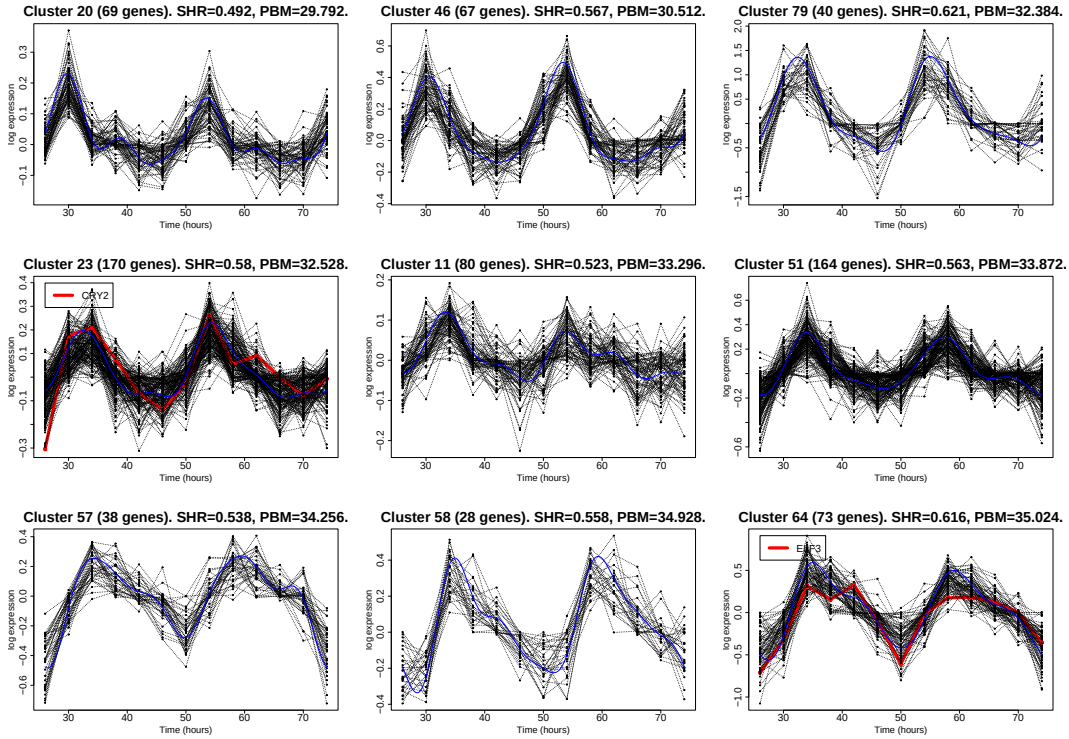

Figure 10: Type IV clusters: sharply rising then drifting back to zero. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

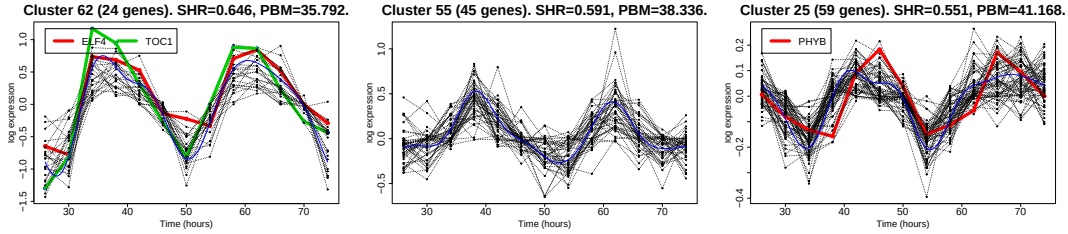

Figure 11: Type IV clusters: sharply rising then drifting back to zero. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

