Functional Data Analysis and QTL Detection Across Time within the Circadian Clock

Sarah Cloud Lauren Lock

MSc by Research

University of York
Biology

September, 2017

Abstract

Circadian data are typically analysed using the Fast Fourier Transform Non-Linear Least Squared method. The purpose of my project was to develop a new approach and to apply this to the analysis of rhythms for Quantitative Trait Locus (QTL) mapping over a time domain. This was achieved by examining the rhythms as functional data. The data used for this process came from what is known as the W9W recombinant inbred line population of *Arabidopsis thaliana* plant. Contrasting conditions of light and dark were also examined. The data consisted of circadian rhythm measurements that were expressed through the *CCR2* gene measured by luciferase imaging of living plants. In order to facilitate exploratory analysis of the way in which the curves behaved, they were smoothed and fitted using basis functions. Through this process the circadian rhythms were transformed from discrete observations to clearly defined functions. From there, derivatives were taken and both velocity and acceleration were examined. This led to the identification of changes in length of period between wild-type and mutant plants, as well as allowing direct comparisons of the curves behaviour between differing light and dark conditions. Having examined the correlation functions of the population free-running in darkness, further exploration was carried out by conducting principal-component analysis. This enabled the main types of variability in the wild type and mutant to be identified and analysed. QTL mapping analysis was then performed in the time domain measuring velocity with plants free-running in darkness. Viewing these data as functions allowed for a thorough and detailed exploration. This led to the discovery of new genetic information that would have otherwise have been overlooked. Overall, the techniques used in this project have opened the door to further study and investigation of the quantitative basis of circadian rhythms.

Acknowledgements

I would like to express my deepest gratitude to Professor Seth Davis and Dr Marina Knight: Professor Seth Davis, for accepting me into his lab and enabling me to complete my research. Thank you so much for this opportunity, for teaching me so much along the way and for all your guidance and support throughout this time. Dr Marina Knight, for setting me on my academic research path and for providing me with support and encouragement throughout this Masters project.

A huge thanks goes to all members of the Davis Lab Group for helping me learn and adjust to biology and for answering all my questions with patience and humour, even the silly ones!

A very special thank you goes to Amada Davis, not only for making me laugh but also for being so patient and for teaching me so much.

Lastly, thanks to my family, in particular my mother who has been invaluable in my development this year and for always believing in me.

Author's declaration

I declare that this thesis is a presentation of original work and I am the sole author. This work has not previously been presented for an award at this, or any other, University. All sources are acknowledged as References.

Contents

List of Figures

List of Tables

1. Introduction

1.1. **Introduction.** The Earth's rotation, on its axis and its orbit around the sun, results in environmental changes in light and temperature across time [96]. In order to anticipate these environmental cues, organisms have developed a timing mechanism termed the circadian clock, which is responsible for the organisms oscillating rhythms [81]. These can be defined as a collection of biological rhythms that occur during a time period that is approximately a 24-hour cycle. In general, to be classed as circadian, a biological rhythm must be maintained in constant conditions and have a free-running period of approximately 24 hours. It must also be possible for the rhythm to reset following exposure to external stimuli, known as entrainment. Lastly, throughout a range of temperatures, the rhythm must maintain its periodicity [36]. Circadian rhythms are apparent in a range of different organisms, such as mammals, birds, plants and algae [126]. A large body of work has been carried out over many years in order to enhance our understanding of how circadian rhythms work and interact within an environment [10].

1.2. **History of plant circadian rhythms.** The first recorded observations of a circadian process was in a plant. This appeared in the fourth century B.C, where the diurnal leaf movements of the tamarind tree was described by Androsthenes[14]. However, the fact that that the origin of these rhythms are endogenous was not understood at that time. The first scientific literature on this subject did not emerge until almost 21 centuries later when, in 1729, Jean-Jacques d'Ortous de Mairan, a French scientist, conducted tests on the sensitive heliotrope plant to ascertain whether a straightforward reaction to sunlight, or something else, was responsible for the daily opening and closing of leaves [53]. He concluded that the movement of leaves was not dependent on the daily light-dark cycle and hypothesised they were controlled by an internal mechanism. In 1930, the German scientist Bünning provided evidence that plants movements were endogenous. He identified two variants within the species *Phaseolus coccineus* where their endogenous period differed, exhibiting cycles of 23 and 26 hours. When crossed, the period lengths of the progeny ranged between the extremes of period lengths of the two parents, this suggests that this particular aspect

of circadian rhythms is a genetically based polygenic trait [18]. This provided strong evidence that circadian rhythmicity derives from an interaction between light and a circadian pacemaker, not from external stimuli alone [117]. Since these first observations, the study of the plant circadian clock has developed and it is now known that many genes covering aspects of the plants development, flowering and growth, as well as its response to environmental stresses or harmful pathogens are controlled by the circadian clock [81]. The analysis of circadian rhythms as been developed over many centuries with plants dominating this field of chronobiology [13]. Whereby chronobiology is defined as the field of biology that studies timing processes in living organisms [36].

1.3. **The plant circadian clock.** A simplified way to visualise the circadian system is as three components. The central component is a self-sustained central oscillator that generates rhythmicity. The other two components are input pathways that enable the clock to be entrained to local environmental day-night cycles, and output pathways which are regulated by the core oscillator [88], [49]. However, in reality the clock is far more complex than this, as multiple levels of feedback occur between each level [87].

A common approach used to explore the circadian clock is monitoring the rhythmic outputs which are controlled by the core oscillator [49]. These have been made relatively accessible by using a luciferase-reporter gene [86]. Luciferase is an enzyme derived from fireflies that catalyses the biochemical reaction that causes fireflies to "glow" [125]. If transgenic plants containing a luciferase transcriptional reporter fused to a circadian regulated promoter are supplied with the substrate luciferin, then the plants emit light when the promoter gene is being expressed [85]. This procedure allows real-time gene expression to be monitored. This has revolutionised the study of plant circadian biology and directly led to many discoveries concerning the plant circadian clock [4],[31],[35],[49]. To demonstrate this visually, Figure 1 shows a plot of a Ws wild-type plant under constant darkness (DD) conditions. Here time in Days (d) is plotted against luminescence measured as counts per second (cps), each point represents the luminescence reading measured at a particular time point.

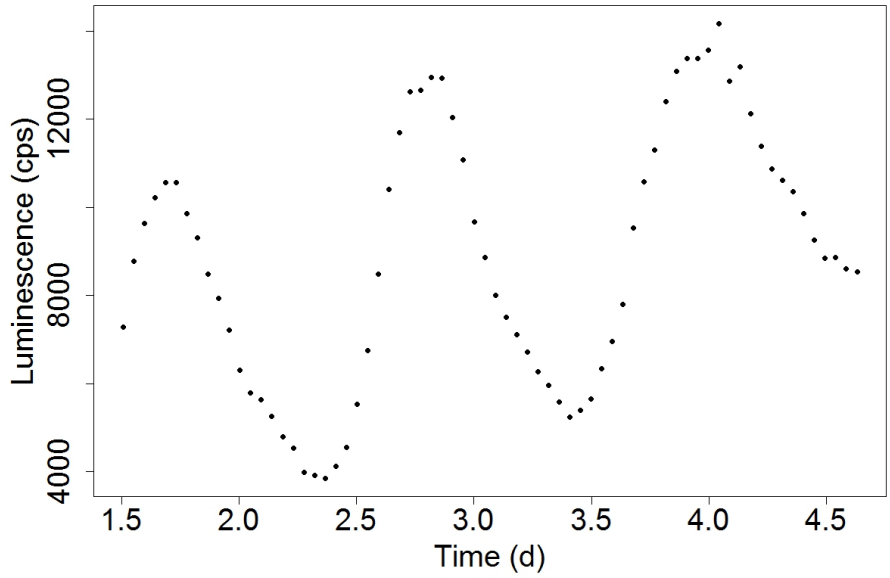


FIGURE 1. The observed luminescence values measured over time of a Ws wild-type plant free-running in DD conditions.

A recent circadian mathematical model published in 2016 [94] proposed the cartoon shown in Figure 2. The core circadian oscillator is divided into three interconnected areas of transcription-translation feedback loops, the core oscillator central loop and a morning and evening loop, as oscillator genes are expressed at different times of the day [56]. As shown in Figure 2 at dawn, CIRCADIAN CLOCK ASSOCIATED 1 (CCA1) and LATE ELONGATED HYPOCOTLY (LHY) repress the expression of the *PSEUDO-RESPONSE REGULATOR (PRR)*-encoding genes, *TIMING OF CAB EXPRESSION (TOC1)*, *GIGANTEA (GI)* and the evening complex (EC) members LUX ARRHYTHMO (LUX), EARLY FLOWERING 3 (ELF3) and EARLY FLOWERING 4 (ELF4). *PRR9*, *PRR7*, *PRR5* and *TOC1* are sequentially expressed and repress the expression of *CCA1* and *LHY*, as well as their own transcription. In the evening, *TOC1* represses all of the previously expressed components in addition to *GI*, *LUX* and *ELF4*. Subsequently, the EC maintains the repression of *GI* and represses *PRR9* and *PRR7*. In reality it is much more complex than this with several feedback loops of gene expression that interact [118].

There are slave oscillators present in the clock and these act as an intermediary between the core oscillators and output pathways [114]. These have two important properties, firstly it holds a negative-feedback loop. These regulate their own

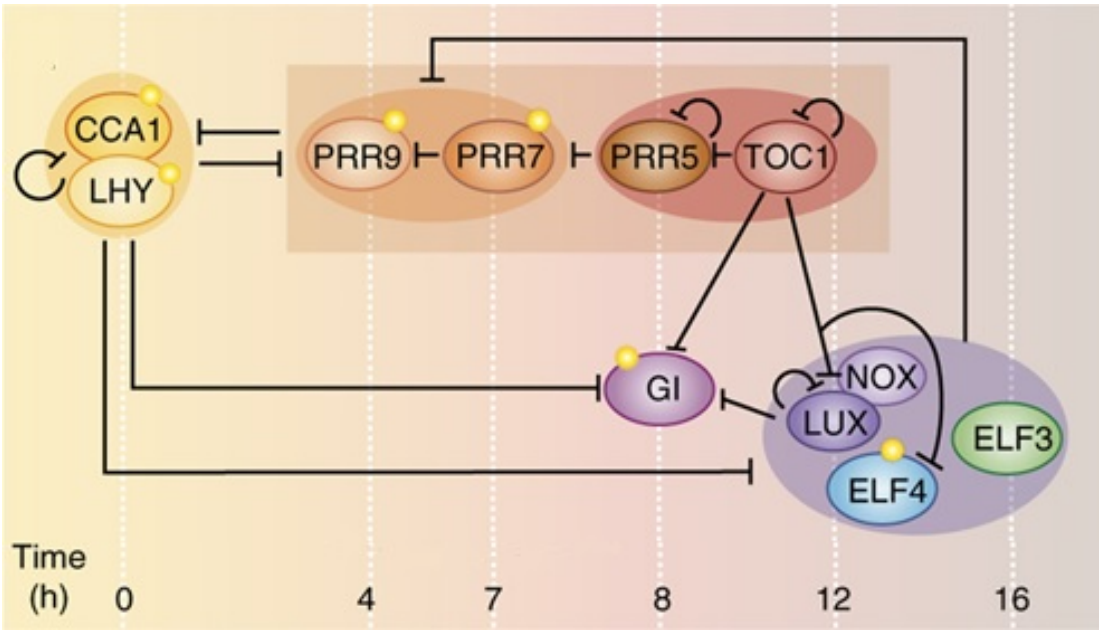


FIGURE 2. Adapted from [94]: The sequential expression of each component throughout the day is shown from left to right and the time of activity is expressed in hours after dawn. The yellow and grey areas represent day and night, respectively. Black bars indicate repression. Ovals represent functional groups. The sun icon depicts light promotion of transcription [94].

expression when constitutively expressed by repressing accumulation of its own transcript. Secondly, it is situated downstream of the core oscillator, such that its rhythm is highly dependent on the core oscillator [114]. An example of a slave oscillator gene in *Arabidopsis thaliana* (Arabidopsis) is *COLD AND CIRCADIAN REGULATED 2 (CCR2)* also referred to as *Arabidopsis thaliana GLYCINE-RICH Protein 7* (AtGRP7). *CCR2* encodes a glycine-rich RNA binding protein whose transcriptional expression is cold regulated with peak expression occurring 8-12h after dawn (introduction of light) [52]. Under free-running conditions for both constant light (LL) and constant darkness (DD), *CCR2* produces robust rhythms, which has made it an ideal choice in clock function detection [122].

1.4. ***Arabidopsis thaliana.*** The small flowering plant Arabidopsis is a model system used in plant circadian biology for identifying genes and determining their functions [83]. It is ideal for use in research as it has a rapid life cycle, is small

in size and has the ability to self-fertilise [123]. In addition, the complete 125-megabase genome sequence is known, and it currently contains 27,416 known genes making it extremely well characterised and therefore a suitable model in plant studies [74], [60].

1.5. **Analysis of plant circadian rhythms.** Over time a number of different techniques have been developed and although these began with simple observations, time series analysis has now become a widely used method [89]. The procedure of choice varies depending on experimental aims, sampling design and thus the type of data that has been collected [111]. Nevertheless, an essential first step in the analysis of any time series consists of visual inspection of a time plot. Visualising data in this way guides the selection of the most suitable procedure to use for further analysis [8].

Similar to other rhythmic processes, circadian rhythms are characterised by certain parameters. These commonly include period, amplitude, phase, waveform and robustness or prominence [111]. One method used to analyse these parameters is cosinor analysis whereby, under the assumption that the parameters of the fitted function reflect the true parameters of the biological rhythm, a cosine function is fitted usings a least squares method to the measured data points [93].

Within the plant circadian community the current standard is the Fourier Transform Non-Linear Least Squares (FFT-NLLS) method [121], the primary aim of this method is the analysis of circadian data obtained from free-running conditions without entrainment. The purpose of this method was initially for use in genetic screenings in order to facilitate identification of mutant organisms that had a altered period, and was first developed by the NSF Centre for Biological Timing in Virginia [136]. The FFT-NLLS method is suitable under the assumption that rhythmic data are comprised of trigonometric functions. As this model has unconstrained periods and a large number of components, it is possible to represent almost any curve. Most parameters are calculated through descriptive statistics and a primary step is to obtain an accurate estimate of the underlying period. There are numerous approaches that can be taken (Fourier analysis,

Enright periodogram, Lomb-Scargle periodogram [111]), all with their own assumptions and differing levels of complexity. As an example a sum of cosine functions can model these data and this is shown in Equation 1 below [136]:

$$f(t) = c + \sum_{i=1}^{N} \alpha_i \cos\left[\frac{2\pi\,(t+\phi_i)}{\tau_i}\right], \tag{1}$$

where τ_i, ϕ_i, α_i are period, phase and amplitude of each cosine component, c is the offset.

The widely used FFT-NLLS method centres on the assumption that circadian rhythms are stationary, *i.e.* that parameters, such as period and amplitude are stable and non-changing across time [46]. However, the literature on circadian rhythms gives very little consideration to the question of stationarity. It is simply taken for granted in the time scale of days/weeks. There has been very little investigation into the effect that overlooking this may have on the representation of circadian data [112].

There are alternative techniques for managing the non-stationarity of circadian rhythms. One such technique is to view data as functional, where even though data consists of discrete-measurements, using this method, their values can reflect a smooth variation that describes behaviour [30],[110].

1.6. **Functional Data.** There are many forms of functional data, however, a defining feature throughout is that the functions they consist of are frequently, though not always, smooth curves [110]. A simple example of this is to consider weather temperature recorded monthly throughout the year. Each time the temperature is recorded, it is done as a discrete value. This data when viewed together, can then show the variation in temperatures during the year as a smooth functional curve [109].

Ramsay and Dalzell (1991) initially coined the term functional data analysis (FDA), though, a number of similar practices and ideas had been in use long before this time [104]. FDA is a branch of statistical analysis that does not have a firm definition, it is continuously shaped by new developments and it has grown exponentially [105]. Nevertheless, in common with many other branches of statistics, there are some common aims to consider when implementing FDA techniques [110], including:

- to represent data in different ways in order to aid further analysis;
- to display data in order to highlight various characteristics;
- to study important sources of pattern and variation among data;
- to use input or independent variable information to explain variation in an outcome or dependent variable.

Established statistical techniques used for analysing data are extended through the use of FDA. By viewing a dataset as a smooth function rather than as as discrete points, enables inferences to be made from the data [104], [105], [110]. This allows many things to be examined that, with more traditional analysis techniques, may be overlooked or missed. It is a fundamental aim to be able to extract and explore as much information as possible from a data set.

1.7. **Quantitative trait loci (QTL) analysis.** A quantitative trait locus (QTL) is defined as a region within a genome that contain genes associated with a particular quantitative trait [80]. A genetic map is created using the genetic markers positioned in close proximity to genes and all genetic markers occupy a known location on a chromosome called loci (singular locus) [25]. Linking this information with displayed phenotypes of a population allows the relationship between phenotypic and genotypic measurements to be analysed, and is commonly known as QTL mapping [79]. The principle of QTL mapping is that through detecting an association between the phenotype and genotype of markers, the genes responsible for the natural phenotypic variation observed within a population can be explored, so providing insight into their effects and interactions [80], [39].

1.8. **QTL mapping methods.** QTL mapping has been well used in plant circadian data, though this is commonly performed on static traits such as hypocotyl elongation, or averaged traits such as period [101],[65]. Estimates for period are often performed using Biological Rhythms Analysis Software System (BRASS), which uses the FFT-NLLS curve estimation method [13], [6]. For example BRASS has been used to address the circadian parameter of phase. Here phase reflects the entrained relationship between the clock and the external cycle. Using BRASS, luminescent rhythms of individual seedlings were plotted as a moving average and the time of the first peak of each seedling rhythm was recorded. Using this

procedure the positions for phase were successfully QTL mapped using interval mapping (IM) and QTLs that have major effects on circadian phase were found [32]. In contrast to the traditional FFT-NLLS method used by BRASS, by viewing circadian data as functional, a different approach for estimating curves can be taken. This creates an opportunity to explore the different outputs from QTL mapping, both in the traits measured and in the static nature of these traits. In order to do this, it is important to consider the first the basic principle of how QTL mapping works.

1.8.1. ***Single marker QTL analysis***. A traditional method for detecting a QTL with mapping data is to consider each marker individually [62]. At any particular marker, all individuals can be categorised according to their genotypes and the phenotypic means for each group can be compared. Statistical tests can then be used to determine whether a QTL is present, a significant value indicates that a QTL is located in the vicinity of the marker [103]. This method does not require a complete linkage map and a disadvantage of this is that the further a QTL is from the marker, the less likely it will be detected statistically as recombination may occur between the marker and the QTL [25].

1.8.2. ***Interval mapping***. Interval mapping (IM) is the most common method of QTL analysis and is particularly useful as it takes account of missing genotype data. The principle is to test for the presence of a QTL at many positions between the markers, making use of a linkage map [62]. The use of linked markers in the analysis compensates for the recombination possibilities between the markers and the QTL. There are several different interval mapping methods, each with differences in the way missing data are treated [17], [21]. Standard interval mapping is frequently used and via the expectation maximisation (EM) algorithm, has the maximum likelihood estimation. Other methods include the Haley-Knott regression method [44].

Given that y_i denotes the phenotype and g_i denotes the QTL genotypes for any given individual i, we assume that $y_i \mid g_i \sim N(\mu_{g_i}, \sigma^2)$. An individuals i probability density function can be given by $\Sigma_j p_{ij} \phi(y_i; \mu_j, \sigma^2)$ where p_{ij} denotes the mixing proportions and are derived using known marker data. The density of the normal distribution (See Appendix 3) is given by ϕ and the sum is over the

possible QTL genotypes (denoted j). These are calculated for each individual using $p_{ij} = Pr(g_i = j \mid \mathbf{M}_i)$, where $\mathbf{M}_i$ is the multipoint marker genotype for individual i (for these calculations R package `"R/qtl"` was used) [15].

μ_j and σ were estimated by maximum likelihood; that is, we take those values for which the observed data is most probable as our estimates. The likelihood function is:

$$L(\boldsymbol{\mu}, \sigma) = \Pi_i \Sigma_j p_{ij} \phi\left(y_i; \mu_j, \sigma^2\right)$$

Where the sum is over the possible QTL genotypes [64].

The maximum likelihood estimates (MLEs) cannot be obtained without using an iterative algorithm. This is where the EM algorithm is used. The EM algorithm estimates the parameters of a model iteratively, beginning at a chosen starting point. Each iteration consists of an Expectation (E) step and a Maximisation (M) step. It can be shown that each such iteration either improves the true likelihood, or leaves it unchanged [82]. The E step finds the distribution for the unobserved variables as well as the current estimates of the parameters, while the M step re-estimates the parameters under the assumption that the distribution found in the E step is correct in identifying those with maximum likelihood [92].

Once the maximum likelihood estimates of the μ_j and σ have been obtained, a test statistic know as a logarithm of odds ratio (LOD) is used to detect QTL presence [95]. A LOD score is a transformation of the likelihood ratio statistic and provides a ratio of the likelihood that there is a QTL present, and the likelihood under the null hypothesis that is there is no QTL anywhere in the genome [132]. A LOD score is calculated as follows [17];

$$LOD = \log_{10}\left(\frac{\Pi_i \Sigma_j p_{ij} \phi\left(y_i; \hat{\mu_j}, \hat{\sigma^2}\right)}{\Pi_i \phi\left(y_i; \hat{\mu_0}, \hat{\sigma^2}\right)}\right)$$

Where $\hat{\mu_0}$ and $\hat{\sigma_0}$ are the average and standard deviation of y_i.

1.8.3. ***Determining a threshold value.*** A LOD score indicates the statistical presence of a QTL, the higher the LOD score the more strongly the presence of the QTL is indicated [17]. QTL mapping methods rely on identifying an appropriate threshold at which a LOD score becomes significant. The test statistics are estimated from the data set and are compared in order to determine whether

a significant QTL exists. The distribution of LOD scores under the null hypothesis depends on a number of factors, including the number of typed markers, size of the genome, the number of individuals and the phenotype distribution. [100]. There are many ways to approach the problem of estimating a significance threshold, one is to use permutation tests, whereby the phenotypes relative to the genotype data are "reshuffled" [103]. By randomly reshuffling the data, and so which phenotype corresponds to which genotypes, the original marker QTL association is eliminated. Permutation tests then generate a new data set under the null hypothesis. The expectation of the null is that there is no association between phenotype and genotype. This process is repeated a number of times resulting in an empirical distribution of the test statistic LOD scores under the null hypothesis, whereby the $95th$ percentile is used as an estimated LOD threshold value [23].

The advantages of using the permutation test approach to estimate the LOD score threshold are its simplicity, its distribution-free nature and its ease of use as it is not dependent on population structure. Although the computational workload is extensive, with the standard minimum number of permutations being 1000 [95], this is made relatively easy with the help of programming languages such as R [132].

