Grant ideas: QTL

Daphne[1], seth.davis[1], and marina.knight[1]

[1]Affiliation not available

January 20, 2020

Questions:

- Which seasonality phenotypes do we care about/have good preliminary data for?
- What would be a good example of photoperiod transition? (how subtle/realistic can we get while still seeing the 'blip' in CCR2 expression? Would this be necessary preliminary data?)
- Do we need any extra preliminary evidence that circadian clock parameters are associated with seasonality phenotypes?
- What do we know about ODE models of photoperiod sensing vs. circadian clock?
- Homework for me: read more about interval mapping

Introduction

Although all plants on earth experience a 24-hour day-night cycle, there is substantial variability in the free-running period of the circadian clock among different varieties of plants. Intriguingly, these period lengths have been correlated with phenological characteristics in the plant, including agriculturally relevant ones such as X and Y. However, under most natural conditions, plants are *always* exposed to entrainment conditions, such as light and temperature cycles. It is unclear why the properties of the circadian clock in the absence of entrainment have any bearing on the phenotype of a plant that is entrained daily.

One hypothesis that has been proposed is that the properties of the free-running circadian clock will impact how plants perceive and respond to changes in photoperiods. Current evidence for this is indirect: for instance, we can observe that the period of the circadian clock under free-running conditions is correlated to phenotypes that are season-dependent, such as flowering time (IS THIS TRUE? CITATION).

One of the challenges in detecting how plants respond to photoperiod transitions is that the changes in gene expression when photoperiod is adjusted are transient 'blips' (see Fig 1A) that are difficult to characterise using the language that many circadian biologists employ (period, amplitude, peak, trough). Instead, we plan to characterise these 'blips' in gene expression dynamics through the lens of Functional Data Analysis (FDA). FDA is a field of statistics that deals with data sampled from curves, and it includes many techniques for directly analysing the properties of curves without having to extract specific pre-defined 'features' such as period and amplitude.

In this proposal, we aim to identify specific quantitative trait loci (QTLs) that are associated with characteristics of the circadian clock, response to photoperiod shifts, and seasonality phenotypes. If we can identify specific genes that are involved in all three of these processes, then we will be able to identify a

clear mechanistic path that explains the genetic basis for variability in the circadian clock and its relation to seasonality-related phenotypes.

Moreover, the identified genes will be good targets for crop breeding. Already, we have identified a circadian clock variant that has been used as a target for barley breeding. If we can explain *why* a specific clock-related genetic variant impacts agriculturally-relevant phenotypes, then that variant would be more likely to produce expected outcomes in field trials than a variant that is identified with no clear mechanistic pathway.

Figure 1: (A) An example of a 'blip' in gene expression, following a change in photoperiod. (B) An example of two gene expression curves that have the same amplitude and period, but that have different shapes. This information would be lost using standard feature extraction software used by circadian biologists. (C) As a proof of principle, we have developed a time-series QTL analysis method, in which we initially fit the experimental data to a smooth curve using a Fourier basis, then extract the slopes at discrete time points, for each strain in the RIL population. At each discrete time point, we find QTLs that are associated with the slopes using interval mapping. (D) In this proposal, we will develop a more statistically rigorous and scalable method. In particular, we will calculate the slope over time for each strain (by taking the derivative of the expression curves). Then for each genetic locus, we will calculate the LOD-scores as functions over time. (E) Data analysis workflow. (F) The experimental set-up: three photoperiod patterns will be considered, and plants from the RIL population will be exposed to each pairwise combination of photoperiod shifts. (G) Two approaches for evaluating the hypothesis that the circadian clock is associated with seasonality phenotypes because the circadian clock impacts the ability of a plant to respond to photoperiod transitions. (H) We will look for QTLs associated with the circadian clock, response to photoperiod transitions, and seasonality phenotypes.