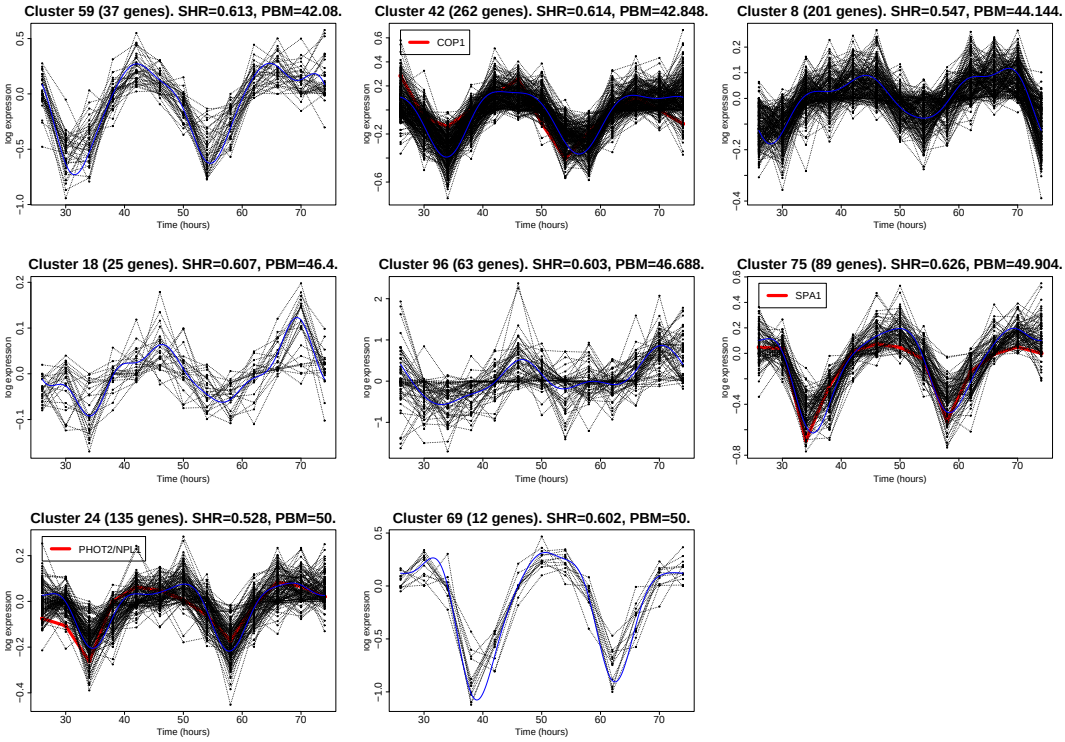

Figure 12: Type V clusters: sharply falling then drifting back to zero. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

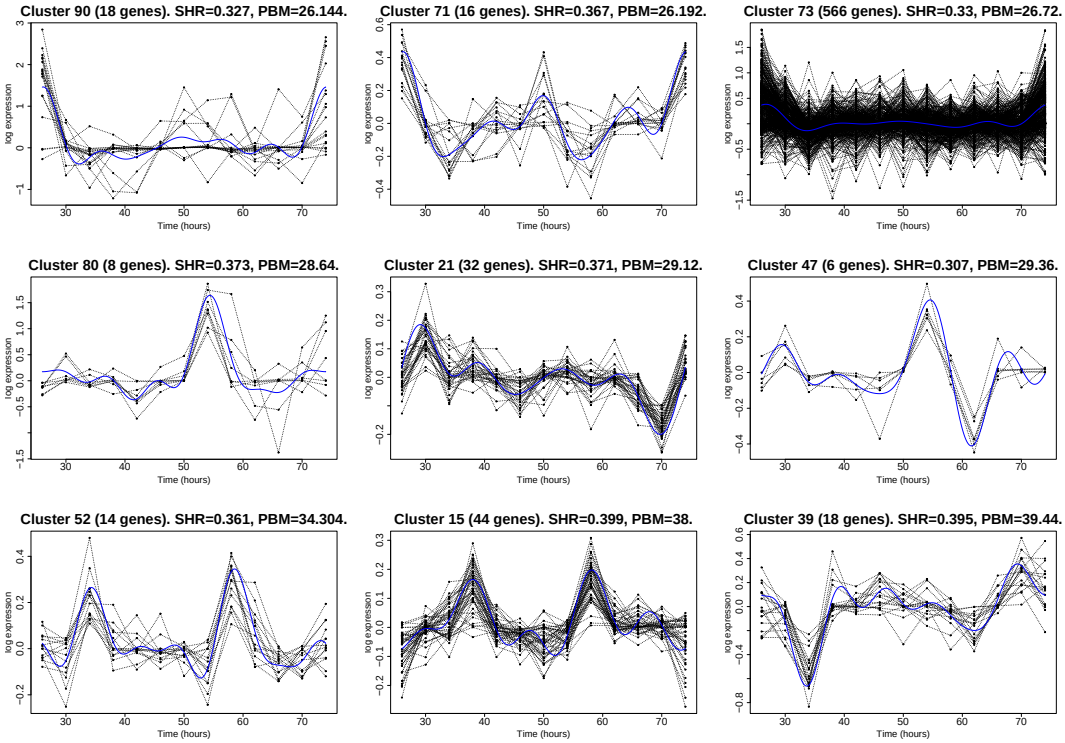

Figure 13: Type VI clusters: potential junk. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