2. Generation of data

2.1. **The plant population.** A recombinant inbred line (RIL) is formed by crossing two inbred strains followed by repeated selfing, in this case to create a new inbred line whose genome is a mixture of the parental genomes [16]. A visual example of a population derived in this manner is shown in Figure 3. The Davis lab group generated the RIL population that was used in this project [34]. It was created from two different parent ecotypes (Columbia (Col-0) *hsp90.2-3* mutant and Wassilewskija (Ws-2) *CCR2:LUC* wild-type [26]) and the resulting BC1F7 generation used is referred to as the W9W RIL population. The Ws-2

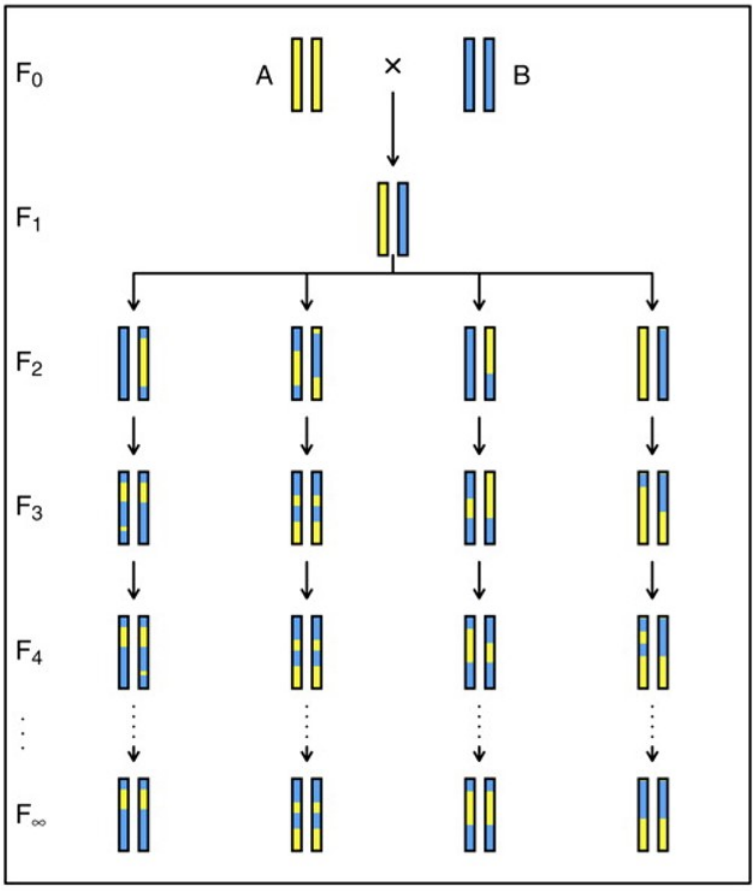


Figure 3. A RIL formed by crossing two inbred strains and repeated selfing.

here on termed 'Ws' line was created by inserting a Luciferase fused to a reporter gene under the control of the promoter for cold and circadian regulated (*CCR*) genes. The Col-0 strain contained a homozygous mutation at *hsp90.2-3* that causes a mutation in the cytosolic heat shock protein. These two parent strains

were crossed and then this F1 line was grown and it recurrently backcrossed to the Ws parent. This BC1 line was then used for selfing six times to create a F7 population. From this F7 population 48 homozygous mutant (*hsp90.2-3*) and corresponding 48 wild-type (without the mutation at *hsp90.2-3*) strains were selected. Hsps are a family of proteins widely distributed throughout animals, plants and fungi. Plants response to stress and disease and their development, are linked to the Hsp90 chaperone proteins [133]. Identification of a number of *HSP90* genes has shown that changes in salinity, temperature and metals all produce a strong effect [133]. Arabidopsis contains seven identified Hsp90 proteins (Hsp90-1 to Hsp90-7) [71]. The W9W mutant plants contain a mutation of the Hsp90.2 protein known as *hsp90.2-3* which is associated with an increase in stochastic variation [113].

The W9W RIL contains 75% Ws and 25% of Col-0 background in its genome. In a previous project, these 96 strains were genotyped, the physical mapping of the chromosomes of theses 96 genotypes was performed using 102 SSLP markers, and a genetic map of the resulting population was generated [34].

Each of the plants in the population (both the wild and mutant strains) are homozygous for the *CCR2:LUC* construct. Therefore, in the presence of luciferin the plants will luminesce allowing for gene expression of the *CCR2* and thus allow the core oscillator of the circadian clock to be measured. Luciferase expression was measured using a Topcount NXT imaging system (Perkin Elmer) where the average luminescence reading of each well is measured over 5 seconds approximately every hour, giving units of counts per second (cps) [47].

2.2. **Materials and methods.**

2.2.1. ***Plant growth media.***

Muraskige and Skoog Basal Salt (MS), containing 3% sucrose (MS3)

For 1 litre MS3;

- 4.4g Muraskige and Skoog Basal Salt
- 30g sucrose
- 0.5g MES free acid

- 15g phytoagar
- Adjust pH to 5.7 using 1mM potassium hydroxide (KOH)

2.2.2. ***Surface sterilisation of seeds.***

Reagents

- Ethanol
  - 100% Ethanol
- Bleach solution
  - 33% Klorix Bleach in 0.02% Triton X-100
- Sterile water
- Agar water
  - Sterile water with 0.1% phytoagar

2.2.3. *Surface sterilisation of seeds.* For each topcount experiment, appropriately 120 seeds were sterilised per genotype. First the seeds were rinsed with $600\mu l$ of ethanol. After the ethanol was removed, $600\mu l$ of bleach solution was added. The bleach solution was then removed and the seeds were rinsed with $700\mu l$ of sterile water. Afterwards, the seeds were suspended in $240\mu l$ of agar water, this was then placed on MS3 medium with 15μg/mL hygromycin antibiotic to select for the transgene *CCR2:LUC*.

These plates were then kept in the 4°C for 2 days for seed stratification, before being transferred to the light/dark (LD) growth cabinet.

2.3. **Topcount Set-up.**

2.3.1. *Reagents.*

- Ethanol
  - 100% Ethanol
- Luciferin solution 1mM
  - 50mM Luciferin with 0.01% Triton X-100 as 1:49 ratio

2.3.2. *Plate set-up.* To transfer seedlings to microtite plate, tweezers were washed in 100% ethanol then the excess ethanol was burnt off to sterilised. Once tweezers were cooled they were used to transfer a seedling into each well of the plate. Luciferin 1mM solution was filter sterilise using 0.45 micron filter. Using a pipette, $15\mu l$ of luciferin solution was placed into each well of the plate. Each plate was covered with a plastic plate seal and a small hole was pierced over each well.

2.4. **Growth Conditions.** Within the scope of this project three data sets are used both the DD and constant light (LL) data sets were collected previous to this project by the Davis lab group [34]. The short day (SD) data set was generated by myself;

2.4.1. ***DD data.*** Plants entrained by subjecting them to cycles of 12 hours light, followed by 12 hours dark (12L:12D) for 10 days at $\sim$ 20°C then placed under constant darkness. Each genotype had 48 plants measured. The non rhythmic or non luminescence plants were "removed" before analysis took place, a table of these removed plants can be seen in Appendix 1.

2.4.2. ***LL data.*** Plants entrained at 12L:12D for 10 days at $\sim$ 20°C then placed under constant light.

2.4.3. ***SD data.*** Plants entrained at 12L:12D for 11 days at $\sim$ 20°C then placed under 6L:18D.

2.5. **Analysis software.** All analysis was carried out using programming language R [102], specifically the "fda" package for all FDA work [106]. QTL mapping was performed also in R using the "R/qtl" package. Two input files were generated containing the quantitative data, genotypes of all of the 96 individuals, and the marker locations. IM via the EM algorithm was performed for the velocity (Section 5.1) of the population across the experiment at every minutes and for the velocity of the wild-type and mutant-type separately. All QTL analysis was performed on data collected under free-running in DD conditions. Permutation tests were used to determine the significance threshold of the LOD score for any given minute during the experiment. 1000 permutations were performed in each permutation test and the 95% significance threshold was taken.

For ease of explanation throughout this report most analysis steps, unless stated otherwise, are demonstrated on a Ws wild-type plant (seen in Figure 4) or in some cases the genotype as a whole. In some chapters comparison of the Ws wild-type and Ws *hsp90.2-3* mutant type is a useful tool to gain information. These plants will now be referred to as simply wild-type and mutant.

3. Representing Functional Data

The following explanation of methods used in the analysis is taken from Lock 2016 with new developments added [78].

The basic principal of FDA is to view observed data as outputs of functions rather than individual observations. In relation to observed data, the term functional gives reference to the structure of the data, not its actual form. As such, it is generally observed and recorded as discrete pairs (t_i, y_i) with $i = 1, \ldots, n$, where n represents the number of observations recorded and y_i is the response measured at time t_i [105].

The plots shown in Figure 4 provide a visualisation of how discrete observations can lead to a continuous function. I first plotted each observation of a wild-type plant free-running in DD conditions as its luminescence reading against time (Figure 4 left). I then simply linearly connected the observed points together (right), this clearly showed that a particular function, here called g, can be assumed to be responsible for the underlying pattern shown by the observed data. Three distinct peaks and two troughs are seen in these data showing clear rhythmic activity over the time frame. The raw continuous line data for the whole

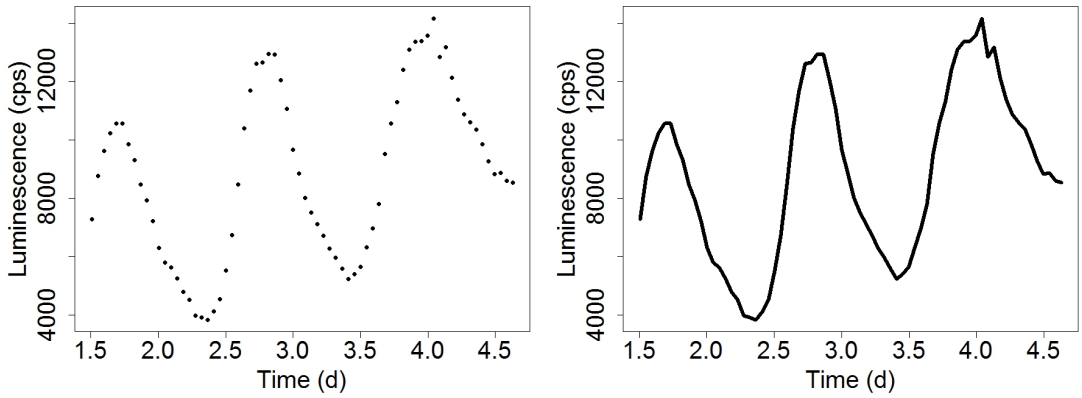


Figure 4. The observed values of a wild-type plant free-running in DD conditions. Plotted as discrete points (left) and a continuous line (right).

Ws wild-type genotype can be seen in Figure 5. This acts as a visual aid for looking at the appearance of replicates of the genotype and gives information on the structure of the group. It is shown that identical plants do not produce identical curves as demonstrated by the amount of variation present in the plot. This

emphasises the reasons for using FDA, so to investigate the modes of variation within data and being able to represent data in different ways. In order to find

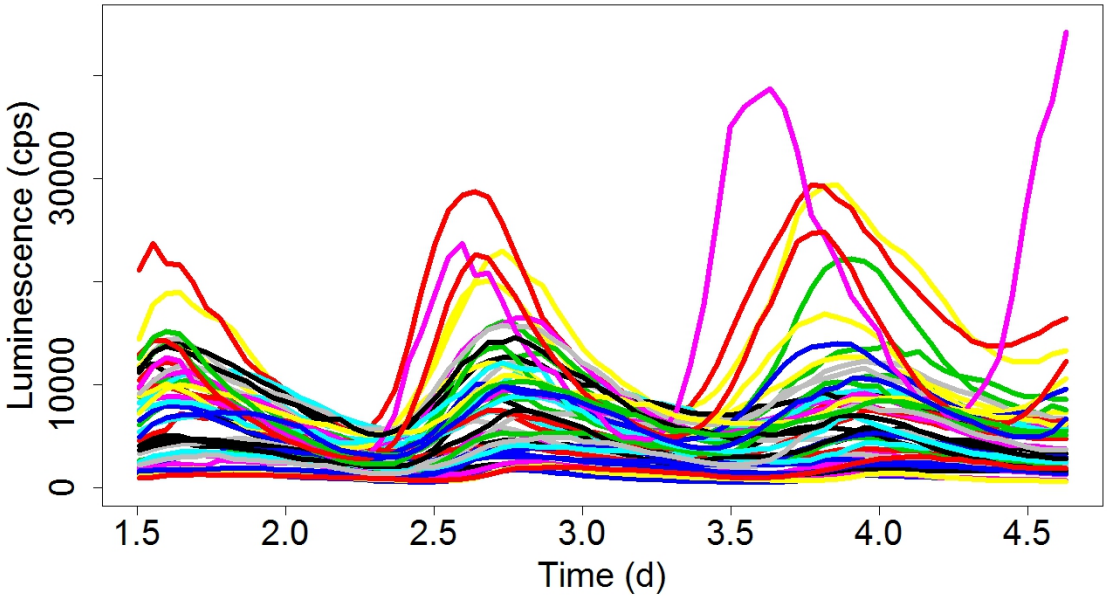


FIGURE 5. The observed values of the Ws wild-type genotype free-running in DD conditions. Plotted as continuous lines.

the correct underlying curve, measurements for every possible value of t would need to be taken. This is not practical. However, this problem is overcome by viewing the discretely measured observed data as functional, hence enabling an accurate estimate for the underlying curve to be obtained. If the observed discrete values are a reflection of an errorless function then the process of estimation would simply be interpolation. That is the creation of new data points with the set of known discrete data points [108]. If however the data has incurred some observational error that needs removing, then smoothing may be required in order to make the conversion from discrete data to functions. The term smoothing means to create an approximate function that will capture important features of the data while excluding noise, thus resulting in a smooth and so differentiable function [124]. It is clear from Figure 4 and the nature of circadian data that in order to make the curve suitable to be used in analysis, some form of smoothing does need to be applied.

Based on the principle that at a particular time point t_i, an observation y_i, is the result of an observation from a function g at a particular time point, $g(t_i)$ for

$i = 1, \ldots, n$, then it can be assumed that y is the realisation of the curve $g(t_i)$ given by;

$$y_i = g(t_i) + \epsilon_i, \tag{2.1}$$

where ϵ_i is the random error (noise), and adds a roughness to the raw data.

For ease the Equation (2.1) is presented using vector notation where, $\boldsymbol{Y}$ is the vector containing the results of the observations at a particular time point t_i, made up from the vector of observations of a function $g(t_i)$; $\boldsymbol{g(t)}$, and the vector of contributing noise $\boldsymbol{e}$. All vectors are of length n so rewriting Equation 2.1 gives:

$$\boldsymbol{Y} = \boldsymbol{g(t)} + \boldsymbol{e} \tag{2.2}$$

Commonly, when using a standard statistical model, the distributions of the random error ϵ_i would be identically and independently distributed with a mean of zero and variance σ^2. However, when looking at functional data this is too simplistic, as the variance of the residuals changes dependent upon t. The matrix in vector notation $\boldsymbol{\Sigma_e}$ will refer to the covariance structure of the errors. It shows over repeated examples that are identical that the residuals vary, apart from the error variation. Consequently Equation (2.2) will turn into:

$$\boldsymbol{Y} = \boldsymbol{g(t)} + \boldsymbol{\Sigma_e}$$

3.1. **Representation on a Basis Function.** In FDA a critical step is estimating $g(t)$. There are various approaches to this, the most common being, basis expansion and smoothing penalties. A basis function consists of a set of known basic functional building blocks and the desired function $g(t)$ will be represented by a linear combination of K groups of basis functions . A formal definition of a basis can be described as:

Definition 3.1.1. A basis for the vector space $\mathcal{V}$ is the set of vectors satisfying both of the following [120]:

- The set is linearly independent
- The set spans the vector space

In the case of FDA the vector space $\mathcal{V}$ is in fact the space $\mathcal{L}^2[\mathbb{R}]$, known as the *Hilbert space* [135] .

The Hilbert space is the accepted way for enabling the number of dimensions to become infinite whilst maintaining the geometry of an Euclidean space [45], [3]. Constructing the exact function of $g(t)$ would be a complex, and computationally expensive, process without using basis functions [120]. A linear expansion of g, represented though K known basis functions ϕ_k is given by [110]:

$$g(t) = \sum_{k=1}^{K} c_k \phi_k(t)$$

Parameters $c_1, \ldots, c_K$ represent the coefficients of the expansion. It is termed $\phi(t)$ is a basis system for g.

3.1.1. *Examples of basis functions.*

Example 1. A Fourier basis system is such a system where functions consist of sine and cosine functions with increasing frequency and is shown in the following equation:

$$\begin{aligned}
\phi_1(t) &= 1 \\
\phi_2(t) &= \frac{1}{2i}\left(e^{i\omega t} - e^{-i\omega t}\right) \\
\phi_3(t) &= \frac{1}{2}\left(e^{i\omega t} + e^{-i\omega t}\right) \\
\phi_4(t) &= \frac{1}{2i}\left(e^{i2\omega t} - e^{-i2\omega t}\right) \\
\phi_5(t) &= \frac{1}{2}\left(e^{i2\omega t} + e^{-i2\omega t}\right) \\
&\vdots
\end{aligned}$$

Here the constant ω is related to period τ by $\omega = \frac{2\pi}{\tau}$.

Example 2. The exponential basis function is another example of a basis system, comprised of:

$$\phi_1(t) = 1$$
$$\phi_2(t) = e^{-t}$$
$$\phi_3(t) = e^{-2t}$$
$$\phi_4(t) = e^{-3t}$$
$$\phi_5(t) = e^{-4t}$$
$$\vdots$$

Example 3. A further example of a basis system is the polynomial basis system, the most common is the monomial basis and is shown in the example below:

$$\phi_1(t) = 1$$
$$\phi_2(t) = t$$
$$\phi_3(t) = t^2$$
$$\phi_4(t) = t^3$$
$$\phi_5(t) = t^4$$
$$\vdots$$

Figure 6 provides a visual representation of Examples 1, 2 and 3 as well as the polynomial basis function for up to $K = 11$ basis functions.

It should be noted that, other than those shown above, there are a number of other basis systems that could be used, for example, the power basis and wavelets [22]. The basis system used will be based on which one best represents the data. Figure 7 shows an example of the way in which each of the basis systems shown in Figure 6 fits an example set of data. The example data comes from a wild-type plant free-running in DD conditions, and each basis system has been applied to the data using $K = 11$. The monomial basis (shown top right), the exponential basis (shown top left) and the polynomial basis (shown bottom left) all provide

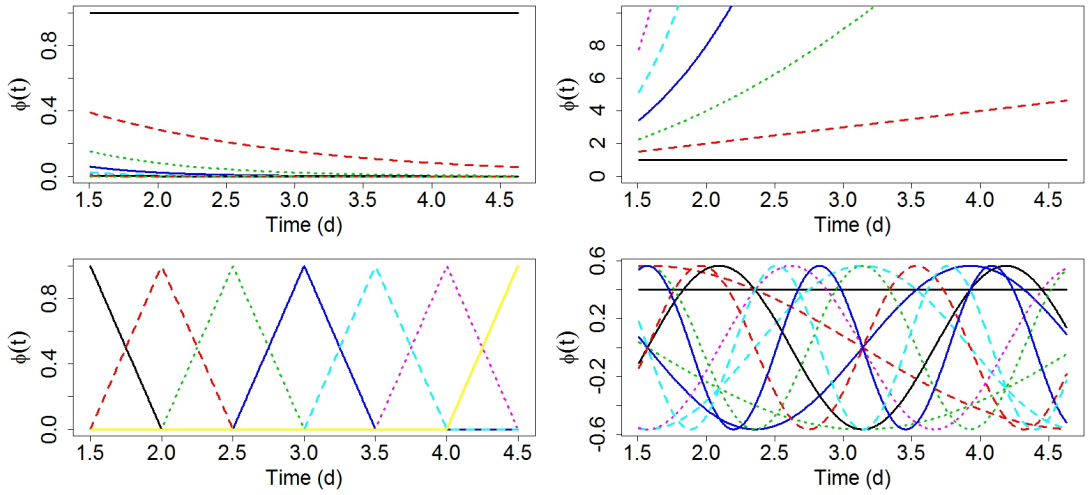


FIGURE 6. The exponential basis (top left), monomial basis (top right), polynomial basis (bottom left) and Fourier basis (bottom right) all with $K = 11$ basis functions.

quite poor representations of the data and generally some important features are missing. However, the Fourier basis system (shown bottom right) provides a generally more appropriate estimation, it gives a smooth function whilst still retaining the main features of the data at $K = 11$. Consequently, this is the most suitable basis system to use for fitting this type of data.

There are a number of advantages in using the Fourier basis system, (i) it is well established and easily accessible and was the only available alternative to a polynomial basis until the mid twentieth century; (ii), it has excellent computational properties, with equally spaced observations this is particularly apparent; (iii) it is an excellent way to describe circadian data, which is periodic, owing to its nature [54]. It is also common to use wavelets which are also ideal for use with periodic data. There are differences between these systems, the central difference is that wavelets are localised in both time and frequency whilst Fourier is confined to frequency. For these reasons, the Fourier basis system will be the focus of this project.

It should be noted that that its orthonormality properties are an important feature of the Fourier basis. Using $\langle , \rangle$ to denote the inner product space (for definition see Appendix 2) a definition follows [40].

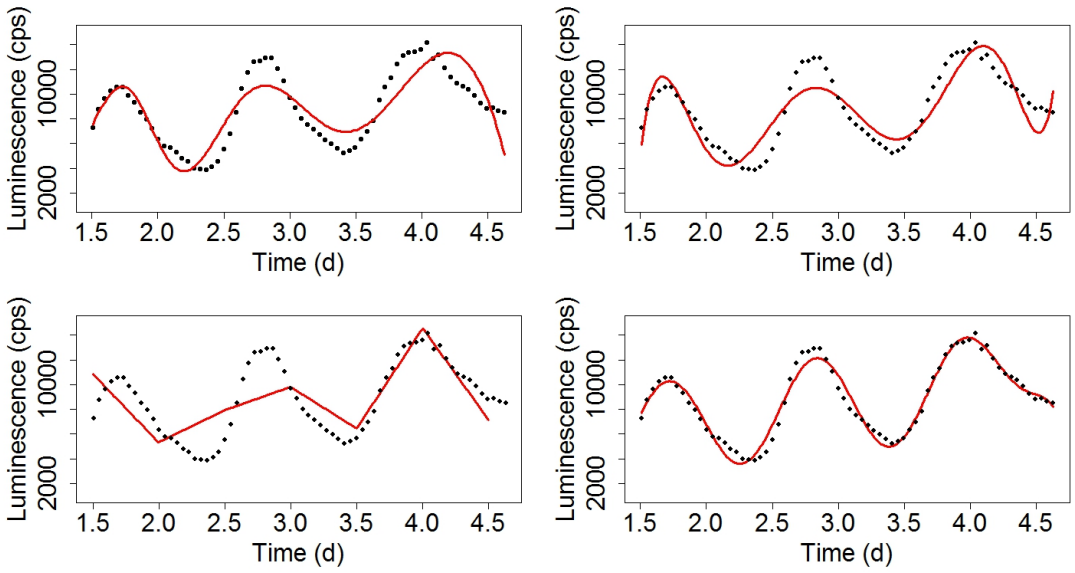


FIGURE 7. Curves estimated for a wild-type plant free-running in DD conditions using an exponential basis (top left), monomial basis (top right), polynomial basis (bottom left) and Fourier basis (bottom right) all with $K = 11$ basis functions.