Work plan

Experimental system

Overview: This analysis will be performed on an existing set of recombinant-inbred lines (RILs) between Arabidopsis ecotypes that are adapted to living at different latitudes. Each of these lines also contains a luciferase tagged CCR2 gene, which we have already used to track the circadian clock over time. As part of the proposed project, we will also need to collect additional expression data of CCR2 over time under photoperiod transitions. In addition, we will need to collect phenotypic data about each of these lines, under different photoperiods.

Preliminary results: Development of RIL lines

Preliminary results: Measuring CCR2 expression over time in continuous light conditions after X entrainment

Measuring CCR2 expression under photoperiod transitions

Measuring phenotype

Statistical method development

Overview: A main aim of the project is to find associations between QTLs and gene expression patterns over time. It is possible to use existing statistical techniques to find associations between QTLs and specific *features* of expression over time, such as period length and amplitude. However, the expression dynamics are much richer than this– if we were to focus exclusively on period and amplitude, we would be throwing away a lot of data about the shape of the entire gene expression profile (see Fig 1B). Instead, we plan to develop new statistical methodology to identify QTLs that are associated with the entire pattern of CCR2 expression over time. If the method is successful, we would be able to input into the model a set of QTLs and the model would be able to generate a gene expression profile over time that would match as closely as possible with the expression profile measured by an experiment. This will be developed into an R package that we envision will become the primary method that will be used to associate QTLs to time series data, whether related to agricultural or medically-relevant time series phenotypes.

Preliminary results: expressing circadian clock curves in terms of a Fourier basis

Preliminary results: A preliminary strategy for performing QTL analysis for time series data

See Figure 1C

Statistical approach 1: A more rigorous approach to finding QTLs associated with local curve features (such as rate of change of gene expression)

In the previous section, we have demonstrated that it is possible to identify QTLs that are associated with a local feature of a gene expression curve– the slope (i.e. change in gene expression over time). In this preliminary approach, a separate QTL analysis was performed at many different discrete time points. However, because we perform a large number of separate QTL analysis, we need to correct our p-values to account for multiple testing. This means that the more dense set of time points in which we run the QTL analysis, the less likely we would be to identify QTLs that are significantly associated with the slope. This indicates that the simple approach does not scale well; however, a more statistically rigorous strategy would overcome this obstacle.

In particular, it is important to remember that the slope at nearby time points are highly correlated to one another, because the gene expression profiles of CCR2 are continuous and smooth. Because of this, we would expect that QTLs that are associated with the slope at one time point should also be associated with the slope at adjacent time points. In other words, each QTLs' logarithm of the odds ratio (LOD score) should be a smooth and continuous function over time. Additionally, the threshold at which the LOD score is deemed significant should also be a smooth and continuous function over time.

Instead of performing a separate QTL analysis at different time points, we will perform all these QTL analyses simultaneously, modelling the QTL LOD scores and the LOD significance thresholds as smooth and continuous functions. This strategy will decrease the false-positive rate of QTL identification and will also overcome the scaling issues with the preliminary method we developed.

An output of this analysis will be a set of QTLs that are associated with the shape of the circadian clock curve, as well as an indication of what time of day each QTL is most important for predicting the slope of the gene expression curve.

Statistical approach 2: Finding QTLs associated with global curve features

The previous proposed method will identify QTLs that are associated with local features of a curve (such as the slope). However, some QTLs might be associated with global patterns in gene expression, such as phase variations. While circadian researchers do use tools like X and Y to compare curves and find phase and amplitude variation, these methods assume that the gene expression pattern is stationary. In other words,

they assume that the amplitude or phase shift is uniform across the entire time series. We have previously shown that this assumption is not true (show example).

Instead, for each curve, we will express the amplitude and phase variations as functions over time (see Tucker paper, SRSF registration followed by fPCA). Then, we will identify QTLs that are associated with these curves, using a technique called functional regression. However, because adjacent QTLs are highly correlated to one another, we won't be able to use existing functional regression methods out-of-the-box, but rather we would need to adapt them to be more like the interval mapping approach used in traditional QTL analysis.