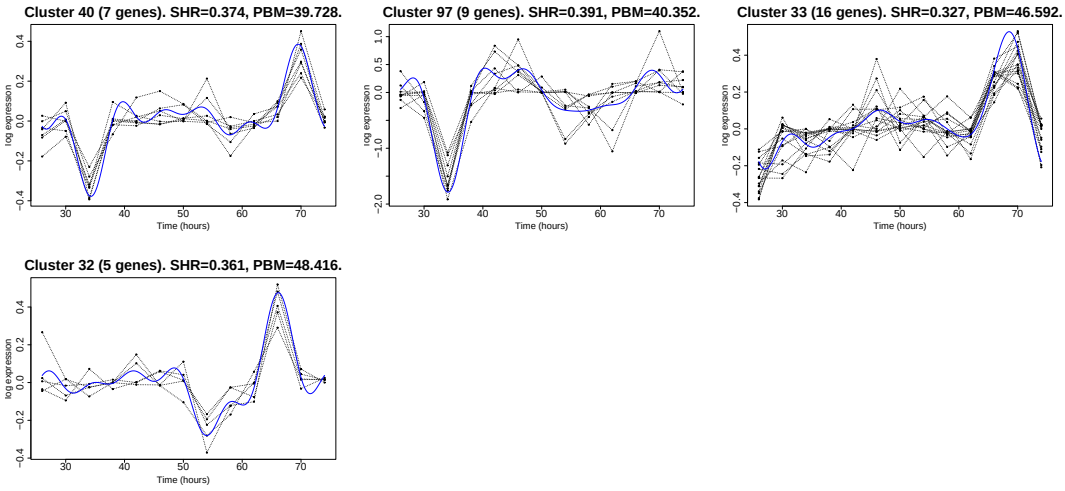

Figure 14: Type VI clusters: potential junk. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

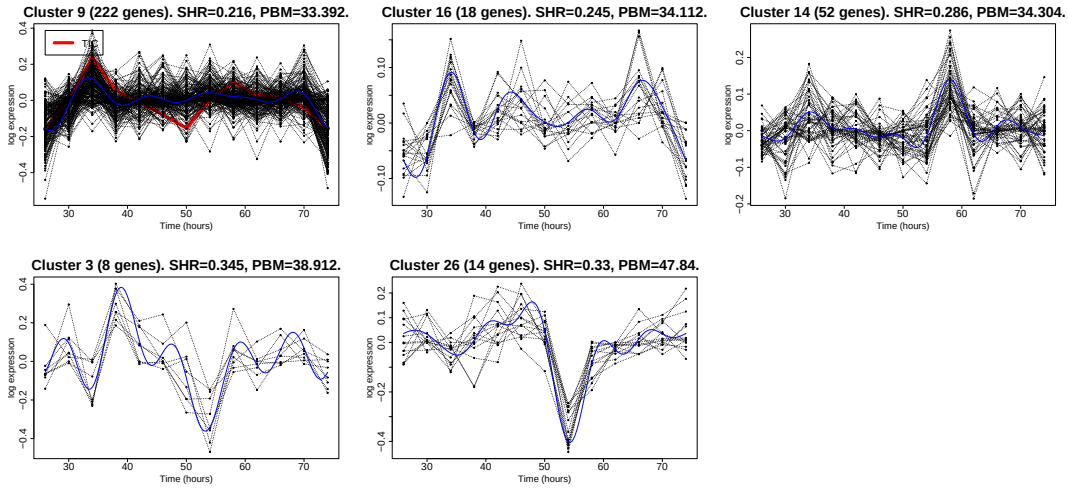

Figure 15: Type VII clusters: other. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster. Genes marked as interesting by the biologists are highlighted in thicker lines of red, green and blue.