Definition 3.1.2. A subset $\{v_1, \ldots, v_k\}$ of a vector space $\mathcal{V}$ is called orthonormal if both of the following are satisfied:

- $\langle v_i, v_j \rangle = 0$ when $i \neq j$
- $\langle v_i, v_i \rangle = 1$ when $i = j$

Summarised, each vector must be perpendicular and have unit length 1.

For this project, the basis functions of the Fourier basis used, form a complete orthogonal system over the vector space $[-\pi, \pi]$ [75].

Estimation of the coefficient vector. Writing in vector notation where the vector of coefficients c_k of length K is denoted $\boldsymbol{c}$ and the vector of basis functions $\phi_k(t)$ of length K is denoted $\boldsymbol{\phi}$, then $g(t)$ can be written as:

$$g(t) = \sum_{k=1}^{K} c_k \phi_k(t) = \begin{bmatrix} c_1 & c_2 & \ldots & c_k \end{bmatrix} \begin{bmatrix} \phi_1(t) \\ \phi_2(t) \\ \vdots \\ \phi_k(t) \end{bmatrix} = \boldsymbol{c}^T \boldsymbol{\phi} \tag{2.2}$$

3.2. **Smoothing Functional Data.** As mentioned previously, the functions must be smooth for analysis to take place. A well recognised method of smoothing data is to minimise the sum of the squared errors using the equation below [33]:

$$SSE = \sum_{k=1}^{K} \left(Y_k - g\left(t_k\right)\right)^2 \tag{2.3}$$

In vector notation where $\boldsymbol{Y}$ represents the vector $(Y_1, \ldots, Y_k)$ and using Equation 2.2 representation of $g\left(t\right)$, $\boldsymbol{\phi}$ is the k by K matrix containing the values of $\phi_k\left(t\right)$, Equation 2.3 in functional form is:

$$min_{\boldsymbol{c}} \left(\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\phi}\right)^T \left(\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\phi}\right) \tag{2.4}$$

This is an example of homoscedastic noise, meaning that for any given value of the independent variable, (in this instance it is t), the conditional variance of residuals in the data set is assumed to be constant [24]. Consequently, this provides a poor level of accuracy, as it is assumed that the errors $\boldsymbol{e}$ are not normally distributed with a mean of zero and variance σ^2 (see Section 3), and in reality, they change depending on t. Thus it is reasonable as an estimation technique to introduce weighting of the residuals. As a result, the observations will then be weighted proportionally to the reciprocal of the error variance for that particular observation, hence overcoming the issue of non-constant variance [84]. Defining the reciprocal of each variance, σ_i^2, as the weight, $w_i = \frac{1}{\sigma_i^2}$, then $\boldsymbol{W} = \boldsymbol{\Sigma_e^{-1}}$ is the symmetric matrix containing these weights. When it is not possible to estimate the complete variance-covariance matrix, for ease, as shown below, it can be assumed that the covariances among the errors are zero:

$$\begin{pmatrix} w_1 & 0 & \cdots & 0 \\ 0 & w_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_n \end{pmatrix}$$

The weighted sum of the squares is now given by:

$$min_{\boldsymbol{c}} \left(\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\phi}\right)^T \boldsymbol{W} \left(\boldsymbol{Y} - \boldsymbol{c}^T\boldsymbol{\phi}\right) \tag{2.5}$$

As the independent variable t changes so too does the variance of the residuals; this is heteroscedastic noise and an example is shown in Equation (2.5) [24]. It is

important to note that in Equation (2.4), in the homoscedastic case the weighted vector $\boldsymbol{W}$ is also there, though this is simply the identity matrix. Consequently, in Equation (2.4) $\boldsymbol{W} = \boldsymbol{I}$.

Taking the derivative with respect to $\boldsymbol{c}$ of Equation (2.5), then rearranging, leads to the weighted least squares estimate of the coefficient vector $\boldsymbol{c}$ denoted as $\hat{\boldsymbol{c}}$:

$$\hat{\boldsymbol{c}} = \left(\boldsymbol{\phi}^T \boldsymbol{W} \boldsymbol{\phi}\right)^{-1} \boldsymbol{\phi}^T \boldsymbol{W} \boldsymbol{Y}$$

Hence leading to the estimate of the function $g(t)$:

$$\hat{g}(t) = \hat{\boldsymbol{c}}^T \boldsymbol{\phi} = \boldsymbol{\phi}^T \hat{\boldsymbol{c}} = \boldsymbol{\phi}^T \left(\boldsymbol{\phi}^T \boldsymbol{W} \boldsymbol{\phi}\right)^{-1} \boldsymbol{\phi}^T \boldsymbol{W} \boldsymbol{Y}$$

3.2.1. ***Choosing*** K ***Basis Functions***. When smoothing the data and therefore ensuring the fit of the estimated function, the choice of K is of central importance. If K is too large then problems of overfitting can arise and the function can become too smooth, conversely, if K is too small, then the estimation may be inaccurate and it will not provide a true representation of the data, possibly resulting potentially important aspects of the data to be missed. In essence, the error from bias is the difference between the models expected prediction and the correctly predicted value, thus the issue is a choice between variance and bias. The error from the variance is the amount of variation between different outcomes of the model and predictions made for a given point. These are both shown below [58]:

$$Bias\left[\hat{g}(t)\right] = \mathbb{E}\left[\hat{g}(t) - g(t)\right]$$

$$Var\left[\hat{g}(t)\right] = \mathbb{E}\left[\left(\hat{g}(t) - \mathbb{E}\left[\hat{g}(t)\right]\right)^2\right]$$

The variance also becomes progressively high when dealing with the higher values for K, and with $K = n$ there will be an unacceptably high estimate of variance. When the number of basis functions, K, are reduced, this also reduces the variance. However, this raises a new problem, as the bias will reach an unacceptable level if K becomes too small. A balance between the two can be achieved through use of the mean squared error, this is essentially just a combination of variance and bias.

$$MSE\left[\hat{g}(t)\right] = Bias^2\left[\hat{g}(t)\right] + Var\left[\hat{g}(t)\right] = \mathbb{E}\left[\left(\hat{g}(t) - g(t)\right)^2\right]$$

Variance increases rapidly when the data is over fitted as a result of using too many basis functions. Bias does not always decrease as a result of increasing K; it is a general assumption that the sampling variance goes up when bias goes down. In order to achieve a good estimate of the smooth trend in the data, it is necessary to accept that there will be some bias [110].

There are a variety of methods that can be used to determine the ideal number of basis functions to use [90]. The two most logical and straightforward methods for choosing K are *stepwise variable selection* and *variable-pruning*. Taking each method in turn, using stepwise variable selection begins with a very small K, then basis functions are added one at a time and after each addition the fit is tested to determine whether it improves significantly and whether the previously added functions are still effective [98]. An example (with the original data superimposed) can be seen in Figure 8. Here consecutive plots of the estimated function of a wild-type plant free-running in DD conditions with increasing values of K is shown. This clearly shows the effect that increasing K has on the fitting of the Fourier basis expansion. When the values of K are low, the estimated curve does not correspond to the shape of the original data, but as the number of K increases, so the estimated function begins to reflect the shape of the original data.

The second method, variable pruning, is often used for data with high dimensions [19]. It is essentially a reversal of the stepwise variable selection method. This method begins by using a high number of K basis functions and reducing them one by one and testing the estimation after each reduction in order to assess whether it holds a significant fit to the data. Thus depending on the amount of variation it contributes, each basis function is dropped, so reducing K, until the majority of the variation within the data is determined.

There are a number of advantages to using basis functions to represent functional data: they can easily store large quantities of data and are ideal to use for large computations [110]; they are helpful to use in the interpolation of functions; they enable data to be expressed in a matrix for which is particularly useful. A particular issue with using basis functions is that the smoothness of the underlying function is largely dependent on the basis function. This results from the process of choosing K discretely and so suggests that the level of smoothing is

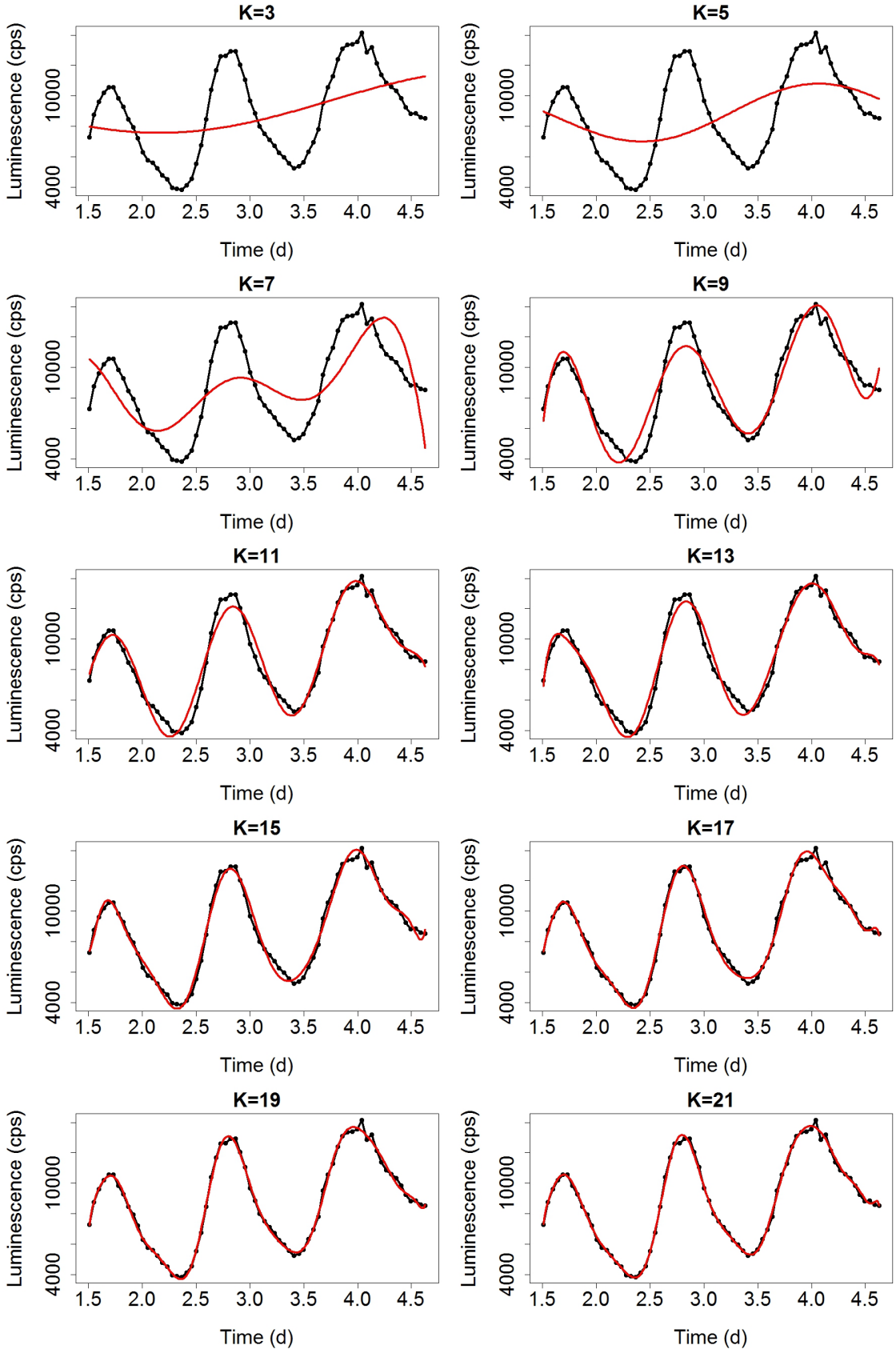


FIGURE 8. Successive plots of the estimated function of a wild-type plant free-running in DD conditions using an increasing number of K Fourier basis functions, with the original data superimposed.

irregular [110]. However, this problem can be rectified with the addition of a roughness penalty.

3.3. **Roughness Penalty.** Using the roughness penalty, also known as penalized least squares regression or the Tikhonov regularization [9], [129] offers a balance between avoiding unnecessary roughness to the curve and fitting the data closely. In addition, it overcomes the limitations and, compared to the weighted least squares method, often gives better results as the smoothness is continuously controlled. The general definition for smoothness in this context is that a given function g has at least one derivative, $D^m g$ refers to the m^{th} derivative and $D^m g\,(t)$ is the value of the derivative taken at t. The roughness penalty is signified by $J\,[g]$ and measures the roughness of g, with λ as the smoothing parameter [51], adjusting Equation (2.3) gives:

$$PENSSE = \sum_{k=1}^{K} \left(Y_k - g\,(t_k)\right)^2 + \lambda J\,[g] \tag{2.6}$$

The degree to which the roughness penalty influences the estimate is determined by the smoothing parameter λ, this is responsible for controlling the bias and variance trade-off (see Section 3.2.1) and the quality of the estimate is dependent on the choice of λ. The roughness becomes progressively more compromised as as $\lambda \to \infty$ and more importance is placed on smoothness of $g\,(t)$ [131]. Conversely, $\lambda \to 0$ increases variability in the curve as the roughness is penalised less, so $g\,(t)$ improves the fit of the data. A common way to quantify the roughness $J\,[g]$ it to examine the square of second derivative of a function g at t. At t, the curvature is measured by $\left[D^2 g\,(t)\right]^2$, it is expected that interest will not be confined to the second derivative square and consequently the following gives a roughness measure that is more general:

$$J_i\,[g] = \int \left[D^i g\,(t)\right]^2 dt, \qquad i = 1, \ldots, m$$

In vector notation, using $g(t) = \boldsymbol{c}^T\boldsymbol{\phi}(t)$ from Equation (2.2) the roughness penalty becomes:

$$\begin{aligned} J_i[g] &= \int \left[D^i g(t)\right]^2 dt \\ &= \int \left[D^i \boldsymbol{c}^T\boldsymbol{\phi}(t)\right]^2 dt \\ &= \int \boldsymbol{c}^T D^i\boldsymbol{\phi}(t)\, D^i\boldsymbol{\phi}^T(t)\, \boldsymbol{c}\, dt \\ &= \boldsymbol{c}^T \left[\int D^i\boldsymbol{\phi}(s)\, D^i\boldsymbol{\phi}^T(s)\, ds\right] \boldsymbol{c} \\ &= \boldsymbol{c}^T\boldsymbol{R}\boldsymbol{c} \end{aligned}$$

where $\boldsymbol{R} = \int D^i\boldsymbol{\phi}(s)\, D^i\boldsymbol{\phi}^T(s)\, ds$.

Now the $PENSSE$ Equation (2.6) can be rewritten in matrix form with the added roughness penalty, this gives:

$$PENSSE = (\boldsymbol{Y} - \boldsymbol{\phi c})^T \boldsymbol{W} (\boldsymbol{Y} - \boldsymbol{\phi c}) + \lambda \boldsymbol{c}^T\boldsymbol{R}\boldsymbol{c}$$

To find an expression for the estimate of the coefficient $\hat{\boldsymbol{c}}$ we first take the derivative with respect to $\boldsymbol{c}$:

$$\frac{\partial}{\partial \boldsymbol{c}} PENSSE = -2\boldsymbol{\phi}^T\boldsymbol{W}\boldsymbol{Y} + \boldsymbol{\phi}^T\boldsymbol{W}\boldsymbol{\phi}\boldsymbol{c} + \lambda\boldsymbol{c}^T\boldsymbol{R}\boldsymbol{c} = 0 \tag{2.7}$$

Rearrangement of Equation (2.7) then gives:

$$\hat{\boldsymbol{c}} = \left(\boldsymbol{\phi}^T\boldsymbol{W}\boldsymbol{\phi} + \lambda\boldsymbol{R}\right)^{-1} \boldsymbol{\phi}^T\boldsymbol{W}\boldsymbol{Y}$$

The hat matrix $\boldsymbol{H}$ is what maps the vector of response values to the vector of fitted values:

$$\hat{\boldsymbol{Y}} = \boldsymbol{\phi}\left(\boldsymbol{\phi}^T\boldsymbol{W}\boldsymbol{\phi} + \lambda\boldsymbol{R}\right)^{-1} \boldsymbol{\phi}^T\boldsymbol{W}\boldsymbol{Y} = \boldsymbol{H}\boldsymbol{Y}$$

3.3.1. ***Choosing the Smoothing Parameter***. Choosing the smoothing parameter λ is important in the same way as the choice of K number of basis functions as it is responsible for controlling the trade-off between the bias and variance of the function. Ideally, averaged over all the data points, λ will minimise the true mean square error [27]. This can be achieved using the Cross validation (CV) method [115]which works by holding back part of the data which is only used for validation and is known as the validation set. The remaining

data is used to train the data set, this is known as the training set. In this way the validation set can be tested to determine the extent to which the model fits the data. To ensure inaccurate results are avoided, the model is tested using different data from that used to train the model, this is achieved by splitting the data set into two groups. When choosing λ the technique of leave-one-out cross validation is used. This is an extreme version of cross validation where validation sample consists of a single observation that is held back and the model is trained using the remaining data. The fitted value for the left out held back observation can then be calculated. By repeating this across the data set and using a different observation as the validation sample each time, it is possible to calculate the cross-validated error sum of squares across a variety of values, the value that produces the minimum then becomes the chosen λ value [110]. Cross validation is an effective and accurate way to determine the smoothing limits however there are two limitations. Firstly, it is computationally expensive so not feasible to employ this method with large sample sizes. Secondly, as CV is minimised, under smoothing of the data can occur as noisy types of variation are favoured, even though it would be better if they were ignored.

Generalized Cross Validation (GVC) is a more popular method that is often used and it gives more realistic results [42]. It is a well established as a simplified version on CV and it overcomes the tendency of CV to under smooth. Using this method the estimate $\hat{\lambda}$ that minimises is selected,

$$GCV\left(\lambda\right) = \left(\frac{n}{n - df\left(\lambda\right)}\right)\left(\frac{SSE}{n - df\left(\lambda\right)}\right)$$

here $df\left(\lambda\right) = trace\boldsymbol{H}$

In respect of λ, in order to minimise GCV trial and error must be used on a large quantity of values for λ and is a similar task to that employed when choosing K basis. However, by using numerical optimisation algorithms or grid searches it is possible for this process to be performed more quickly by [110].

4. Curve registration and alignment

In common with any method of analysis, analysing functional data poses a number of difficulties, one of the most notable issues is the problem of misaligned curves [57]. Biological functional data involves looking from individual to individual and although the typical shape of the curve may be similar, there are differences in dynamics and intensity. Largely as a result of how data is collected. For example, all seeds are sterilised under the same conditions and plated onto MS3 media, but it is not be guaranteed that all seedlings get equal nutrients, some may be subject to slightly more sugar causing them to grow faster and therefore bigger. Similarly, in both the growth chambers and the Topcount imaging machine there is an unavoidable light gradient causing plants to experience slightly different light intensities. Again, the amount of luciferin pipetted into each well is set to 5 μl, however, slight calibration errors can result in slightly different quantities in each well. The process of transferring the seedlings can also cause variation in the results. Great care must be taken when transferring seedlings into the Topcount microtiter plates owing to their delicate nature, as even the slightest damage can cause stress responses that only some of the plants experience. Although individually all these difference are small, they can cause substantial differences in luminescence readings, or timing differences, between identical plants. Consequently the process itself can cause errors in data acquisition.

Differences are reflected in the resulting data represented by the two distinctive types of variation, phase variation and amplitude variation [67]. Figure 9 provides a simple demonstration of these types of variation using plots of the first derivatives of a Gaussian density. The top panel shows phase variation, whilst the bottom panel shows amplitude variation. In this project, variability in phase is attributed to the timings of the features without considering their size, where as amplitude variation shows the difference associated with the intensity in luminescence. In general, observed data are rarely uniform and often these two types of variation are "muddled" together. Although all types of variation play an important role in the analysis and interpretation of data, in some cases it is best to separate these two types of variation as together they further complicate

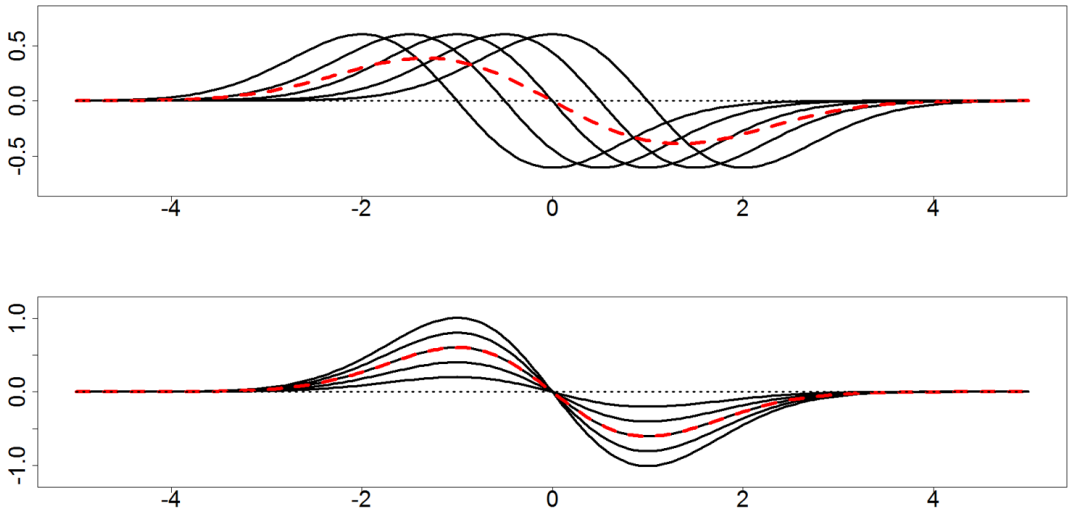


FIGURE 9. The top panel shows five first derivative Gaussian density curves varying only in phase (different μ values) The bottom panel shows five first derivative Gaussian density curves varying only in amplitude (different σ values). The dashed line in each panel indicates the mean of the five curves.

the analysis of curves. Separation is also useful in order to obtain the best fit in relation to a typical curves behaviour as taking an average, when these two types of variation are mixed, is not necessarily representative of data. This is highlighted in Figure 9 where the mean curve is shown as the red dashed curve in both the top and bottom panel. Firstly, focusing on the top panel, it is clear that taking an average over changes in phase does not produce a curve that is representative of a typical curves behaviour. It is apparent that the maximum and minimum of this mean curve is lower than any of the curves in the group. The mean has a dampened shaped compared to any other curve in the group. The x axis cooridinate at which the mean curve hits it maximum is also not representative of what would be the mean x coordinate of the maximum points in the group. The middle x coordinate of the 5 curves maximum points should be at $x = -1$ where as the mean red curve actually places the maximum point at $x = -1.25$. Consequently, when compared to any of the other curves, the mean curve does not reflect the true behaviour of a typical curve. Secondly, the bottom panel shows changes in amplitude, here the mean curve shows a peak at

the correct x coordinate as with all other curves in the group. In addition to this, the amplitude of the peak is in the middle of all curves in the group so providing a good estimate of the behaviour of a typical curve within that group. Overall, Figure 9 clearly illustrates that in order to obtain a strong representation of a typical curves behaviour it may be best to separate these two types of variation.