Data analysis plan (see Fig 1E)

Phase 1: Associations between circadian clock parameters and seasonality-related phenotypes in RIL populations

For each RIL, we will have CCR2 expression curves (i) under continuous light, entrained using either photoperiod X or photoperiod Y (ii) under photoperiod X –> photoperiod Y (iii) under photoperiod Y –> photoperiod X (see Fig 1F). We will analyse the association between these curves, using functional regression. Functional regression is just like normal linear regression, except that instead of finding linear associations between two sets of numbers, functional regression finds linear associations between two sets of functions (function-to-function regression) or between a set of numbers and a set of functions (scalar-to-function or function-to-scalar regression). In this case, since we have multiple time series measurements, we will employ function-to-function regression, and this will help us find what features of the circadian clock gene expression curves are most informative for predicting how a plant will respond to photoperiod transitions.

Phase 2: Is the association between the circadian clock and seasonality-related phenotypes mediated by a plant's ability to respond to photoperiod transitions?

First, we will need to determine how much of the differences in seasonality-linked phenotypes in the RIL population can be explained by differences in the ability of the plant's to respond to photoperiod transitions. To do this, we will perform a function-to-scalar regression, and we can also try alternative regression strategies that ignore the time-series nature of photoperiod transition data, such as random forest regression and support vector regression.

In parallel, we will build regression models that predict the seasonality-linked phenotypes in the RIL population, based exclusively on the free-running circadian CCR2 expression curves.

Then, we will need to quantify how "redundant" these models are: If both models (circadian clock and photoperiod transition) provide the same information about the phenotypes, then this lends credence to the hypothesis that the free-running circadian clock is associated with seasonality phenotypes because it impacts how plants respond to photoperiod transitions.

We will test this in two ways (see Figure 1G). First, we will build a model that predicts the phenotypes using CCR2 expression from *both* the circadian clock and photoperiod transition curves. Then, we will evaluate whether this combined model performs significantly better than each individual model. If the combined model is not significantly better, then these models are redundant.

In the second approach, we can use the model from the previous section to predict CCR2 gene expression curves after photoperiod shifts, using only the circadian clock CCR2 gene expression curves. Then, we can input these 'predicted' curves into the model that we previously trained to predict phenotypes from CCR2 expression curves under photoperiod transitions. If the subsequent predictions are accurate, then this would

suggest that the correlation between the circadian clock and seasonality phenotypes are mediated by the ability of the plant to respond to photoperiod transitions.

Alternatively, we may discover that CCR2 expression under free-running conditions and under photoperiod transitions provide significantly different contributions to seasonality phenotypes. If so, this suggests that there may be an alternative mechanism that contributes to the association between the circadian clock and phenotypes.

Phase 3: Detection of QTLs associated with circadian clock parameters, photoperiod transitions, and seasonality-related phenotypes

Finally, we will apply the new statistical methodology to find QTLs associated with CCR2 expression under the free-running circadian clock and photoperiod transitions. We will also use standard interval mapping approaches to associate QTLs directly with seasonality phenotypes. The result of this analysis will be three distinct sets of QTLs, associated with the circadian clock, response to photoperiod transition and seasonality phenotypes. We can compare these QTL lists and use them to suggest specific potential mechanisms (see Figure 1H).

Multi-scale modelling: linking genetic variation to circadian clock re-wiring to photoperiod adaptation to seasonality phenotypes

I know that there are some well-developed ODE models of the circadian clock in Arabidopsis. Are there equivalent ODEs for photoperiod adaptation? Are the links between these models known? (I guess via FT, GI, PIFs, etc, but I'm not sure what actual models have been developed). A nice cherry-on-top of the grant might be to use the QTLs to adapt the model of the circadian clock/photoperiod adaptation to reflect our observations. I don't know enough about what has already been done to know if this is feasible.