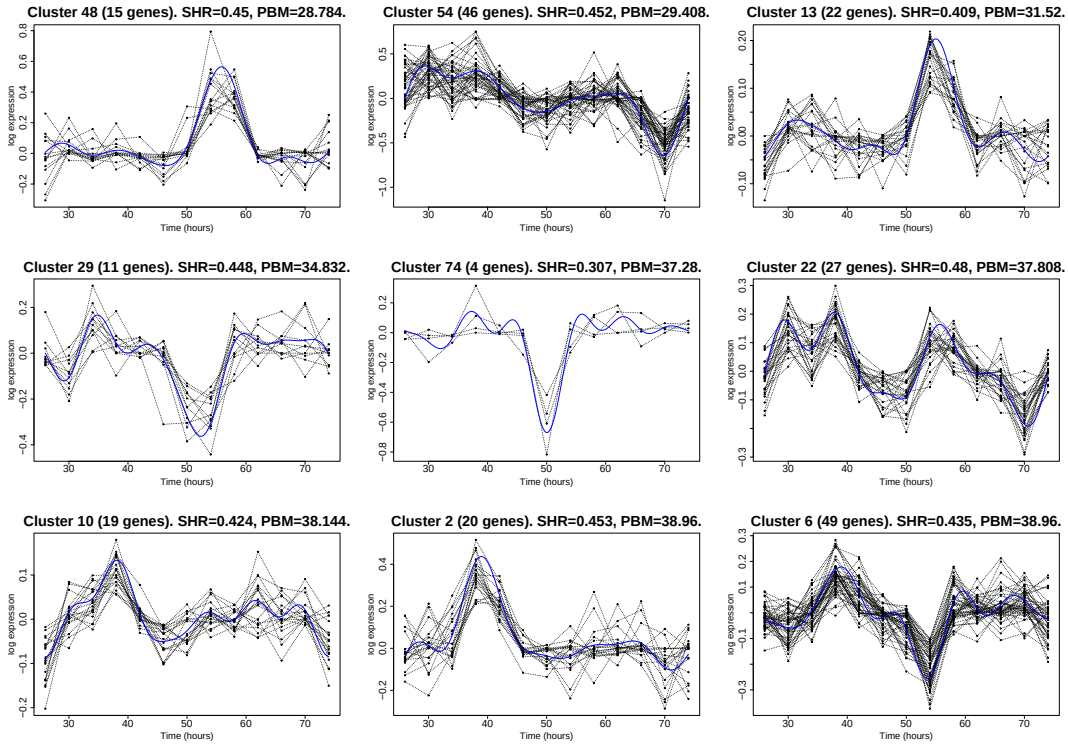

Figure 16: Type VIII clusters: potentially circadian, but not repeated. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

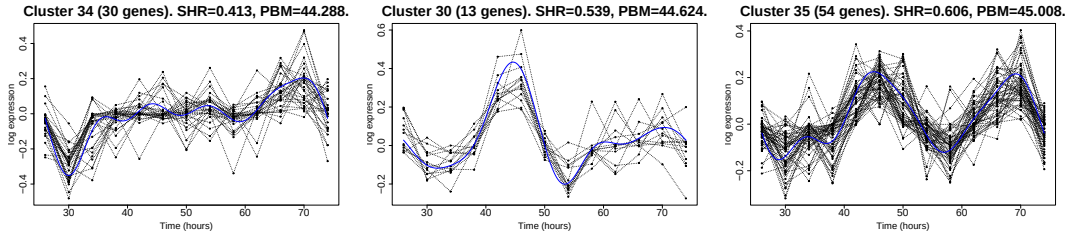

Figure 17: Type VIII clusters: potentailly circadian, but not repeated. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

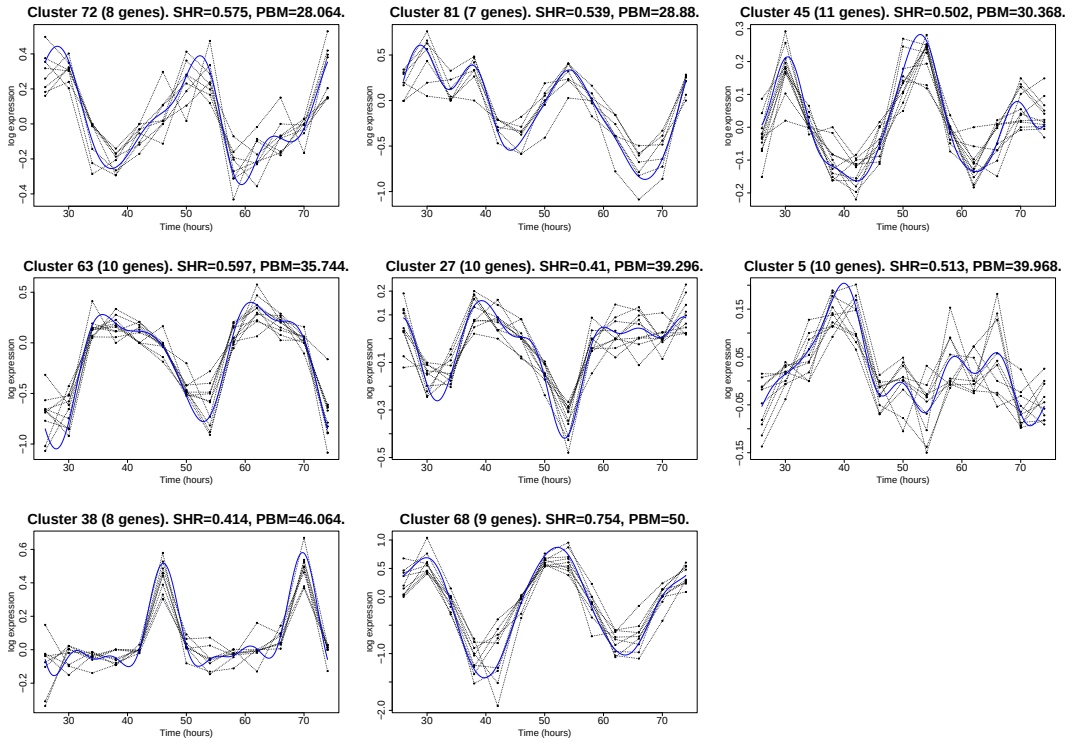

Figure 18: Type IX clusters: outliers. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