Although the plants are all subjected to the same conditions, they are still affected by environmental factors that cannot be controlled. Within each genotype tested, there is no genetic variation between each of the 48 replicates, therefore it is reasonable to assume that any variation observed relative to the timing of the features of the curves, is due to some external environmental factors.

Looking at a set of N functions $g_i(t_j)$, where $i = 1, \ldots, N$, the values of any two given functions can differ due to the two types of variation present. Firstly, differences could be attributed to amplitude variation, whereby $g_1(t)$ and $g_2(t)$ may simply, at time t, exhibit different values, while having the same shaped features as displayed in the bottom panel of Figure 9. Secondly, the functions may differ due to phase variation, whereby functions g_1 and g_2 should not be compared at the same time point t as they do not exhibit the same behaviour, an example of this is shown in the top panel of Figure 9. Hence when dealing with phase variation, in order to compare the two functions appropriately, the time scale itself has to be distorted or transformed [110].

4.1. **Landmark registration.** One way to eliminate phase variation is to use a technique called landmark registration. This monotonically transforms the domain for each curve so that features of the curves can be aligned to a specific time argument [68]. This is a well-known method that allows a landmark to be defined as some feature of a curve that can be associated with a specific argument value.

For each curve this process begins with g_i and the corresponding argument values t_{if}, this is $g_i(t_{if})$, where $i = 1, \ldots, N$ and $f = 1, \ldots, F$ where f is each identified landmark feature [110]. This requires computation of a set of smooth strictly monotonic functions h_i of the curve called time-warping functions, these will transform the curves to align the functions to a common argument [41], [68]. When identifying the landmarks to be used for registering, the features

chosen must be shared throughout the data set, irrespective of timing. These can often be identified either visually or alternatively through taking derivatives. For both the DD and LL data sets, 5 landmarks were chosen. These were the first 5 maxima and minimas occuring after 1.5 Days. In particular, these were chosen as the features are present in all curves and are easily recognisable, either at the derivative level, or the actual curves themselves [77]. Figure 10 shows an example of a wild-type plant free-running in DD conditions with these five landmark features identified, for consistency and ease the first derivative was used to identify the landmark occurrences at which $\frac{dg}{dt} = 0$.

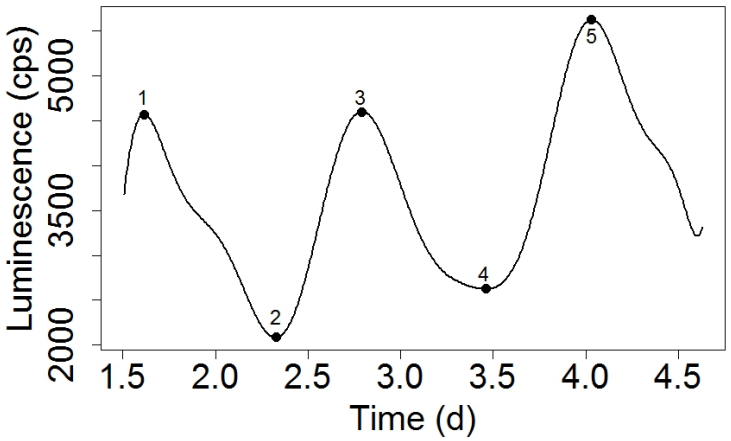


FIGURE 10. Estimated function of a wild-type plant free-running in DD conditions where the characters 1-5 identify the 5 features used for landmark registration.

4.2. **Warping functions.** A warping function h is an element of a convex space $\mathcal{H} \subset \mathcal{W}^m [0, 1]$ also known as Sobolev space [2] [43]. The time warping function must satisfy certain criteria, firstly both the start and end time must be the same as the other curves. Secondly, as the timing of events remains in the same chronological order, irrespective of the timescale, h_i must be a strictly increasing function. Lastly, h_i must be invertible, this ensures that for the same feature, the time points on different timescales correspond to each other [108]. Summarised correspondingly; let g_i be the set of functions defined on some interval $[T_s, T_e]$. Also let $h_i(t)$ be the time warping functions of t for functions i over the common interval $[T_s, T_e]$ and where the operator $/circ$ denotes function composition [116].

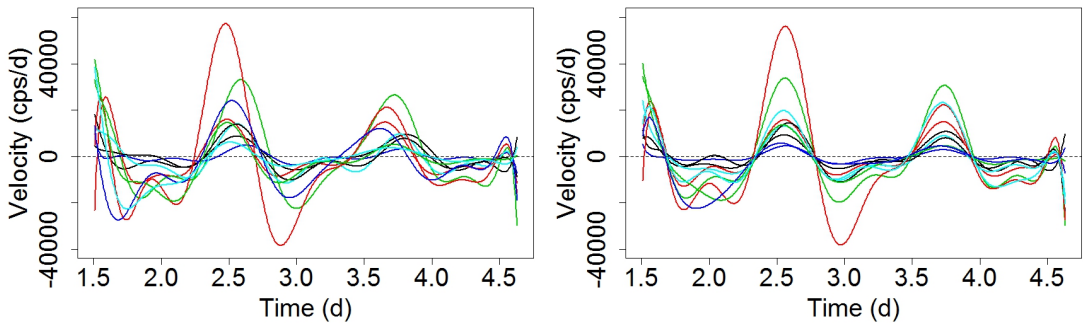
- $h_i(T_s) = T_s$;

- $h_i(T_e) = T_e$;
- $h_i(t_1) > h_i(t_2)$ for $t_1 > t_2$;
- $h^{-1}[h(t)] = (h^{-1} \circ h)(t) = t, \quad \forall\ t$ is uniquely defined.

Let g^* be a fixed function defined over interval $[T_s, T_e]$ then,

$$g^*(t) = (g \circ h)(t) = g[h(t)] \tag{4.1}$$

In order to complete landmark registration, the estimated warping function $h_i(t)$ must be used as shown in Equation 4.1. This requires two steps: firstly, the inverse warping function $h^{-1}(t)$ subject to $h^{-1}[h(t)] = t$ must be computed; secondly, the relationship between $h^{-1}(t)$ on the horizontal axis (abscissa) and $x(t)$ on the vertical axis (ordinate) must be smoothed. Simple interpolation can then be used to obtain the values of the registered function [41]. Figure 11 shows an example of landmark registration carried out on 10 velocity curves of the wild-type parent group free-running in DD conditions. The left panel (Figure 11a) shows the unregistered 10 velocity curves, while the right panel (Figure 11b) shows the registered curves where the 5 landmarks of crossing at zero were used. The act of registering curves, and therefore the reduction/complete cutting out of



(A) Unregistered curves (B) Landmark-registered curves

FIGURE 11. The left panel (A) gives the first derivatives of 10 wild-type plants free-running in DD conditions. The right panel (B) shows the landmark-registered curves corresponding to these, where five crossings at zero were used as landmarks (corresponding to maximum and minimum points in original curves as seen in Figure 10).

phase variation within a group, allows for a much more accurate representation

of a typical curves behaviour when computing the mean as shown in Figure 9. Figure 12 shows the mean curves computed from both the unregistered (black) and registered (red) curves in Figure 11. Corresponding to this and relating back to the original functions, Figure 13 provides the mean estimated functions for the wild-type parent group for the first 10 curves for both unregistered (black) and landmark registered curves (red). It can be seen in both Figure 11 and 13, that averaging curves using the registered group, provides a much more realistic depiction of the behaviour and features of a typical wild-type plant free-running in DD conditions. The black curve shown in both Figure 12 and Figure 13 has far more dampened features as, when the curves are unregistered, computing the mean curve effectively borrows from amplitude to accommodate the variation in phase. Therefore, as would be expected, the mean of the registered curves seen in red show much sharper and well defined features.

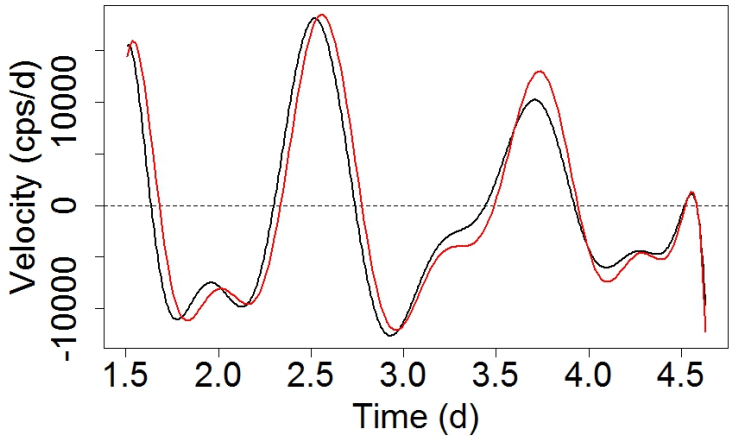


FIGURE 12. The first derivative mean curves corresponding to the unregistered curves in Figure 11a (black) and the landmark registered curves in Figure 11b (red).

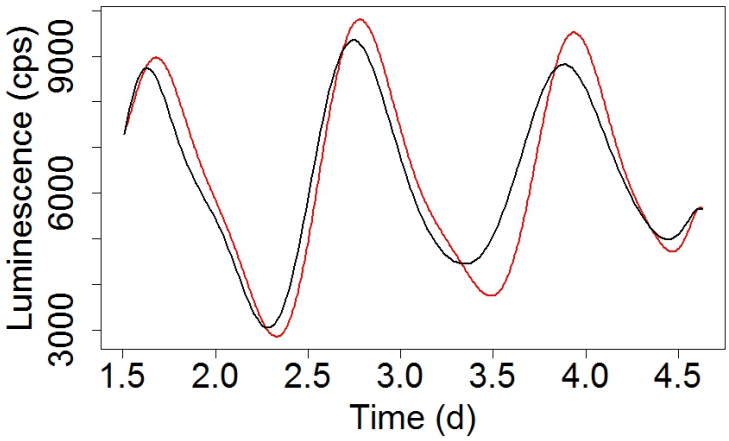


FIGURE 13. The mean curves corresponding to the first 10 curves in the wild-type parent group free-running in DD conditions for both unregistered curves (black) and the landmark registered curves (red).

5. Derivative Analysis

As previously mentioned in Section 3.3 at time t, the m^{th} derivative of a function g is given by $D^m g(t)$, also represented by $\frac{d^m g}{dt^m}$. The velocity is given by the first derivative Dg, acceleration is given by the second $D^2 g$.

5.1. **Velocity and acceleration.** The application of FDA techniques allows data to be represented as smooth differentiable functions. The first derivative, velocity, enables the rate of change of behaviour to be observed easily. It provides the opportunity to gain insight into the behaviour of the curve in relation to whether it is increasing or decreasing and by how much. In addition, further information about the curve can be seen by examining the second derivative, acceleration $D^2 g$. This provides information concerning the rate of change of velocity in relation to time. If the first derivative (velocity) is increasing, then a positive second derivative (acceleration) will present a convex, upward facing curve. Conversely, if the first derivative (velocity) is decreasing, then a negative second derivative (acceleration) will present a concave, downward facing curve.

An example wild-type curve from the DD data set can be seen in Figure 14. This shows the luminescence curve (top) and both the corresponding velocity (middle) and acceleration (bottom) curves. In order to explore and identify the characteristics of the data it is important to be able to examine the derivatives. This is apparent in the example shown in Figure 14, between 1.5 and 2.5 Days. Focusing on the top panel, between 1.5 and 2.5 Days. The portion of the curve between the maximum and minimum appears relatively straight, only a small deviation can be seen. It may not seem of importance however, it becomes apparent that this is a prominant feature when the curves are viewed as both their first and second derivatives. Focusing on the velocity curve in the middle panel, immediately preceding 2 Days, it is clear that the velocity begins to increase, it then makes a slight decrease before it continues to increase. This pattern is also reflected in the acceleration curve shown in the bottom panel. Here, as the curve is increasing, it crosses the y axis, it then crosses back down the y axis at approximately 2 Days and finally continues to increase after crossing the y axis again. This highlights the importance of using derivative plots. It allows

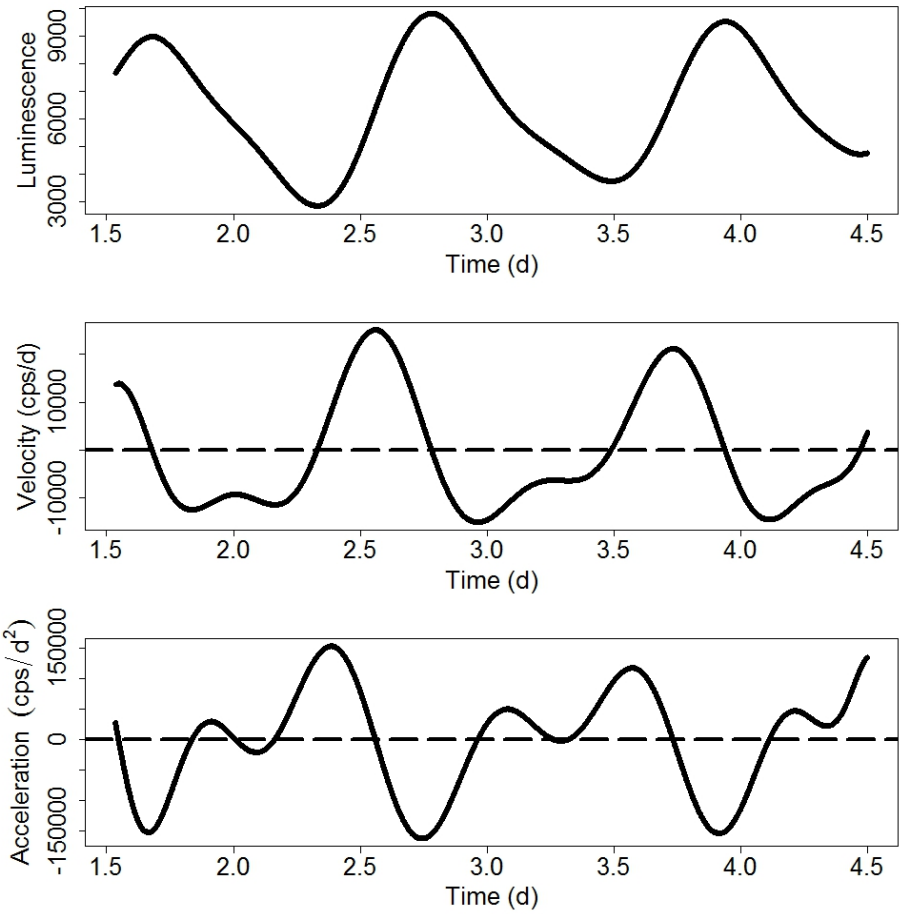


FIGURE 14. An example of wild-type plant subject to DD conditions plotted as luminescence against time (top), first derivative curve relative to time (middle) and second derivative curve relative to time (bottom).

characteristics, which may otherwise have been overlooked, to be easily identified and explored.

Across all genotypes this feature is prominent. In fact it is present in all data sets, although to different degrees of intensity and a dampening of this effect does occur across each of the DD and LL experiments.

5.2. **Phase-plane Plots.** In order to compare the relationship and trade-off between velocity and acceleration they are plotted against each other. The resulting plot is known as a phase-plane plot and provides a 2-dimentional diagram with velocity and acceleration being the x and y Cartesian axes respectively (example seen in Figure 15). What is considered to be potential energy is associated with

the vertical axis and kinetic energy the horizontal axis [104]. Consequently, for the mechanical system, a particular point in the phase-plane plot, at any given time t, has a certain velocity and acceleration.

Two types of energy can be attributed to an object, kinetic energy from the movement of the object and potential energy from its position within a system [66]. By plotting the derivatives in this way, the behaviour of a system that may be difficult to interpret can be easily visualised. It can show how a function moves between different energy states. This allows for easy measurement of the time at which the curve most active (also called acrophase), and the time of greatest change (also known as inflection phase) [111]. A plot showing a perfect circle centring around $(0, 0)$ denotes the precise periodic motion of a function, therefore, the energy path of the function for one period is shown. Thus, relationships that exist between and within functions can be identified through this critical aspect of functional data analysis. In Figure 15, the phase-plane plot of $\sin(t)$ has a period of length 1. A point of constant velocity is apparent where the horizontal axis is crossed and indicates a maximum or minimum point of the velocity curves. Similarly, when the vertical axis is crossed it corresponds to a maximum or minimum point on the original luminescence curves.

Figure 16 displays how phase-plane plots can be used to demonstrate differences between unregistered and landmark registered curves. Here the phase-plane plot of a wild-type mean function for both unregistered (black) and landmark registered (red) free-running in DD conditions is shown. For this project, each of the phase-plane plots that are shown include markers that appear along the length of the curve and each of these markers corresponds to a particular time. For example, the velocity and acceleration of the curve at 1.5 Days (or 36 hours) is indicated by the point on the curve that is labelled 1.5. In addition, a cycle signifies one period of the curve and refers to the path of the original function from a minimum to a minimum point. Registration affects the relationship between the kinetic and potential energy by drawing in and so producing a tighter ellipsis, so providing a more detailed, sharper representation of the mean curve within the group which is representative of the curve alignment.

Figure 17 displays the luminescence graph (left). The wild-type (black) and mutant (red) from the DD data set is seen. The difference in amplitude between

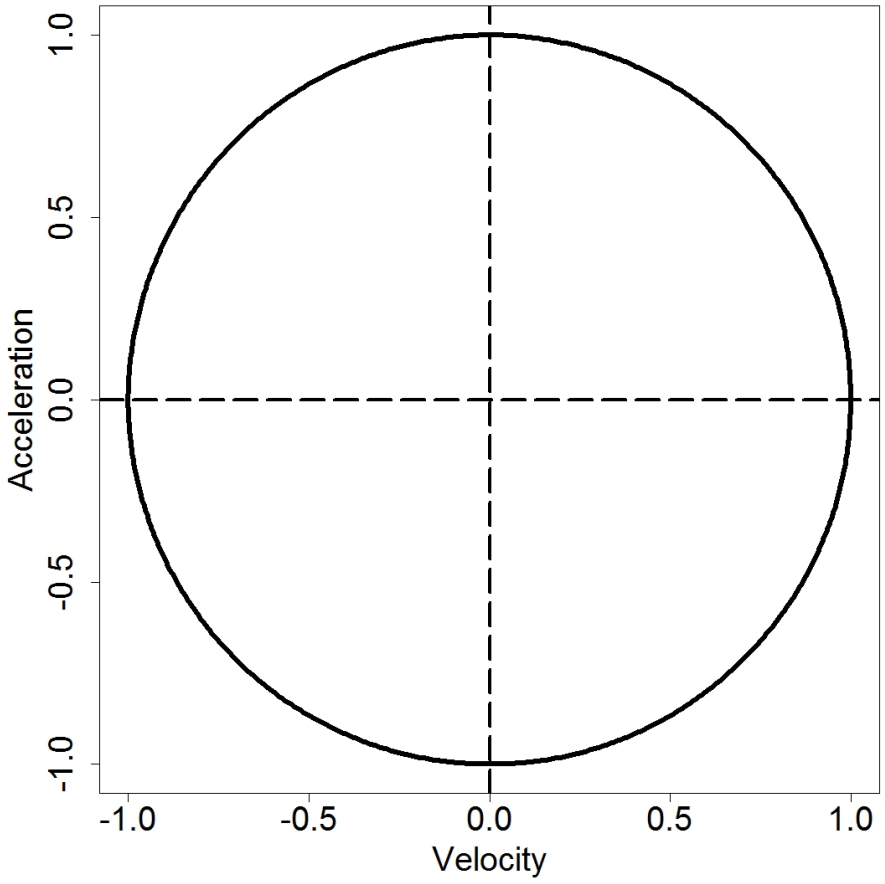


FIGURE 15. A phase-plane plot of harmonic function $sin\,(t)$. Kinetic energy is maximised when acceleration is 0, and potential energy is maximised when velocity is 0.

the two types is clearly apparent. It is important to note that there are also changes in phase, though these are less easily identified. Nevertheless, exploring relationships within the population is made much clearer and easier through the use of phase-plane plots. As shown in the plot on the right in Figure 17, using the phase-plane technique is the best way to illustrate the differences between these two curves. In this case, behaviour of the period of the curve is determined by the shape of the ellipse around $(0, 0)$, maximum velocity and acceleration are signified by a perfect circle, so indicating they are proportionally equal, consequently the two energy types (velocity and acceleration) are equal. Period lengthening, or a slower circadian clock, results in an elliptical shape in the horizontal plane. On the other hand, period shortening, or a faster running circadian clock, results in an elliptical shape in the vertical plane

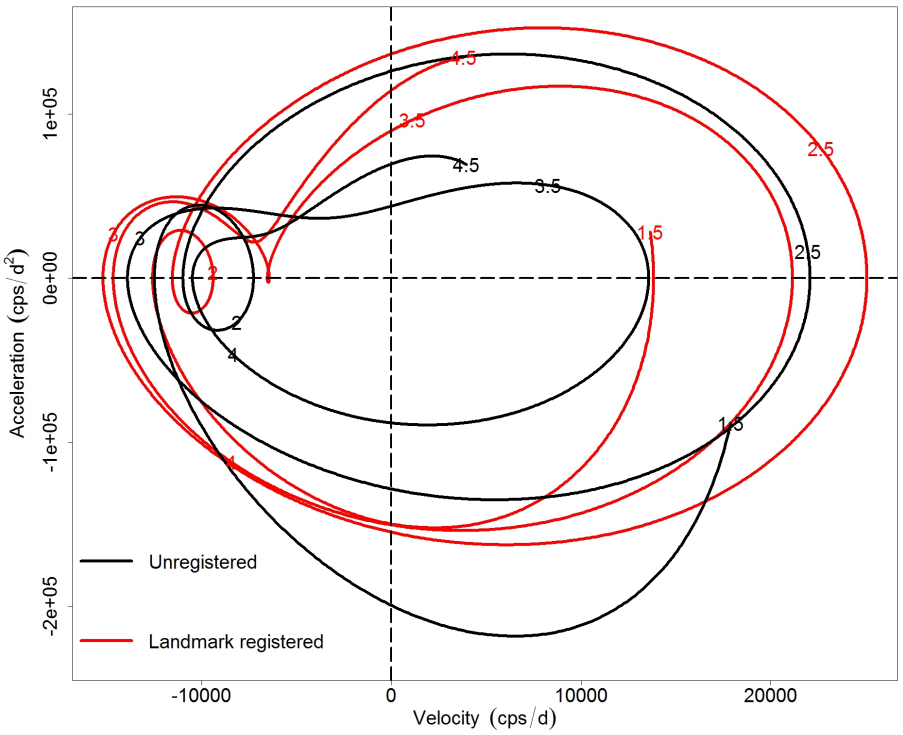


FIGURE 16. A phase-plane plot of a wild-type mean function for both unregistered (black) and landmark registered (red) curves over approximately 3 days (72 hours). Along each curve labels are placed every 0.5 days (12 hours).