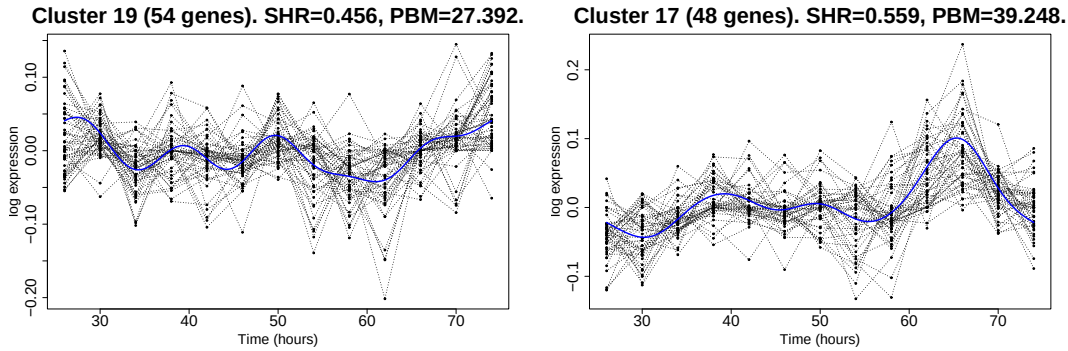

Figure 19: Type X clusters: not interesting. The black dots are the data points (joined by dotted lines) and the blue line is the posterior mean for the cluster.

Appendix B: Robustness analysis

To illustrate the problems of using AHC unmodified, consider the case when we optimise the log marginal likelihood by varying the parameter v. The scores for a range of values from $v = 0.1$ to $v = 10,000$ are shown in figure 20, with a close-up on the values between $v = 0.4$ and $v = 0.6$ in figure 21. The corresponding optimal number of clusters are given in figure 22.

Because of its explicit algebraic form, we should expect the log-marginal likelihood (as a function of the hyperparameter) to be continuous and differentiable with a moderate second derivative. The numerical results don't seem to accord with this. Thus we have strong (if circumstantial) evidence that the AHC isn't consistently finding a global maximum, since when the hyperparameter is perturbed slightly the score can take quite different values. Such a phenomenon has encouraged other authors ([21, 22]) to develop more sophisticated searches through partition space in the related Bayesian CART model. However, in this application, provided we guard against certain types of misclassification in the way we describe, this instability is not too much of a problem because the corresponding influence on the expected utility is small.

This can be seen by considering the pattern of clusters we obtain if we use a very different value of the hyperparameter, say, $v = 10,000$ instead of $v = 0.498$. The value $v = 0.498$ was chosen throughout this paper as it provided the highest log marginal likelihood on clustering of another set of genes. (Namely, the final set of 2,845 genes obtained with $v = 10,000$ at each stage of the clustering process.) By using the same v at each stage, the PC, PJ and O sets were generated consistently.

For the same set of 5,696 genes, the value $v = 10,000$ gives us far fewer, but larger, clusters (34 instead of 100) so that we get intrinsically different solutions. Despite this, the estimated profiles of most genes do not radically differ under changes in v. This means that the sets of genes identified to have interesting profiles do not change greatly over large ranges of v. Furthermore, genes having similar profiles remain close. The fact that clusters with profiles which are similar are combined doesn't have a big impact in either our choice of promising genes for future analysis, or the classification of genes as potentially circadian.

Figures 23 and 24 show two phase plots of the clusters for the $FINAL$ gene set clustered with the extremely different hyperparameters $v = 0.498$ and $v = 10,000$ respectively. (Their meaning is described in the figure captions. Note that the phase by maximum of a cluster is the maximum value of the posterior mean in the first 24 hours.) The spots are distributed similarly (including the relative sparseness of large clusters between 26 and 28, and 36 and 42 hours) and the larger spots remain in the same sectors of the plot indicating that the structure of the two partitions is broadly the same.