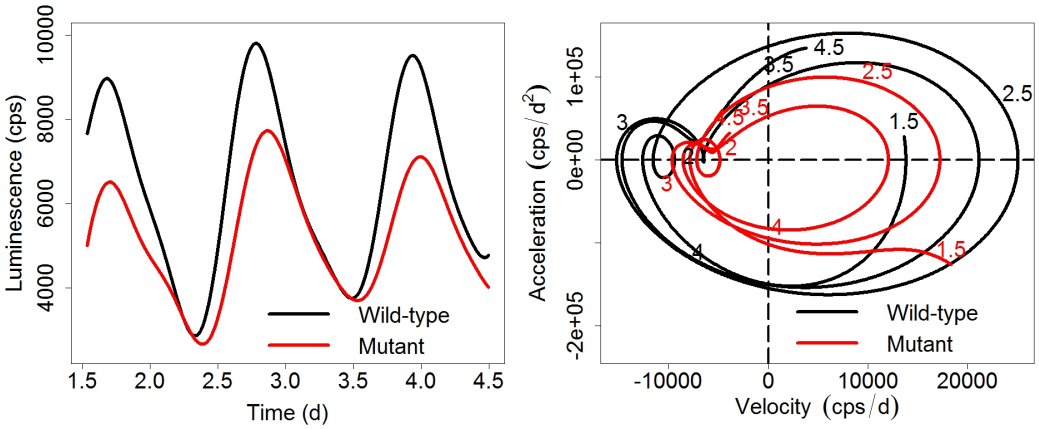


FIGURE 17. Estimated mean landmark registered functions (left) and a phase-plane plot of the first two derivatives of the mean landmark registered functions (right) for wild-type (black) and mutant (red) free-running in DD conditions measured over approximately 3 days (72 hours). Along each curve in phase-plane plot (right) labels are placed every 0.5 days (12 hours).

This effect is shown more clearly in Figure 18 which again shows both the wild and mutant plants, however here only the first cycle is shown, from $1.5 - 3$ Days. In this case the ellipsis was found not to be centred around $(0, 0)$, but rather it is shifted to the right where it primarily lies within positive velocity and is representative of the increasing amplitude of the curves over time. Once again this feature of the data becomes much more apparent when plotting in the phase-plane. The wild-type curve in Figure 18 begins in positive acceleration and

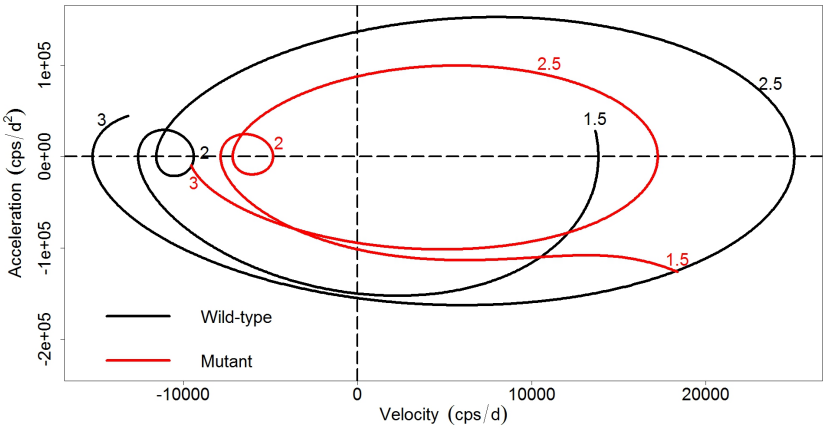


FIGURE 18. A phase-plane plot of a wild-type (black) and mutant (red) mean landmark registered function measured over approximately 1.5 days (36 hours). Along each curve labels are placed every 0.5 days (12 hours).

velocity before moving clockwise where it crosses the vertical axis corresponding to the first peak seen in the left plot of Figure 17, this represents the start of the first cycle. The curve then loops round crossing the horizontal axis, it then immediately culminates in a cusp at 2 Days. From this point the cycle then increases in potential energy, it crosses the vertical axis and reaches the maximum potential energy before 2.5 Days. It then curves downwards where it crosses the horizontal axis before it loops round to complete the cycle by crossing the vertical axis once again with negative acceleration. This corresponds to the second peak in the left plot in Figure 17 where the curve is at its maximum potential energy, the period of this cycle is 1.101 Days (approx 26:25:40).

Focusing on the *hsp90.2-3* mutant type curve, the first cycle starts to cross the vertical axis at a lower potential energy than the wild-type, it loops round to cross the horizontal axis also with a lower kinetic energy than the wild-type. The cusp

that occurs around 2 Days is also present here and the cycle continues clockwise where it crosses the vertical axis again and reaches its maximum potential energy just prior to 2.5 Days. It continues round and reaches maximum kinetic energy after 2.5 Days where it crosses the horizontal axis. The cycle then loops round to cross the vertical axis and so completes the first cycle, the period of this first cycle is 1.164 Days (approx 27:56:27) which is an increase in period length compared to the wild-type plants, This confirms that the *hsp90.2-3* mutation is a period lengthening mutant as it causes the clock to run slower.

It is interesting to note that the mini loop occurs when the plants would normally be expecting to make a dark to light transition, under entrainment conditions this is the time point that dawn would begin. This shows that even though the plant has been subject to DD for 48 hours, under the entrainment conditions the circadian clock works in anticipation of dawn, this indicates that the onset of dawn alone does not drive the plants rhythm. It is instead a joint effort and results from all input pathways working together to entrain the plants rhythms to a particular schedule that is relative to the environmental cycles of light and temperature. This means that the way the sequence relates to the environmental cycle alters, although the internal sequence of events does not, that is, organisms are allowed by the circadian system to anticipate these cycles relative to environmental cycles [88].

As previously mentioned, viewing data in the phase-plane enables identification of particular features of the data that would otherwise be overlooked. Figure 18 shows a good example of such a feature, where mini loops can be seen clearly in both the wild-type and mutant around 2 Days. This feature is present in the first and second cycle for all genotypes for the DD data set. However, in the second cycle, the degree to which this mini loop occurs is somewhat reduced and in some genotypes it is not present at all, in fact this dampening effect generally increases throughout each cycle in the experiment. Nevertheless, the shape of the cycle is still not uniform and it does show a cusp or pinch, in the types of energies present. When directly compared to the wild-type, mutant genotypes show a particularly dulled mini loop, or pinch. This is illustrated in Figure 19 which shows the second cycles for both the wild-type and mutant plants, here it can clearly be seen that the mini loop is dampened in the wild-type and disappears in

the mutant type leaving only a slight pinch in the cycle. This highlights, not only that the plants ability to control their clock in constant darkness deteriorates over time as the anticipation of dawn effect decreases, but also that the ability of the clock to tell the time itself, deteriorates over time. The expected time of dawn under entrainment conditions would be 3 Days (72 hours), but it is clear from the plot in Figure 19 that this feature takes place after the expected time of dawn, in fact, for the wild-type this occurs at 3.260 Days (approx 78:14:00), 6 hours and 14 minutes later than would normally be expected. For the mutant type the pinch point occurs at 3.326 Days (approx 79:50:00), which is 9 hours and 50 minutes after expected dawn. This result is consistent with the *hsp90.2-3* mutants direct effect on the clock to tell the time and so causes the clock to run slower. A

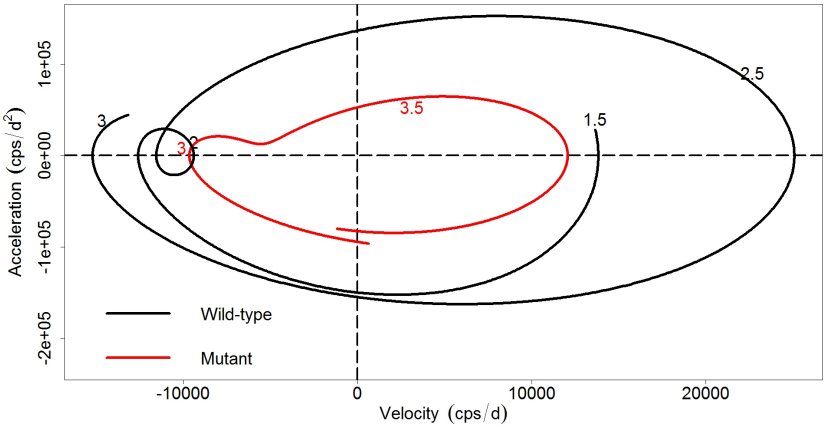
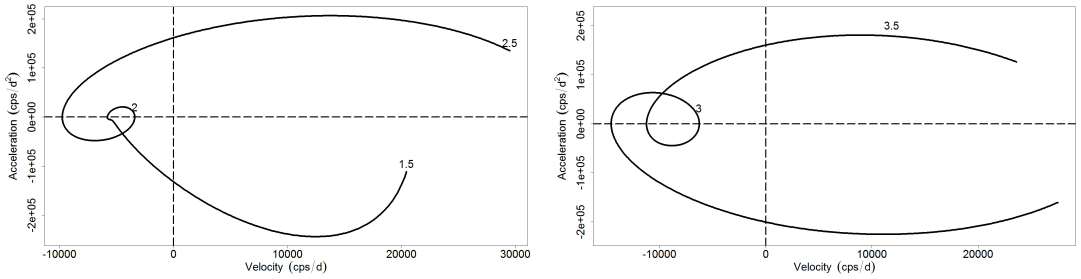


FIGURE 19. A phase-plane plot of a wild-type (black) and mutant (red) mean landmark registered function measured over approximately 1.5 days (36 hours). Along each curve labels are placed every 0.5 days (12 hours).

further illustration the clocks ability to anticipate dawn is shown in Figure 20 from the LL data set, where Figure 20a shows this feature during the first cycle for the wild-type and Figure 20b shows the feature during the second cycle, also for the wild-type. The loop appearing in the first cycle occurs at around 2 Days, similarly to the Ws wild-type curve from the DD data set. Again this happens when the plant under entrainment conditions would normally experience a dark to light transition, however in this case the plant is already in constant light. By examining Figure 20b it is clear that this feature is still prominent and is more defined compared to the second cycle of the DD data. Alongside this, the time

of the feature in the second cycle is at 3 Days suggesting that the clock is more robust under LL conditions.



(A) (B)

FIGURE 20. Phase-plane plots of a wild-type mean landmark registered function under LL conditions. Both the first (A and second (B) anticipated light cue regions shown. Along each curve labels are placed every 0.5 days (12 hours).

In order to provide a comparison of this feature in both the DD and LL data sets, Figure 21 shows the phase-plane plot of the first cycle from the SD data set whilst still under entrainment conditions, that is the wild-type was still subject to external light dark cycles. This shows a perfect example of the relationship between the potential and kinetic energy during a cycle where the mini loop feature is large and well defined. As expected this occurred just prior to 2 Days in anticipation of the onset of dawn at exactly 2 Days. Using this plot as a comparison highlights that the ability of the clock to anticipate these light cues becomes increasingly dampened when it is subjected to constant conditions, this effect is at its strongest free-running in DD conditions. It also emphasises the fact that the clocks ability to tell the time is also adversely affected over time when it is subjected to free-running conditions, such that, with each cycle the time the anticipation feature is observed lengthens. Once again this was observed most strongly in the DD data set with the wild-type being over 6 hours later than the time of dawn under entrainment.

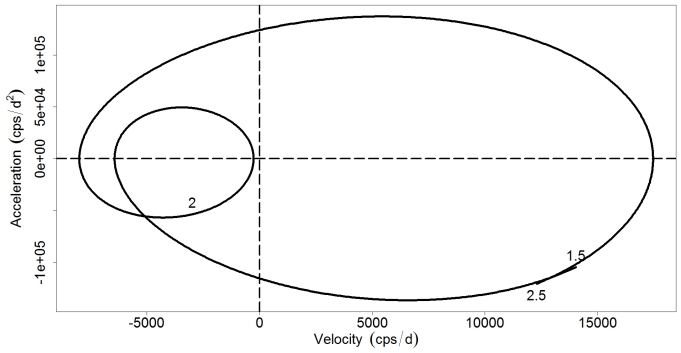


FIGURE 21. Phase-plane plots of a wild-type mean landmark registered function from SD data set whilst still under 12L:12D entrainment conditions. Measured over approximately 1 day (24 hours). Along each curve labels are placed every 0.5 days (12 hours).

6. Variance - Covariance and Correlation Analysis

The following explanation of methods used in the analysis is taken from Lock 2016 [78].

As previously noted, there is variation both across functions within genotypes and between functions in different genotypes. The variance indicates how widely individuals within a group vary from each other. Whilst the classical definition of covariance is the measurement of the degree to which two variables are linearly associated, in a functional approach, covariance measures the association between different time values within a group. Analysing the variance and variance-covariance functions within a group allows for identification of the main types of variability, this is useful as it provides further insight into the groups behaviour. To find the variance between a group of N curves $g_i(t)$ the equation below is used :

$$Var_g(t) = \frac{1}{N-1}\sum_{i=1}^{N}\left(g_i(t) - \bar{g}(t)\right)^2, \qquad \text{where } i = 1, \ldots, N$$

The covariance between curve values $g_i(t_1)$ and $g_i(t_2)$ at times t_1 and t_2, where $i = 1, \ldots, N$, is specified by the bivariate covariance function and is given by the following equation:

$$Cov_g(t_1, t_2) = \frac{1}{N-1}\sum_{i=1}^{N}\left(g_i(t_1) - \bar{g}(t_1)\right)\left(g_i(t_2) - \bar{g}(t_2)\right) \tag{3.2}$$

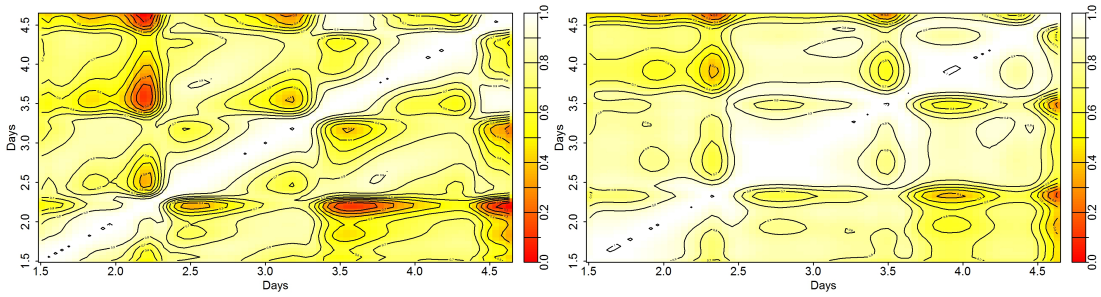
As stated above, important insights concerning the variability within a group can be gained through the varience-covarience functions, however, there are frequently measured on different scales so leading to difficulty with interpretations as relative comparisons are difficult. Using a correlation function is a much more accessible tool for interpreting the relationship within groups and can overcome this problem [76]. Correlation shows both the degree to which variables have a tendency to move together and, whether they are related positively or inversely. In functional terms, the correlation coefficient provides an indication of the association between the two functional observations $g_i(t_1)$ and $g_i(t_2)$ of the same quantity [107]. The correlation function that is equivalent Equation (3.2) is:

$$Corr_g(t_1, t_2) = \frac{Cov_g(t_1, t_2)}{\sqrt{Var_g(t_1)\,Var_g(t_2)}}$$

Thus the cross-correlation function consists of estimated correlations between the set of function values at time t_1 and the set of functions values at time t_2. The value of the interval $[-1, 1]$ is always taken on by the correlation coefficient. A value that is positive indicates high values of $g(t_1)$ are related to high values of $g(t_2)$ or equally low values of $g(t_1)$ are related to low values of $g(t_2)$. Conversely, a value that is negative indicates high values of $g(t_1)$ are related to low values of $g(t_2)$ or vice versa [38]. Likewise it is apparent there is no association between $g(t_1)$ and $g(t_2)$ if the coefficient value is zero. Although this indicates that these is no linear relationship, it does not necessarily indicate that they independent [20].

This information can be displayed graphically as shown in Figure 22, where the contour plotted correlation function values of the wild-type plants free-running in DD conditions are shown for both the unregistered (left) and landmark registered (right) curves. Complementary to this Figure 23 shows the corresponding contour plotted correlation function as a surface over the plane of possible pairs of time.

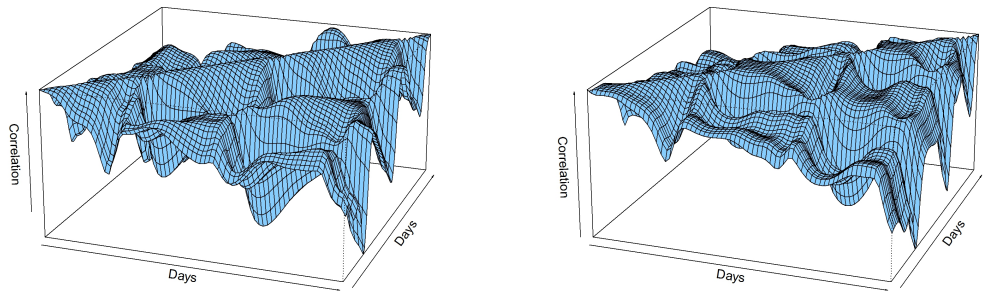
In the contour correlation plots, the line running along the diagonal from the



(A) (B)

FIGURE 22. The estimated correlation surface of wild-type free-running in DD conditions presented as contour plots. The left panel (A) represents the unregistered curves, the right panel (B) represents the landmark-registered curves. Correlation values are represented through the colour key ranging from 0 (red) to 1 (white).

bottom left corner $(1.5, 1.5)$ to the top right $(4.5, 4.5)$, represents, within the groups, the correlation at the matching points in time i.e. $Corr_g(t_j, t_j)$, therefore, 1 are the expected values. The speed at which the two time arguments become separated is illustrated by line that runs perpendicular to this diagonal and as



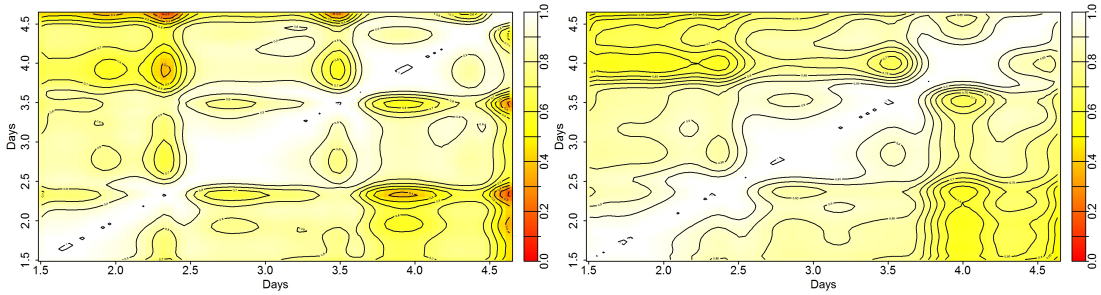
(A) (B)

FIGURE 23. The estimated correlation surface of wild-type free-running in DD conditions presented as perspective surface plot over the plane of possible pairs of time. The left panel (A) represents the unregistered curves, the right panel (B) represents the landmark-registered curves.

expected, Figure 22a shows that across the diagonal there is high correlation and from this point, moving outwards, although the correlation decreases it does remain positive. Around 2.5 and 3.5 Days there are areas of fast separation moving perpendicular to the diagonal, in between these areas the separation is less intense. Past these areas of fast separation, parallel to the diagonal, is further very high area of positive correlation that can also be identified in Figure 23a as the fairly flat highest areas. This is representative of the correlation between functions at different times, whereby the first peak is positively correlated to the second and third peak. In comparison, the landmark registered group of curves shows similar behaviour with a strong correlation across the diagonal, but with a much wider area of high correlation when moving away from the diagonal. Again this highlights the important role that landmark registration has in aligning functions and extending the amount to correlation within a group, the large areas of high positive correlation are due to the functions all moving positively together.

Figure 24 shows the contour plotted correlation function values of the land mark registered wild-type (left) and mutant (right) free-running in DD conditions, and Figure 25 shows the corresponding contour plotted correlation function as a surface over the plane of possible pairs of time. By firstly inspecting the mutant curves in Figure 24b there is, as expected, a very strong correlation across the

diagonal, and highly positively correlated areas remain when moving perpendicularly outwards. Around 4 Days there is an area of fast separation , along with dampened separation around 2.5 and 4.5 Days. These areas are all consistent with the maximum and minimum points observed in Figure 17. Presenting data

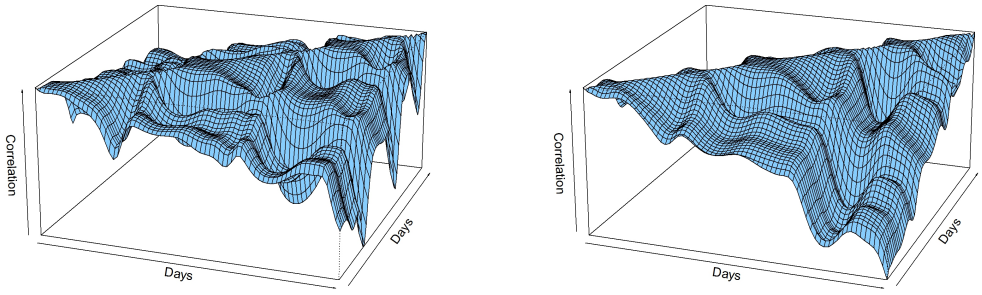


(A) A contour plot of wild-type registered curves estimated correlation surface free-running in DD conditions.

(B) A contour plot of mutant landmark registered curves estimated correlation surface free-running in DD conditions.

FIGURE 24. The estimated correlation surface of the landmark registered wild-type (A) and mutant (B) subject to DD conditions landmark registered curves presented as contour plots. Correlation values are represented through the colour key ranging from 0 (red) to 1 (white).

is in this way allows for further comparisons between the wild and mutant type to be explored and looking between the landmark registered curves of the wild and mutant type plots, the difference between them is clear. In general, the mutant type plants show more positive correlation across the surface. This suggest that the *hsp.90.3-2* mutant causes the plants to behave more similarly to one another. The comparison of wild and mutant plants can also be identified in the corresponding contour plotted correlation function as a surface over the plane of possible pairs of time seen in Figure 25, where notably the wild-type perspective is higher in positive correlations and is reflected in the much smoother surface.



(A) (B)

FIGURE 25. The estimated correlation surface of the landmark registered wild-type (A) and mutant (B) free-running in DD conditions presented as perspective surface plot over the plane of possible pairs of time.

7. Principal Component Analysis (PCA)

The following explanation of methods used in the analysis is taken from Lock 2016 [78].

Principal component analysis (PCA) has its origins in multivariate analysis, it is an unsupervised learning method that is used in exploratory data analysis to detect patterns or grouping in data [72]. Karl Pearson first described (PCA) in 1901 [97] and in 1933 Harold Hotelling was responsible for furthering its development [55]. Typically, a data set consists of many interrelated variables and the primary function of PCA is to preserve as much of this variation as possible whilst reducing the dimensionality of the data set [59]. The principle aim of this method is to identify the types of variation contained within the data and in order to achieve this, it is changed into principal components which consist of a new reduced set of variables providing an effective summary of the observations, whilst still retaining all the important information that is contained in the data. Thus the principal components are selected in order to ensure that the maximum amount of information concerning the original variables is added with each consecutive component [28] and is reached from the eigenvalue decomposition of the data covariance matrix. A principal component in multivariate data analysis can be defined as follows [63]:

Definition 7.0.1. Let $\boldsymbol{g}^T = (g_1, g_2, \ldots, g_m)$ be a vector with mean $\bar{g} = \mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}$. The main aim of this is to identify, in order, the most informative k linear combinations of a new set of variables $p_1, p_2, \ldots, p_k$. The principal components for this are as follows:

$$
\begin{aligned}
p_1 &= a_{11}g_1 + \cdots + a_{1m}g_m = \boldsymbol{a_1}^T\boldsymbol{g} \\
p_2 &= a_{21}g_1 + \cdots + a_{2m}g_m = \boldsymbol{a_2}^T\boldsymbol{g} \\
&\vdots \\
p_k &= a_{k1}g_1 + \cdots + a_{km}g_m = \boldsymbol{a_k}^T\boldsymbol{g}
\end{aligned}
$$

Focusing on the first principal component which is $p_1 = \boldsymbol{a_1}^T\boldsymbol{g}$, the spread of the p value across all the observations determines the choices of coefficients $a_{11}, \ldots\ a_{1m}$.