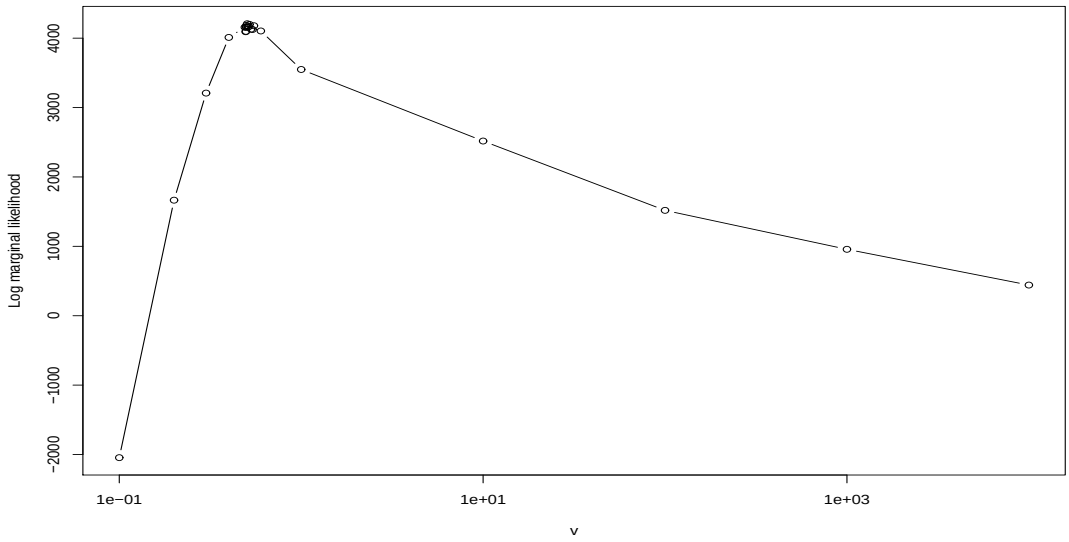

Figure 20: The log marginal likelihood (score) of the $FINAL$ set clustered with a range of v values.

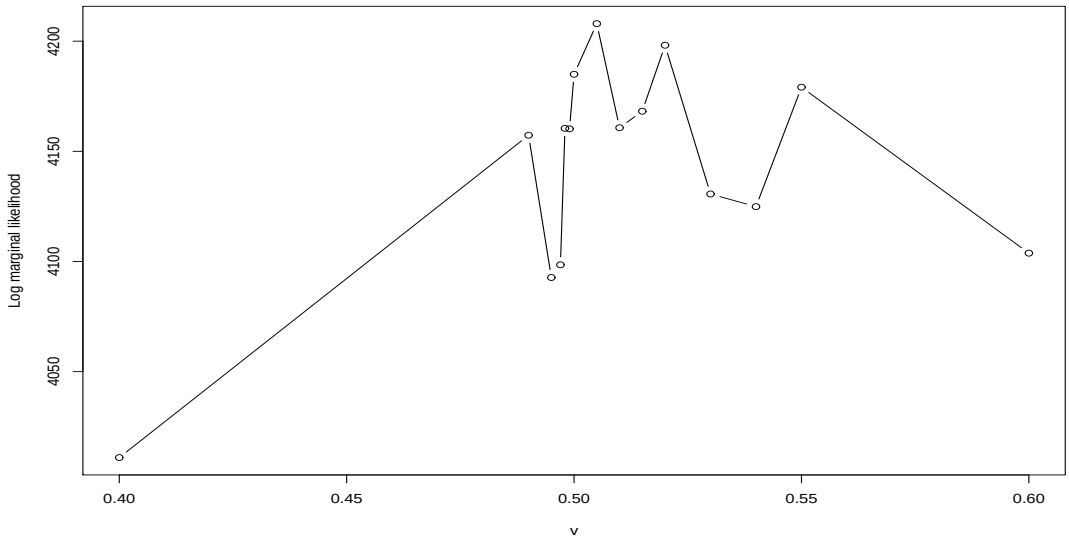

Figure 21: The same data as figure 20, shown over a smaller range of v values.