The variance is used to measure this spread and is given in the form;

$$Var\left(p_1\right) = \sum_{i,j=1}^{m} a_{1i}a_{1j}\sigma_{ij} = \boldsymbol{a_1}^T\boldsymbol{\Sigma a_1}$$

Conditional upon the constraint $\boldsymbol{a_1}^T\boldsymbol{a} = 1$ it is chosen to maximise $\boldsymbol{a_1}^T\boldsymbol{\Sigma a_1}$, this is achieved by using Lagrange multipliers [12]. The Lagrangian function is defined by,

$$L\left(\boldsymbol{a_1}\right) = \boldsymbol{a_1}^T\boldsymbol{\Sigma a_1} - \lambda\left(\boldsymbol{a_1}^T\boldsymbol{a_1} - 1\right) \tag{4.1}$$

Differentiating Equation (4.1) gives,

$$\boldsymbol{\Sigma a_1} = \lambda\boldsymbol{a_1}$$

As a result it is shown that $\boldsymbol{a_1}$ should be selected to be an eigenvector of $\boldsymbol{\Sigma}$, for example, called $\boldsymbol{e}$, with an eigenvalue λ. Supposing that eigenvalues of $\boldsymbol{\Sigma}$ are ranked in decreasing order $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ then,

$$\begin{aligned} Var\left(p_1\right) &= \boldsymbol{a_1}^T\boldsymbol{\Sigma a_1} \\ &= \lambda\boldsymbol{a_1}^T\boldsymbol{a_1} \\ &= \lambda \end{aligned}$$

As a consequence of this, in relation to the largest eigenvalue λ_1 of $\boldsymbol{\Sigma}$ $\boldsymbol{a_1}$ should be chosen as the eigenvector $\boldsymbol{e}$ so maximising $Var\left(p_1\right)$.

A similar method is used to calculate the second principal component, however, as it is important that the subsequent principal components are not correlated, an additional constraint is applied to the calculation as follows: $\boldsymbol{a_2}^T\boldsymbol{a_1} = \boldsymbol{a_1}^T\boldsymbol{a_2} = 0$. The subsequent principal components are also calculated in this way, consequently reducing the problem to solving [134],

$$\boldsymbol{\Sigma a} = \lambda\boldsymbol{a}$$

If the covariance matrix $\boldsymbol{\Sigma}$ has the eigenvalue-eigenvector pairs $\left(\lambda_1, \boldsymbol{e_1}\right), \left(\lambda_2, \boldsymbol{e_2}\right), \ldots \left(\lambda_p, \boldsymbol{e_p}\right)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$, the k^{th} principal component is then represented by;

$$p_k = \boldsymbol{e_k}^T\boldsymbol{g} = e_{j1}g_1 + e_{j2}g_2 + \cdots + e_{jm}g_m, \qquad j = 1, 2, \ldots, m.$$

7.1. **Functional PCA (FPCA).** Variance-covariance and correlation functions do not always provide a clear and understandable view of the variability of the data and so can be difficult to interpret. This problem can be tackled by using functional PCA (FPCA) which continues the fundamental purpose of PCA by following through into an uninterrupted functional form [110].

When dealing with functional data, it is first centred. Therefore the mean curve is subtracted from each of the relevant original curves. When using FPCA an additional change arises as the vectors that were previously observed in the multivariate case can now be represented through the summation:

$$p_i = \boldsymbol{a_i}^T \boldsymbol{g} = \sum_{i,j=1}^{m} a_{ij} g_j$$

are replaced with functions, these linear functions of the curves can be represented by integrals [59]:

$$p_i = \int a\left(t\right) g_i\left(t\right) dt$$

Thus, the inner product space of a sequence of numbers $\langle \cdot, \cdot \rangle_{l^2}$ is replaced by the functional version $\langle \cdot, \cdot \rangle_{L^2}$. Here the inner product of functions say $g\left(t\right)$ and $h\left(t\right)$ denoted $\langle g, h \rangle_{L^2}$ is:

$$\langle g, h \rangle_{L^2} = \int g\left(t\right) h\left(t\right) dt$$

Deciding on the weight function $\alpha_1\left(t\right)$ to maximise is the initial step in FPCA;

$$\frac{1}{n-1} \sum_{i=1}^{n} p_{i1}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left[\int \alpha_1\left(t\right) g_i\left(t\right) dt \right]^2, \qquad i = 1, \ldots, n$$

subject to the unit sum of squares constraint $\int \alpha_1\left(t\right)^2 dt = 1$ [110]. The idea that each principal component is orthogonal from the previous principal component is known as the orthogonality constraint and in the same way as with PCA, the orthogonality constraint must be satisfied in each successive step of the weight function p_i. In functional data terms, this is measured through the inner product, whereby $\langle g, h \rangle_{L^2} = 0$ Subsequent functional principal components are then defined successively as:

$$p_{ik} = \frac{1}{n-1} \sum_{i=1}^{n} \left[\int \alpha_k\left(t\right) g_i\left(t\right) dt \right]^2, \qquad i = 1, \ldots, n$$

7.1.1. ***Calculating Functional Principal Components***. Solving the functional version of the eigenequation, the eigenfunction, provides a more easily manageable approach to identifying the functional principal components. It is firstly necessary to define the covariance of the sample between $g(s)$ and $g(t)$, this is comparable to the multivariate case where it is assumed that the sample mean is zero, similarly, in the functional case it is assumed that the curves means are zero. This covariance function $v(s,t)$, is as follows [119]:

$$v(s,t) = \frac{1}{n-1}\sum_{i=1}^{n} g_i(s)\, g_i(t), \qquad i = 1, \ldots, n$$

To find the functional principal components, the eigenfunction that needs to be solved is:

$$\int v(s,t)\,\alpha(t)\,dt = \lambda\alpha(s) \tag{4.2}$$

There are a number of ways to solve Equation (4.2),one straightforward resolution is to make the data discrete and perform conventional PCA on the values that were identified from the $(n \times p)$ matrix where the i^{th} row comprises the values $g_i(t_1), \ldots, g_i(t_p)$ [127]. Having completed PCA, the eigenvectors that were identified should be transformed into functional form by renormalizing them before using a suitable smoother to estimate the values [105].

Alternatively, Equation (4.2) can be solved by using known basis functions where, similar to the process shown in Section 3.1, each function g_i is expressed as a linear combination of the aforementioned basis functions, consequently, each function g_i will then have a basis expansion in the form:

$$g_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t) \qquad i = 1, 2, \ldots, n \tag{4.3}$$

In matrix formation it is essential to represent the Equation (4.3) where $\boldsymbol{C}$ is the $n \times K$ coefficient matrix and $\boldsymbol{g} = (g_1(t), \ldots, g_n(t))^T$, $\boldsymbol{\phi}(t) = (\phi_1(t), \ldots, \phi_K(t))^T$ then:

$$\boldsymbol{g}(t) = \boldsymbol{C}\boldsymbol{\phi}(t)$$

So from Equation (4.2) the covariance function in matrix notation becomes:

$$v(s,t) = \frac{1}{n-1}\boldsymbol{\phi}^T(s)\,\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\phi}(t)$$

At this point, taking any eigenfunction α, can be stated in terms of the basis function in Equation (4.3):

$$\alpha(s) = \sum_{k=1}^{K} b_k \phi_k(s) = \boldsymbol{\phi}^T(s)\,\boldsymbol{b}$$

It follows that, using these results, Equation (4.2) gives:

$$\begin{aligned}\int v(s,t)\,\alpha(t)\,dt &= \int \frac{1}{n-1}\boldsymbol{\phi}^T(s)\,\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{\phi}(t)\,\boldsymbol{\phi}^T(s)\,\boldsymbol{b}\,dt \\ &= \frac{1}{n-1}\boldsymbol{\phi}^T(s)\,\boldsymbol{C}^T\boldsymbol{C}\left[\int \boldsymbol{\phi}(t)\,\boldsymbol{\phi}^T(t)\,dt\right]\boldsymbol{b}\end{aligned}$$

The integral $\int \boldsymbol{\phi}(t)\,\boldsymbol{\phi}^T(t)\,dt$ is a $K \times K$ square matrix $\boldsymbol{A}$ whose entries are:

$$a_{k_1,k_2} = \int \phi_{k_1}(t)\,\phi_{k_2}(t)\,dt$$

The basis system used for this project is the Fourier series which is orthonormal, consequently $\boldsymbol{A}$ is just the identity matrix $\boldsymbol{I}_K$ [110]. Therefore, Equation (4.2) becomes:

$$\frac{1}{n-1}\boldsymbol{\phi}^T(s)\,\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{A}\boldsymbol{b} = \lambda\boldsymbol{\phi}^T(t)\,\boldsymbol{b} \tag{4.4}$$

As Equation (4.4) has to hold for all the values available for s, it therefore reduces to the following:

$$\frac{1}{n-1}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{A}\boldsymbol{b} = \lambda\boldsymbol{b} \tag{4.5}$$

It should be noted that when $\int \alpha^2(t)\,dt = 1$ that

$$\begin{aligned}1 &= \int \alpha^2(t)\,dt \\ &= \int \boldsymbol{b}^T\boldsymbol{\phi}(t)\,\boldsymbol{\phi}^T(t)\,\boldsymbol{b}\,dt \\ &= \boldsymbol{b}^T\boldsymbol{A}\boldsymbol{b}\end{aligned}$$

Likewise, in order for the two functions α_1 and α_2 to be orthogonal, the corresponding vectors of coefficients must satisfy the following:

$$\boldsymbol{b}_1^T\boldsymbol{A}\boldsymbol{b}_2 = 0$$

Usually, the eigenvector in an eigenequation is normalised to ensure a unit length norm, this is achieved by defining $\boldsymbol{u} = \boldsymbol{A}^{\frac{1}{2}}\boldsymbol{b}$. It then follows that $\boldsymbol{u}^T\boldsymbol{u} = 1$ and Equation (4.5) can be rewritten as:

$$\frac{1}{n-1}\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{C}^T\boldsymbol{C}\boldsymbol{A}^{\frac{1}{2}}\boldsymbol{u} = \lambda\boldsymbol{u} \tag{4.6}$$

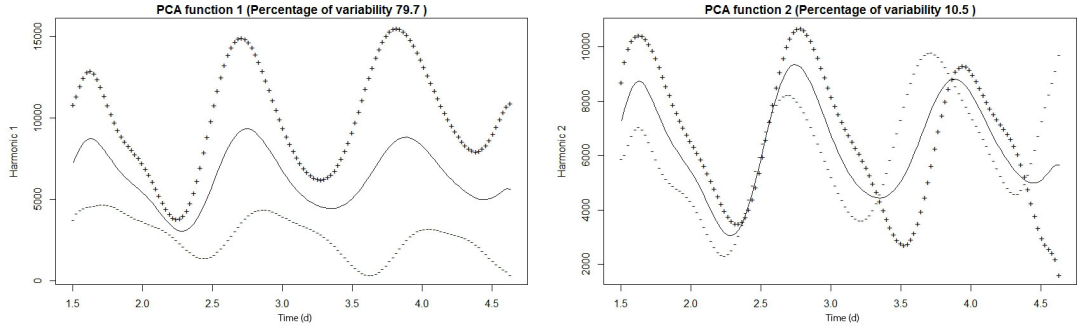
Then, for λ Equation (4.6) is solved and for each eigenvector $\boldsymbol{b} = \boldsymbol{A}^{-\frac{1}{2}}\boldsymbol{u}$ is calculated. As previously explained, this project focuses on examining the particular case where the basis is orthogonal, so $\boldsymbol{A} = \boldsymbol{I}_K$, will mean that $\boldsymbol{b} = \boldsymbol{u}$ will be an eigenvector of $\frac{1}{n-1}\boldsymbol{C}^T\boldsymbol{C}$ [59].

7.2. **Results of FPCA.** FPCA provides additional information concerning both the functions and the data, that may not have been detected using other analytical methods. To help interpret the results of FPCA, perturbations of the mean function for each group is plotted. This is achieved by adding and subtracting a multiple of each principal component function. FPCA was performed on the wild-type and mutant plants free-running in DD conditions using the fda package in R [106].

The number of principal components that should be used can be assessed by examining a scree plot which, as far as multivariate data are concerned, displays the eigenvalues that are associated with a given principal component. This enables the components accounting for the majority of the variability to be identified as the proportion of the total variance accounted for by each principal component is shown [1]. A scree plot works in the same way in the case of FPCA analysis as the eigenvalues are plotted against the principal components and what is known as the elbow of the plot is the point where it flattens out. Each subsequent component will contribute little variability after this point, therefore, the last useful component is often considered to be the component that occurs immediately before the elbow. Nevertheless, this is only considered to be a guideline as there are other effective methods to determine the number of principal components that should be used, such as examining the cumulative variance [37]. In this method, values lying between 75% and 90% are generally thought to account for an acceptable amount of variance [59].

For the unregistered wild-type plants free-running in DD conditions using two principal components accounts for 90.17% of the variability in the data, which is well within the accepted range. One way to visualise the results of FPCA that can help with interpretation, is to examine plots of the mean function of the group and the functions obtained by adding and subtracting a suitable multiple of the principal component functions [110]. Figure 26 provides an example of

this where the solid curve represents the mean function and the dashed (+) or dashed (-) shows the effects of either adding or subtracting a multiple of the principal component of the wild-type unregistered curves subject to DD conditions. This method is a valuable technique as it allows several modes of variation to be viewed together and presents a clearer view of the direct effects of each principal component. It can be seen that the first principal component (Figure 26a) ac-



(A) (B)

FIGURE 26. The mean curves of the unregistered wild-type plants free-running in DD conditions and the effect of adding (+) or subtracting a suitable multiple of each principal components curve. The first principal component (A) accounts for 79.68% and the second principal component (B) accounts for 10.5% of the total variance.

counts for a substantial 79.68% has a positive effect on the mean function which corresponds to a vertical shift in the plot associated with changes in amplitude. The size of the effect of this principal component increases as the experiment goes on, with the largest difference occuring between 3.5 and 4 Days. It can be inferred that this first principal component reflects the high amplitude variation seen within the group, particularly at the end of the functions. It is also apparent that there is a small effect in a horizontal shift where the addition of a multiple of the principal component causes a backwards shift in phase from the mean indicating a faster clock and conversely, the subtraction causes a forward shift in phase from the mean indicating a slower clock.

The second principal component function (Figure 26b) accounts for a further 10.5% and has a more complex effect on the overall behaviour of the curve. Up

until the minimum point of the curve, after 2 Days, the effect is a change in amplitude and the addition of the principal component has a positive effect, between this minimum and the next maximum, between 2.5 and 3 Days, both the addition and subtraction of the principal component does not affect the mean function. After this second peak changes in a time shift are more obvious, that is, the addition of the principal component causes the mean function to express its peaks later, therefore indicating a faster clock. Conversely, the addition of the principal component has a negative effect on the original mean function occurring between the minimum prior to 3.5 Days, the maximum at approximately 4 Days and again at the minimum around 4.5 Days. This is due to the changes in phase between the plants across the experiment where the features of the data do not occur at the same time.

Due to landmark registrations removal of phase variation from the data, it is expected that fewer principal components would have been needed to account for the variation in landmark registered data compared to unregistered data when performing FPCA. For the wild-type landmark registered curves using two principal components would be the most suitable choice and accounts for a total of 94.8% of the variability of the group. As expected, this has accounted for more variability in the same amount of components compared to the unregistered curves.

The first two principal components of the wild-type landmark registered curves free-running in DD conditions are shown in Figure 27a and Figure 27b as perturbations of the group mean function. As previously, a solid line shows the groups mean function whilst the effect on the mean function of adding or subtracting a multiple of the principal component is represented by the (+) and (-) curves. The first principal component function (Figure 27a) accounts for 85.5% of the variability within the group, as expected this is higher than the first principal component of the unregistered curves (79.7%). The effect of the principal component here is simply a vertical shift and is explained just by the changes in amplitude across the genotype. Overall, the registered curves present a more constant variability across the entire experiment compared with the unregistered curves that indicated a large increase in variability towards the end. Again this is due to the effect that phase variation has on the curve overall when it is mixed

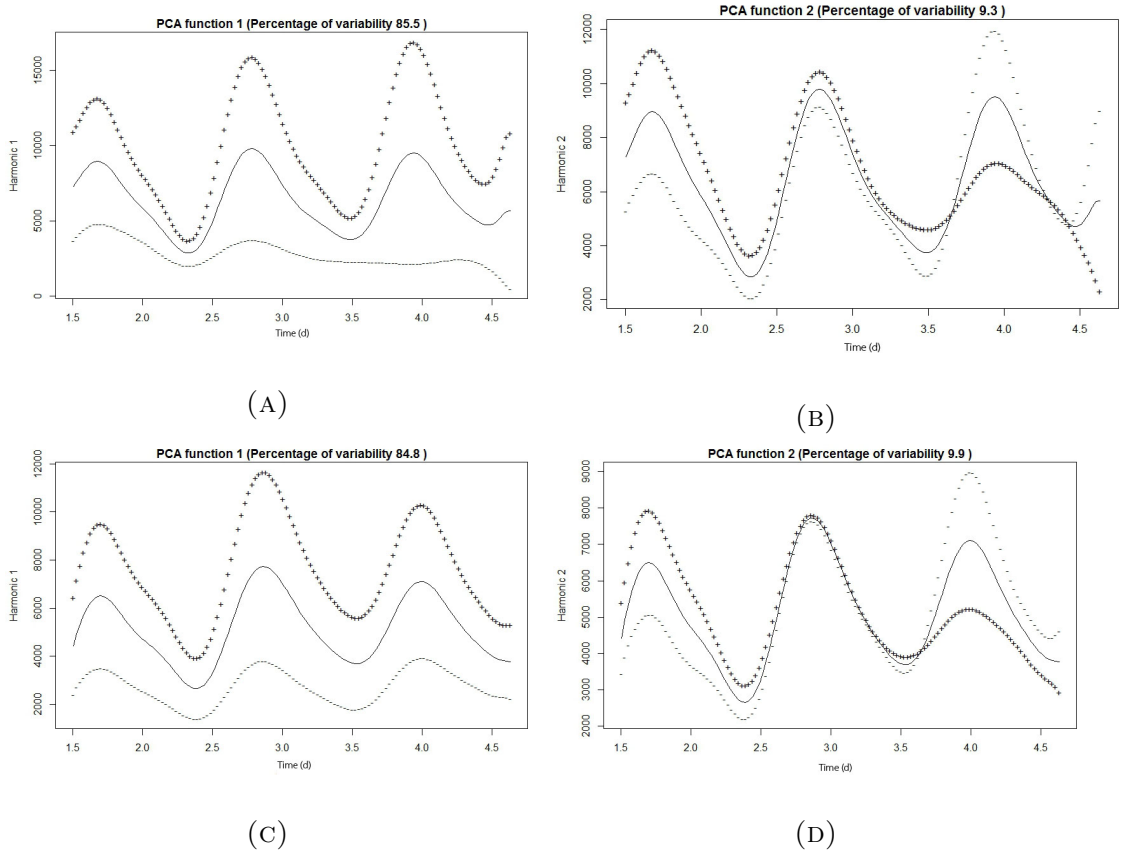


FIGURE 27. The mean curves of the landmark registered wild-type plants (A,B) with the first and second principal components accounting for 85.5% and 9.3% of the total variance and mutant plants (C,D) with the first and second principal components accounting for 84.8% and 9.9% of the total variance. All plants were free-running in DD conditions and the effect of adding (+) or subtracting a suitable multiple of each principal components curve.

with amplitude variation, hence highlighting a further reason for separation of the two types of variation.

The second principal component accounts for 9.3% of the total variation within the group, with most variation shown at the beginning or the end of the experiment and the middle section showing the least. The beginning, from about 1.5 to 2.5 Days, shows the positive effect the addition of principal components has on the mean function in respect of amplitude variation. However, after 3.5 Days, this effect is reversed and the addition of the principal component has a negative effect on the mean function and the subtraction has a positive effect, once again

this is only demonstrated in the form of a downward vertical shift of the variation in amplitude that is seen within the group.

For the landmark registered mutant plants free-running in DD conditions in common with both other groups two principal components are a suitable choice to explain the data. Like the registered wild-type plants more of the variability is explain by these principal components than in the unregistered curves. Using two principal components, accounts for 94.7% of the variability within the group. Figure 27c and Figure 27d show perturbations of the group mean function of the first two principal components of the mutant landmark registered curves free-running in DD conditions. As previously, the mean function of the group is shown by the solid line, whilst the (+) and (-) curves represent the effect that adding or subtracting some multiple of the principal component has on the mean function. Figure 27c shows the first principal component that accounts for 84.81% of the variability within the group and as with the wild-type landmark registered group, there is a fairly constant positive effect of the addition of this principal component. This stems from a vertical shift in the function that occurs due to the amplitude variation seen in the data; this first component shows that slightly less variation is explained compared to the wild-type plants. The second principal component (Figure 27d) also shows very similar behaviour to the wild-type, however, slightly more variation is explained, in this case 9.9%. The majority of the variability is apparent at the beginning and end of the experiment, whereas between approximately 2.5 and 3.5 Days, neither the subtraction nor addition has an effect on the mean function.

Once again, the addition of the principal component has a positive vertical shift effect on the mean function prior to 2.5 Days and after approximately 3.5 Days this is reversed and the addition of the principal component had a negative effect on the mean function. Once more, in common with the wild-type landmark registered curves, there is an increase in variability at the end of the experiment when looking at the second principal component. Viewing Figure 27 as a whole also allows for a direct comparison of mode of variation between the wild-type and mutant plants. As had already been described the first principal component for the wild-type plants holds more variation than the mutant type. However, the distribution of this variation is much more constant in the mutant type, with

the wild-type showing less positive effect when adding the principal component across the minimum points in the mean curve. The second principal components in both types shows very similar behaviour, with the wild-type holding 0.6% less variability within the group than the mutant plants. One area of notable difference is that between 2.5 and 3.5 Days the mutant plants both the addition and subtraction of the principal component to the mean function had virtually no effect. Similarly, both the wild-type and mutant show a switching of effect of adding the principal component after 3.5 Days with this are onwards holding most of the variation for the groups.

Using FPCA provides the opportunity to examine alternative ways to present and explore different types of variability that occur within the data, specifically, comparing mutant and wild-type plants and unregistered and landmark registered curves, brings to light some of the influence these have in affecting the variability within the data. It has been shown that the landmark registered curves hold much more variation in fewer principal components; this is due to the elimination of phase variation, meaning all modes of variation are consequently displayed through amplitude. Additionally, the location of where this variation is located also changes. In both the wild-type and mutant plants, for the first principal component landmark registered curves show that the variation is held evenly across the experiments, whereas the majority of the variation held from 3.5 Days onwards for the unregistered curves. Once more, in respect of the second principal components, all forms of variation in amplitude were apparent for both the landmark registered wild and mutant types, however, this only occurred at the beginning and end of the experiment, this variation reversed with addition of the principal component after 3.5 Days which had the effect of decreasing the mean function. Consequently this presents strong evidence to emphasise how important it is to explore the data thoroughly by using landmark registration to separate the two types of variation. This method allows for easy identification of these types and so making it possible to explore the effect they have on the data so gaining information and furthering understanding of the impact this variation has on behaviour.

One possible way to further explore the results of FPCA is to examine the way in which principal components are interpreted. This can be done by looking at

which of the variables is most strongly correlated with each of the components by plotting the loadings. This would allow for exploration of the relationships between the variables of interest, the importance of each variable in affecting the principal component variability is measured by the loading [1]. As explained, it can be difficult to interpret principal component functions, however, one way to overcome this is by using a rotation method and identifying some more easily interpreted rotated functions, for example a method known as VARIMAX taken from multivariate analysis could be used [29]. The separation of amplitude and phase variation will allow for more in depth analysis of the relationship between these types of variation within groups. For example, accounting for the variation in amplitude would allow for further investigation into phase variation and vice versa. The variation within the velocity and acceleration curves can also be analysed using FPCA [73] and these techniques can also use canonical component analysis, this would specifically look at the correlation between two groups, in order to explore the variation still further [110].

8. QTL Analysis

The genetic map for a RIL set made between Ws wild-type and Col-0 *hsp90.2-3* is shown on Figure 28 where 120 known markers have been placed along the genome. Here the chromosomes are located along the x axis and the position of the markers in centimorgans (cM) along the y axis. This genetic map was termed the W9W population. Figure 38 shows a visual representation of the structure of the population where blue blocks indicate genes from Col-0, yellow from Ws, grey represents either Col-0 or Ws and white indicates no data was available.

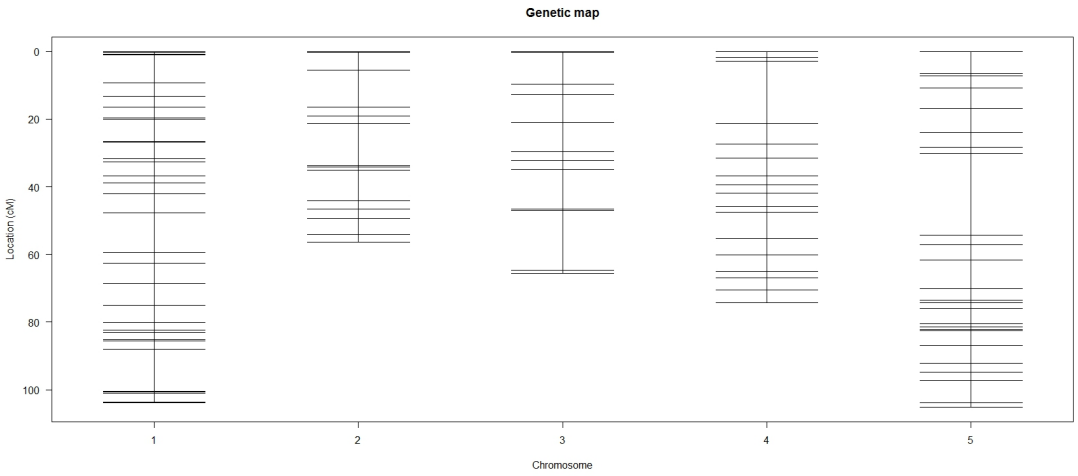


FIGURE 28. The genetic location of markers of W9W population. These data supplied by Amanda M Davis, created by the Davis lab group [34].

8.1. **Phenotyping the W9W population.** In order to perform QTL analysis first a trait that varies between the parents of the population must be measured. Traditionally this is something well defined and easily measured, such as flowering time [70] or hypocotyl elongation [7]. With the combination of the FDA techniques previously discussed, and the ability to QTL map, previously unmeasurable traits can now be measured. Due to the nature of the estimated functions, traits such as velocity and acceleration can be measured and mapped. This can provide a unique insight into the genetic basis for rates of change of the plants rhythms. Not only can these traits now be measured, but also at any given time point during the experiment. Where in typical QTL mapping experiments there would be a single value for a trait across the whole experiment.

After applying the FDA techniques to the provided DD data set, the estimated velocity functions for all genotypes in the population were evaluated every minute across the experiment, so measurements were taken from 37:00:00 to 108:00:00 measurements were calculated. In order to determine that QTL mapping is a suitable form of analysis there needs to be a phenotypic difference between parents [91]. That is the progeny needs to be a segregating population. The natural variation within the population is important for gene detection [69]. For the trait being observed within the population, and in the realms of QTL mapping, the phenotype is assumed to display a normal distribution. An example of this is seen in Figure 29 where a histogram of the phenotype values at 37:03:00, is displayed for the population free-running in DD conditions measuring rate of change as a trait. Here the distribution can be seen to look normal and using a Shapiro-Wilk normality test a p-value of 0.8209 was calculated.

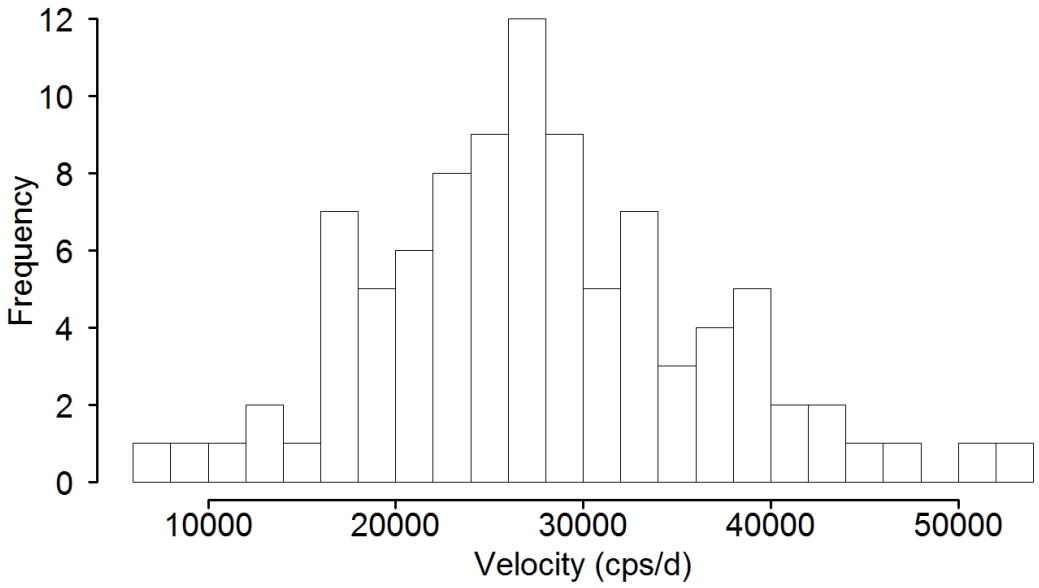


FIGURE 29. Histogram of velocity (cps/d) for all 96 genotypes free-running in DD conditions at 37:03:00.

8.2. **QTL analysis across time.** With the application of FDA techniques applied to the W9W population, traits can be measured at any given time point across the entire experiment allowing real-time gene expression to be monitored. All Figures below are still images from a QTL map video that shows in real time the movement of QTLs across the experiment for the velocity of the population

free-running in DD conditions, in other words looking at the rate of change of the circadian rhythm.

The method of using permutation tests to determine a significance threshold was used. As determining the threshold is dependent on the QTL data presented for any given phenotype, the threshold value is therefore also dependent on time. With use of programming language R, a permutation test was performed at every minute across the experiment, using 1000 permutations and taking the 95% significance threshold. This produced a LOD score ranging between 3.412845 and 11.40345. Chromosome 1 was shown to have significant QTLs along the markers between $21.38 - 26.38$ cM (AtMSQTsnp31; assum. G-w and AtMSQTsnp40; assump. G-w) all occurring between 37:01:00 and 37:58:00 hours, an example of this is shown in the left hand plot in Figure 30. This is of particular interest as this is the time where the plant "missed" its first light cue and is adjusting in free running conditions after expecting it to be in the light. Another significant QTL shows up on chromosome 1 at $40.38 - 41.43$ cM (AtMSQTsnp60), this appears a lot later on in the experiment between 100:49:00 and 104:06:00 and is shown in the right hand plot of Figure 30. Chromosome 2 also showed

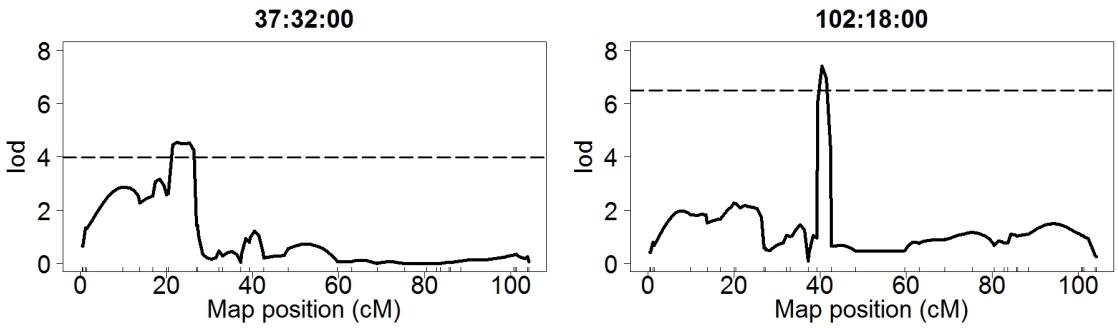


FIGURE 30. The QTL mapping output of chromosome 1 on velocity of the W9W population free-running in DD conditions at time 37:33:00, with 95% confidence threshold at 3.982 and 6.495.

significant QTLs around markers between $9.51 - 13.51cM$ (between LUGSSLP41 and AtMSQTsnp128; assum. G-w), these occur around the time of the first peak (41:00:00-43:00:00) shown in the left hand plot of Figure 31. Again, another QTL is detected between $24.51 - 29.51$ cM (between AtMSQTsnp130 and W9W ii4) later on in the experiment at approximately $93{:}00{:}00 - 95{:}00{:}00$ shown in the right

hand plot of Figure 31. On chromosome 3 some rhythmicity to the QTLs can be

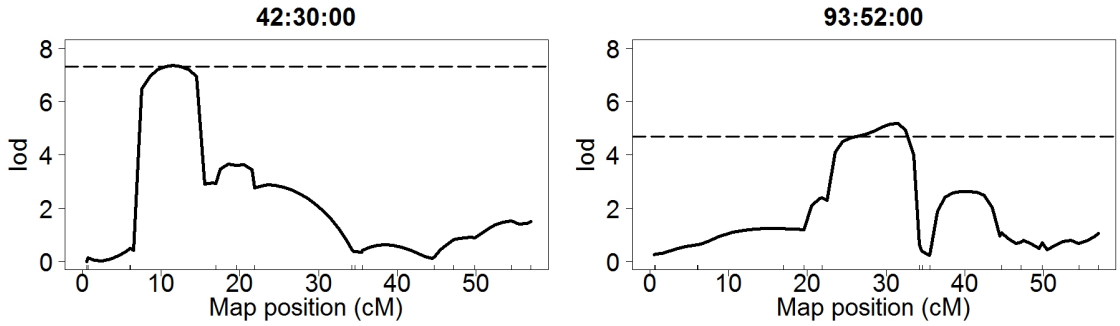


FIGURE 31. The QTL mapping output of chromosome 2 on velocity of the W9W population free-running in DD conditions at time 42:30:00, with 95% confidence threshold at 7.309.

seen, between the markers placed along 50.35 cM to 67.35 cM (between markers W9W iii7 and F27K19). Within this region there are significant QTLs observed between the following approximate times; 37:22:00 − 39:12:00, 62:36:00 − 65:08:00, 88:07:00 − 95:00:00 and 99:30:00 − 105:37:00 these times are all shown in Figure 32 where they pattern in behaviour is clear. Interestingly these times correspond to the general times where the curves are experiencing a peak, it should be noted though the luciferase gene insert is at the top of chromosome 3 which is across the region where these QTLs are being detected, so they could be a result of the luciferase expression being at its highest point.

Chromosome 4 showed the least significant QTLs across the experiment shown in Figure 33, with a QTL being detected early on between 38:53:00 − 38:58:00 so only being significantly present for a few minutes, this is all within the region on the chromosome between 43.29 and 43.47 cM (G3883). The second much stronger QTL was detected along the chromosome between 69.47 and 71.47 cM (AtMSQTsnp306 and AtMSQTsnp310), the significant presence of QTLs here are also occur for a longer amount of time appearing approximately between 41:15:00 and 43:30:00. Chromosome 5 also shows some strong rhythmic activity, this occurs broadly across the area of markers placed between 1.52 and 23.52 cM. More specifically the pattern seen follows a significant QTLs detected on the wider region of 1.5 to 14.52 cM (W9W v1 and nga151a) then shortly after a stronger narrower QTL is detected around the region between 18.52 to 23.53 cM (nga151a

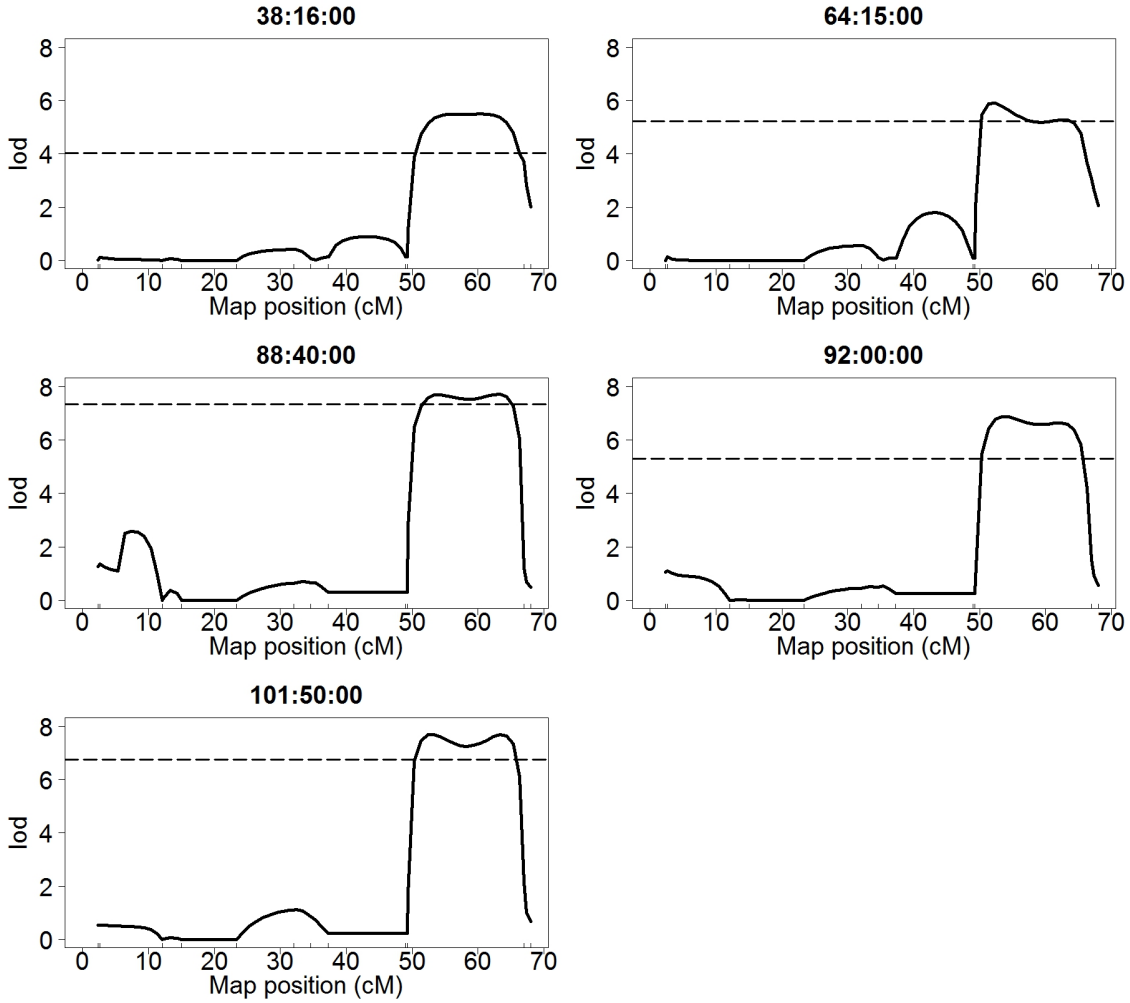


FIGURE 32. Successive QTL mapping outputs of chromosome 3 on velocity of the W9W population free-running in DD conditions at various times, with 95% confidence threshold representative 4.03, 5.24, 7.34, 5.30 and 6.76.

and AtMSQTsnp355), where by the first QTL disappears. Figure 34 shows still shots of the QTL outputs presented from these times and demonstrates clearly the repeated pattern shown. This pattern is seen at intervals across the experiment mainly occurring close to the peaks and troughs observed in the original data. Figure 35 shows all 94 genotypes landmark registered mean curves along with guide lines for the times at which these QTLs appear.

The application of FDA to the data has allowed information to be viewed that otherwise may have been missed. For example, if performing QTL analysis-measuring period as the trait, what is actually being measured is the average

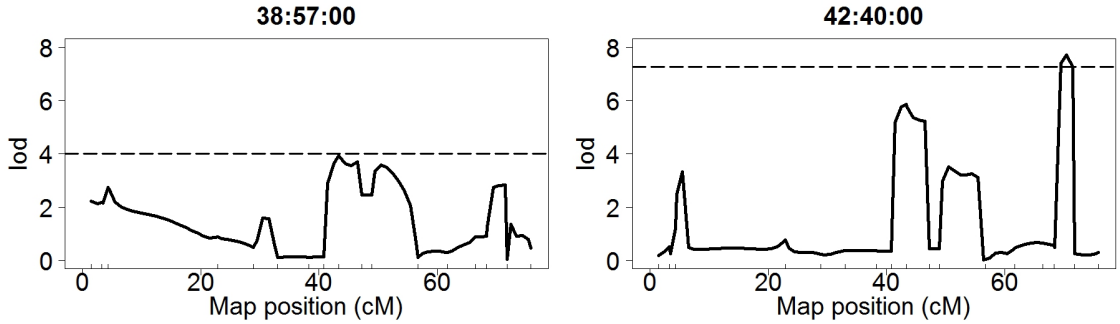


FIGURE 33. Successive QTL mapping outputs of chromosome 4 on velocity of the W9W population free-running in DD conditions at various times, with 95% respective confidence thresholds 4.01 and 7.28.

period across the entire experiment. Another example could be hypocotyl elongation QTL experiments where, the measurement of the hypocotyl is taken once only at a certain time of day and the QTL analysis produced is specific to that one time measurement. Figure 36 shows still shots from the QTL video whereby highlighting that QTLs are dependent on the time of day, the plot of the left shows a QTL present on the top of chromosome 1 at 37:32:00 and the plot on the right shows that at 38:00:00 the QTL is no longer significant and has disappeared. In fact this QTL is present from approximately 37:06:00 till 37:59:00, as only present for under an hour it may not, when averaged over the whole experiment, have reached the threshold value and consequently would not have been noted. Moreover, if the measurements for the QTL analysis were taken out side this window of the QTL being present it would have also been missed. This has highlighted that QTLs are in fact dependent on the time of day the trait is measured, some are even themselves rhythmic and this method provides an option for viewing genetics in a non-static manner. It should be noted that now the ability to see the specific times of day that QTLs appear, means further gene detection work is possible.

8.3. **Future considerations of QTL anaysis.** The analysis shown is from observing the whole population for both wild-type and mutant-types and across the experiment time frame. An alternative to this would be to separate the population into the wild and mutant types as it would actually be expected to see

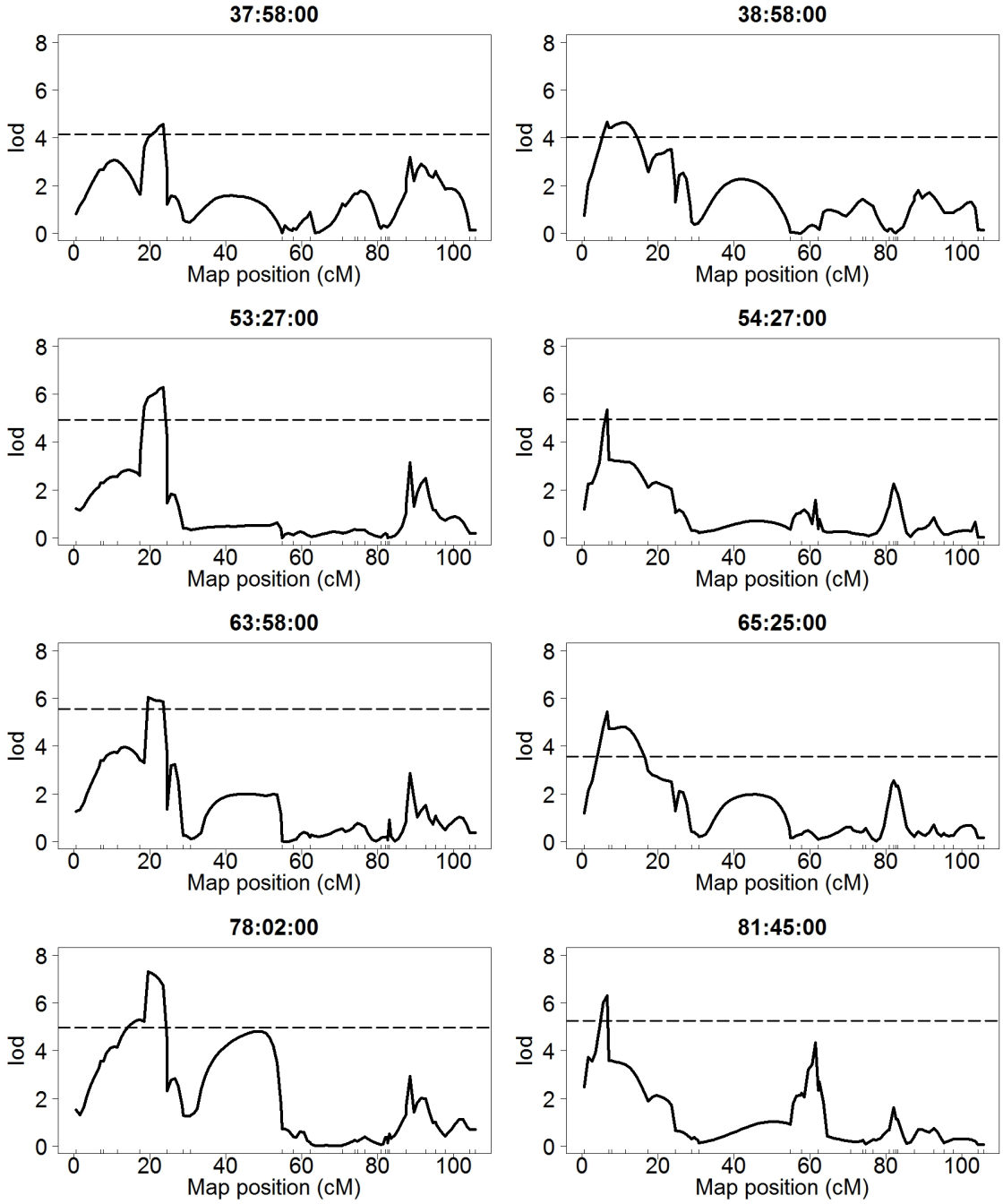


FIGURE 34. Successive QTL mapping outputs of chromosome 5 on velocity of the W9W population free-running in DD conditions at various times, with 95% confidence thresholds.

two separate normal distributions. As an example to this Figure 37 shows the histograms of phenotypes of the population separated into wild-type (left) and mutant type (right). A ShapiroWilk normality test was performed on both sets

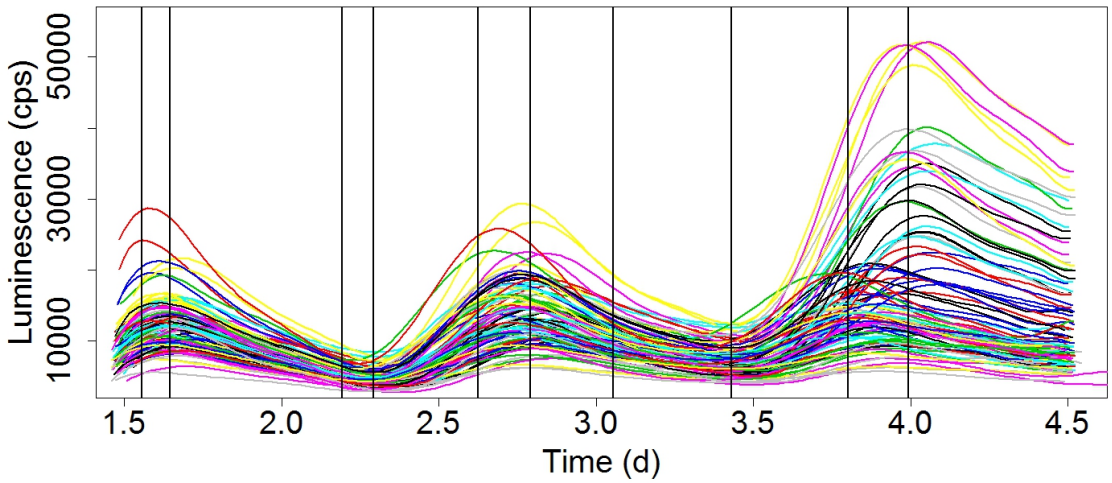


FIGURE 35. Landmark registered mean functions of W9W population free-running in DD conditions. Guide lines indicate the timings of rhythmic QTLs on chromosome 5.