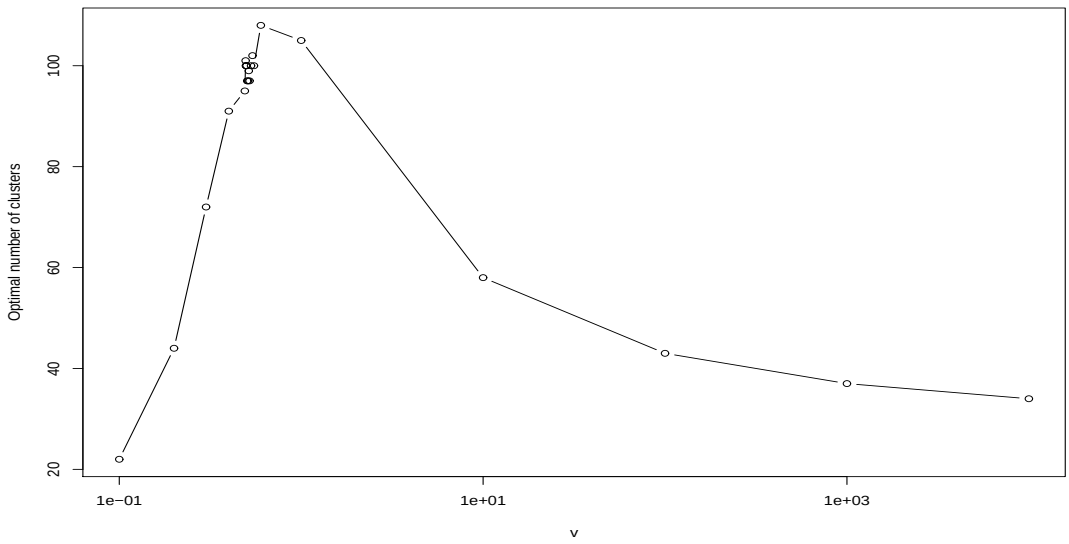

Figure 22: The number of clusters in the $FINAL$ set clustered with a range of v values.

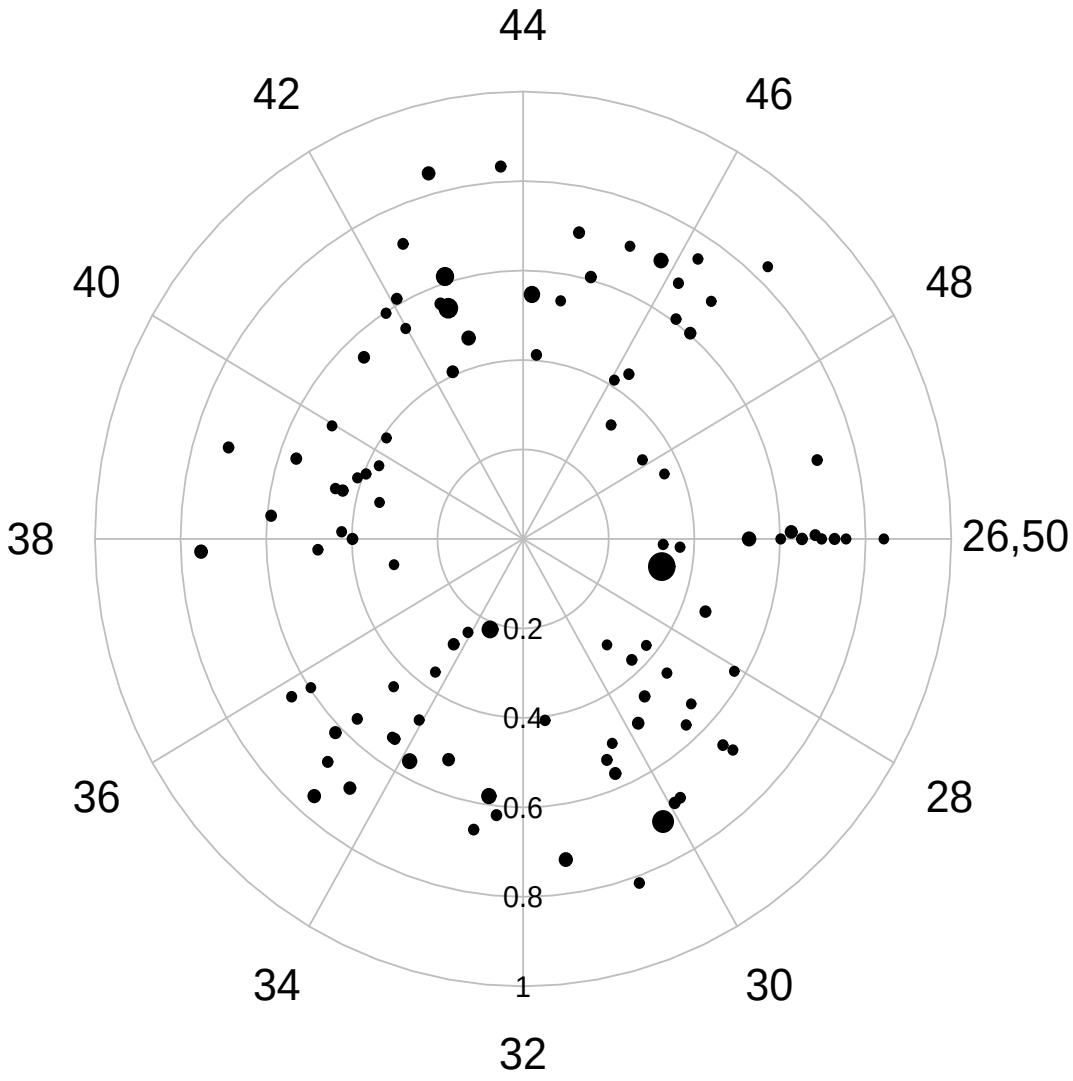

Figure 23: Phase plot of the clusters of the $FINAL$ gene set with $v = 0.498$. Each dot is one cluster, its radius is proportional to the number of genes it contains. Its distance from the origin gives its second harmonic ratio, and the angle indicates the phase by maximum of the posterior mean profile.

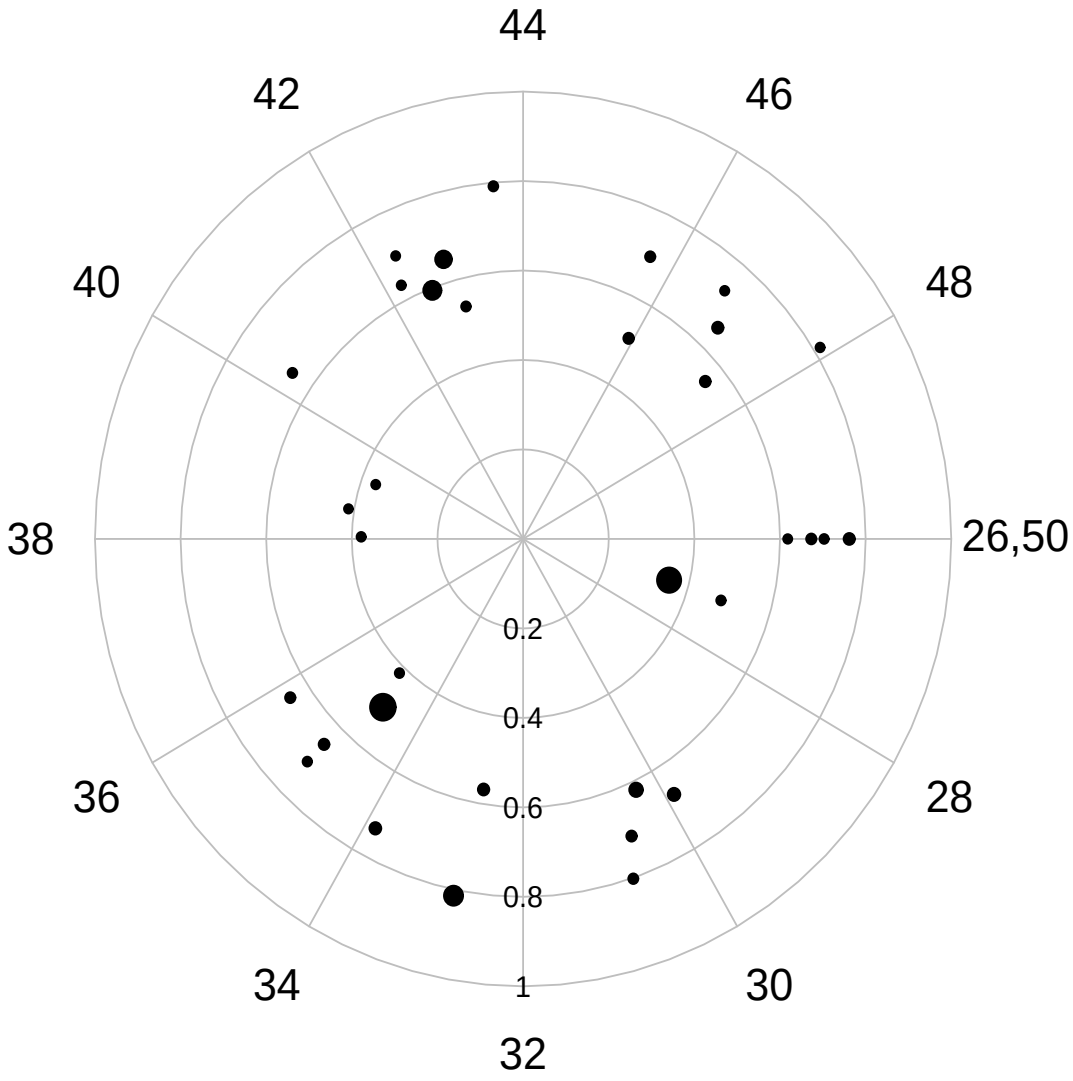

Figure 24: Phase plot of the clusters of the $FINAL$ gene set with $v = 10,000$. Each dot is one cluster, its radius is proportional to the number of genes it contains. Its distance from the origin gives its second harmonic ratio, and the angle indicates the phase by maximum of the posterior mean profile. The structure is broadly similar to that of figure 23.