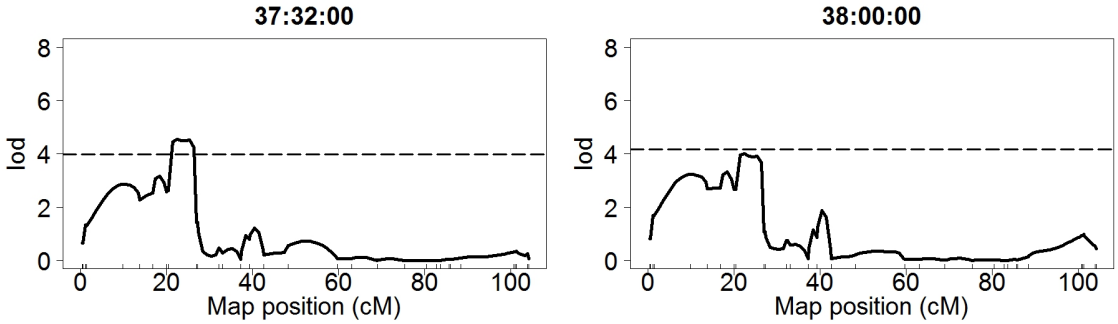
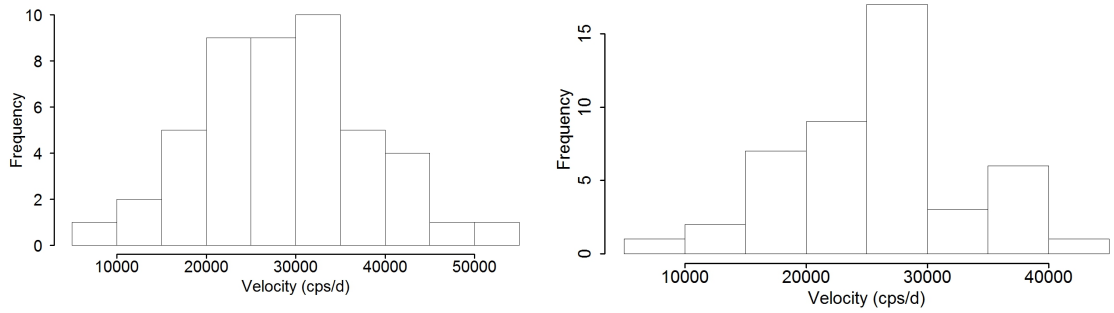


FIGURE 36. QTL map highlighting a QTL on chromosome 1 appearing at 37:32:00 hours and dissappearing at 38:00:00

the wild and mutant type giving p-values of 0.9996 and 0.7428. This was supported as a feature across all time points in the experiment with normality being stronger on average once the wild-type and mutant types are separated. Interestingly the wild-type plants display more typically a stronger normal distribution across the experiment than the mutant plants. This provided further evidence for the *hsp90.2-3* mutant increasing stochastic variation. It can therefore be useful to separate the two plant types to see the direct effect of the mutation on QTL outputs. Though this analysis is beyond the scope of this project, it was opened up a possibility for further exploration.



(A) (B)

FIGURE 37. Histogram of velocity (cps/d) for all genotypes separated as wild-type (A) and mutant-type (B) free-running in DD conditions at 37:03:00.

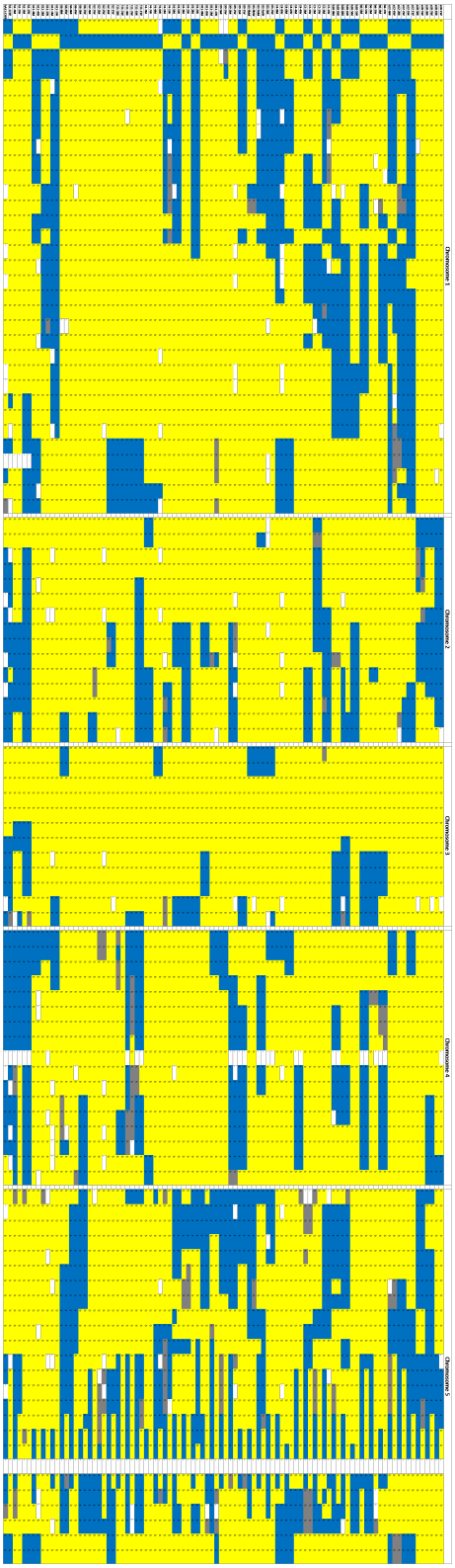


FIGURE 38. Physical map of W9W RIL. Yellow bars represent Ws genes, blue bars represent Col-0 genes, grey bars represent the heterozygous and white bars indicate that no data were available.

9. Conclusion

There is a long history of using mathematical analyses in circadian research in order to increase understanding of mechanisms as an intrinsic rhythmic process and time series analysis is at the forefront of this [111]. Currently, the standard techniques used by the circadian community for time series analysis is the Fourier Transform Non-Linear Least Squares (FFT-NLLS) method [89], [121]. The aim is to analyse circadian data obtained in free-running conditions, without entrainment. Under these free-running conditions, the circadian clock is shown to be nonstationary [48]. However FFT-NLLS works under the assumption that the data are stationary [121]. One way to overcome problem is to develop novel FDA techniques to analyse these non-stationary circadian rhythm data.

Observations of a circadian rhythm (*CCR2*) in Arabidopsis were measured; this was made possible by a luciferase reporter gene that emits luminescence in the presence of substrate luciferin. The RIL plant population used was a BC1F7 generation with 48 homozygous mutant (*hsp90.2-3*) and corresponding 48 wild type (without the mutation at *hsp90.2-3*). There were three data sets used within the project, each subjected the plants to different conditions. They were either set to run in constant darkness, constant light or short day cycles, for each genotype there were 48 replicates per experiment.

A large number of possibilities became apparent when these data were viewed as functions rather discrete points. As a consequence of this, a number of additional details concerning the behaviour of these curves were evident. The transformation of data, from discrete observation to continuous functions, is a critical step and could be approached in a variety of ways, however this project focused on using basis functions and smoothing parameters. The set of 17 known Fourier basis functions were linearly combined to provide an estimated function for the data. Then, in order to maximise use of the functions, a roughness penalty was used to create a smooth differentiable function.

Exploratory analysis of these data began once the discrete data had been fitted into functional form by using the method of basis functions. Initially, through visualising the data in simple plots, it was apparent there were two key areas of variation, amplitude and phase variation. Although strict precautions were taken

to ensure the same conditions for all plants, this cannot be perfectly achievable. Uncontrollable factors could be distribution of media or gradient of light in the growth rooms. Therefore, the variation between genotype replicates does not solely result from genetics, but also from some unknown environmental factors. Although all types of variation are important in the interpretation and analysis of data, due to the nature of the data set and the project aims of investigating the role genes have on the phenotypic behaviour of plants, it was suitable to separate these two types of variation. Landmark registration aims to remove phase variation within a group by monotonically transforming the time domain for each curve so that identifiable features are aligned to a common time argument using time warping functions. For each light condition, the maximum and minimum points were used as landmarks. Using landmark registration and therefore cutting out phase variation has added the benefit of giving a much more accurate representation of a typical curve's behaviour when computing a mean function for each group. This was clearly displayed in Figures 12 and 13 where the registered functions show more clearly defined sharper features compared to the unregistered curves. This is a direct result of the variation in phase effectively stealing from the amplitude when taking an average, so dampening the features.

As the applied FDA techniques created smooth differentiable functions, both first and second derivatives were taken in order to further explore behaviour of the data. The ability to view data as derivative functions played an important role in identifying and exploring particular characteristics of the data. For example, a wild-type plant subject to DD conditions (Figure 16) showed a clear feature of the data in both the velocity and acceleration curves between 1.5 and 2.5 Days, however this feature is not apparent in the original curves. The effect of registering curves is apparent in the phase plane coordinate system where it is shown to draw in both kinetic and potential energy so giving a tighter ellipsis. This shows that landmark registeration provides a more detailed, sharper depiction of the mean curve. Using phase plane plotting it is also shown that mutant-type plants free-running in DD conditions had a longer period (1.164 Days) compared to the wild-type plants (1.101 Days). This is confirmation that *hsp90.2-3* mutation was seen as a period lengthening mutant causing the clock to run slower. This feature noted in Figure 14 was clearly displayed in the phase-plane system in

all experimental conditions. It showed in the phase plane plots as either a mini loop or cusp point occurring at the time the plant would be expected to make a dark to light transition. In the DD data set the plants ability to anticipate dawn deteriorated more quickly than in LL conditions with the second cycle occurring at more of a lag and less prominently so suggesting the clock was more robust under LL conditions. In comparison, an example was shown (Figure 21) under entrainment conditions where the loop is extremely well defined and occurs just prior to 2 Days, as expected, in anticipation of dawn. In summary this demonstrates that the circadian clock works in anticipation of dawn and that this is not what drives the plants rhythms, but rather it is a joint effort from all input pathways to entrain the plants rhythm to a particular cycle.

Variation within the groups is very important, obtaining the correlation functions by using variance and variance-covariance functions allowed for identification of the main types of variability and provided another level of insight into the behaviour of each genotype. The unregistered wild-type plants free-running in DD conditions showed areas of fast separation moving perpendicular to the diagonal at around 2.5 and 3.5 Days. Areas of high positive correlation parallel to the diagonal are representative of the correlation between functions at different times. Highlighting the important role landmark registration plays in analysis of data, the curves showed similar behaviour to the unregistered curves with a strong correlation across the diagonal, but with a much wider area of high correlation when moving away from the diagonal. As expected, the unregistered mutant plants showed a strong correlation across the diagonal and highly positively correlated areas remained when moving perpendicularly outwards. There were two areas of fast separation at around 2.5 and 4 Days, along with dampened separation around 3.5 and 4.5 Days. These areas were all consistent with the maximum and minimum points identified on the original curves. Again, compared to the landmark registered curves, there was an increase to the overall positive correlations across and moving outwards from the diagonal. The areas of fast separation were dampened compared to the unregistered plot, with the only area of relatively fast separation was seen around 4 Days. Using correlation plots to display the data facilitates further comparisons between the wild-type and mutant. Looking at both the landmark registered and unregistered curves, changes between the

wild-type and mutant are clear with a definite positive trend in positive correlation. This suggested that the *hsp.90.3-2* mutant causes the plants to behave more faithfully to one another, that is, it reduces inter-group variation.

FPCA is an extremely effective tool to further develop analysis of the different modes of variation within the groups. For ease of interpretation, perturbations of the mean function for each group are often used and are achieved by adding and subtracting a multiple of each principal component function. Once again, the use of landmark registration is highlighted in FPCA results, when comparing the wild-type plants free-running in DD conditions, the amount of variability accounted for in the first two principal components increases from 90.17% to 94.8%. Also, from plots of perturbations of the means, it is clear that the first principal component of the unregistered curves has a positive effect on the mean function which corresponds to the vertical shift in the plot associated with changes in amplitude. The size of this effect increased as the experiment went on. It can be concluded that this first principal component reflects the high amplitude variation seen within the group, particularly at the end of the experiment. It was also noted that there was a small effect in a horizontal shift where the addition of a multiple of the principal component causes a backwards shift in phase from the mean, so indicating a faster clock. This is mainly results from the mix of phase and amplitude variation seen within the unregistered curves. In comparison, the first principal component of the landmark registered curves accounts for more variation, as well as showing a more even vertical shift and is completely explained by the changes in amplitude across the genotype.

Traditionally a QTL mapping would be performed with one set of data per trait. This trait would be measured once throughout an experiment, (*i.e* hypocotyl length) or as an average across a whole experiment (*i.e* period). The resulting QTL analysis would only provide information for the time point that the measurements were taken, or as an average across a whole experiment, giving no information of how QTLs may be changing across time. The application of FDA techniques that allowed the discrete data to be viewed as continuous functions resulted in the ability to measure, not only an otherwise unmeasurable trait such as velocity (due to data being estimated as smooth functions), but for these traits to be measured at any moment in time across the experiment. Thus giving an insight

into time dependent QTLs. In theory, discrete data could be used to measure a QTL trait across time, but the practicality of time expensive experiments would make it near impossible. FDA applications give the unique opportunity for time dependent QTL traits to be measured in the time domain. In this project rate of change of the circadian rhythm in plants from the DD data set was measured at every minute from the first missed light cue.

The analysis showed that, not only are QTLs dependant on time, but also that some are themselves rhythmic. This was demonstrated on chromosome 3 between the markers placed along 50.35 cM to 67.35 cM (between markers W9W iii7 and F27K19), where throughout the experiment, significant QTLs were found corresponding to times when the original curves would experience a maximum point. This could be due to the Luciferase gene insert, which is at the top of chromosome 3 and so these QTLs could be a result of the luciferase expression being at its highest point. Chromosome 5 also showed strong rhythmic activity between 1.52 and 23.52 cM. The pattern followed the detection of significant QTLs between 1.5 to 14.52 cM (W9W v1 and nga151a), then stronger narrower QTLs were detected between 18.52 to 23.53 cM (nga151a and AtMSQTsnp355) where the first QTL disappears. Looking at the times these took place together with the original functions, showed that the maximum and minimum points correspond. QTLs were also displayed on chromosomes 1, 2 and 4, these did not appear to be rhythmic but are still significant and give reason to further investigate.

This QTL analysis has shown that QTLs depend on the time of day. The methods implemented have allowed for genetics to be viewed in a non-static manner. Overall this has shed new light on the detection of genes and can add to the development of techniques used to gain information concerning what time of day specific genes interact and cause diversity within a population.

9.1. **Further work.** This project focused on applying novel FDA techniques to plant circadian data subjected to different light conditions (DD, LL and SD). This method of analysis provided a unique insight into the exploration of the data sets.

A next step specific to this project would be data collection for the W9W population under 18L:6D (LD) conditions. This would enable a direct comparison

between this and the SD data through means of derivative analysis, correlation analysis, PCA and QTL analysis. This would give further insight to plants behaviour in adjusting to different light-dark cycles. In addition to further explore QTLs that I identified, they would need to be narrow down to smaller regions. To achieve this more genetic markers would be needed in the regions surrounding the QTLs. Once the specific responsible genes are identified, their functions would be further determined in newly created Arabidopsis lines. All of this would allow for more information to be gained about the circadian clock.

The central aim of this project was to investigate the extent to which a circadian rhythm is affected by light-dark cycles, in so doing, it has also provided a solid foundation from which this type of analysis can be used to explore any circadian rhythm. This project has only just begun to examine the wide range of possibilities where the application of functional data analysis could provide new and exciting contributions to this area of research.

Abbreviations

AtGRP7 *ARABIDOPSIS THALIANA* GLYCINE RICH PROTEIN 7

BRASS BIOLOGICAL RHYTHMS ANALYSIS SOFTWARE SYSTEM

CCA1 CIRCADIAN CLOCK-ASSOCIATED 1

CCR2 COLD AND CIRCADIAN REGULATED 2

CV CROSS VALIDATION

DD CONTINUOUS DARKNESS

ELF3 EARLY FLOWERING 3

ELF4 EARLY FLOWERING 4

EM EXPECTATION MAXIMISATION

FDA FUNCTIONAL DATA ANALYSIS

FFT-NLLS FAST FOURIER TRANSFORMATION NONLINEAR LEAST SQUARES

FPCA FUNCTIONAL PRINCIPAL COMPONENT ANALYSIS

GCV GENERALISED CROSS VALIDATION

GI GIGANTEA

IM INTERVAL MAPPING

LD LONG DAY

LL CONTINUOUS LIGHT

LOD LOGARITHM OF ODDS RATIO

LUC LUCIFERASE

LUX LUX ARRHYTHMO

MLE MAXIMUM LIKELIHOOD ESTIMATE

PCA PRINCIPAL COMPONENT ANALYSIS

PRR PSEUDO RESPONSE REGULATOR

QTL QUANTITATIVE TRAIT LOCI

RIL RECOMBINANT INBRED LINE

SD SHORT DAY

TOC1 TIMING OF CAB EXPRESSION 1

Appendix

Appendix 1: Tabel of genotypes and plants removed from the DD data set previous to analysis.

Table 1. Tabel of genotypes and plants removed from the DD data set previous to analysis.

Genotype	Plate reference of plants removed	Genotype	Plate reference of plants removed	Genotype	Plate reference of plants removed
Ws2BC		C4 4W	A7	F4 7W	A4, A6, A7, D5
90.2.3		C4-4M	\	F4-7M	G5
A10-1W	D1	C4-6W	A2, B5	F11-1W	A2, A12, D6
A10-1M	G11	C4-6M	F11, H7, H11	F11-1M	F11, G11
A10-3W	\	C4-8W	A3, A6, A7, A9, B5	F11-3W	\
A10 3M	\	C4 8M	E1, E7, G6, G10	F11 3M	E7, F3, F8, G9, G11, H2
A10-5W	\	D3-3W	A2, A4, A5, A7, A8, A10, A11, B3, B4, B5, B8, C1, C2, C3, C4, C8, C9, C10, C11, C12, D1, D4, D5, D6, D7, D12	F11-5W	A10, B6, B8
A10-5M	\	D3-3M	G8	F11-5M	H2
A12-1W	A1, B9	D3-9aW	A12, D3, D4	G2-1W	B4, C11, D12
A12-1M	E5, G12	D3-9aM	F8, F9, G3, H8, H9, H11	G2-1M	F2, F6, F11
A12-2W	\	D3-9bW	\	G2-2W	B1
A12-2M	\	D3-9bM	\	G2-2M	E1, G8
A12-6W	D7	D9-1W	A4, A8	G2-3W	C5, D3, D6, D7
A12-6M	G9	D9-1M	H8	G2-3M	G2, G3, G10
B6-1W	\	D9-2W	A10, B2, D3, D10	G6-3W	A2, A4, A8, A11, B4, B7, B8, B10, B11, C3, C5-D12
B6-1M	E7, G2, G3, G10, H3	D9-2M	E10	G6-3M	E5, E6, E12, F3, F4, F9, F11, G1, G5, G9, G11, H9, H11
B6-3W	A2	D9-5W	A3, A9, C8	G6-6W	\
B6-3M	G6	D9-5M	G7, G8, H3	G6-6M	F11, G2, H12
B6-5W	C7	D11-1W	A8, B2, B3, C1, C10,	G6-8W	\
B6-5M	E3, E6, E7, F5, F7, F12, G1, G6, H2, H4	D11-1M	E2, E3, E7, F1, F8, G4, G11	G6-8M	E6, F4, H8
B10-5W	\	D11-8W	B9, D9	G11-1W	\
B10-5M	H10	D11-8M	E7, G10	G11-1M	E1, H1, H3, H12
B10-6W	C3	E4-2W	A2, C7, C9, D8	G11-3W	\
B10-6M	E4	E4-2M	E6, E7, E8, E9, F7, G4, H1, H4, H9, H10	G11-3M	E1, G11, G12, H12
B10-8W	D6	E4-3W	A2, A8, A11, C11	G11-4W	A2, A4, D12
B10-8M	F3, G5	F4-3M	G1, G11, H1	G11-4M	F9, F5
C3-1W	A5, C2, C3, C8	E4-5W	\	H1-1W	B9, D1, D12
C3-1M	E12, F4, G3, G6, G9	E4-5M	F4, G8, G12, H8, H9	H1-1M	E9, E10, F3, F4, G1, G11
C3-2W	\	F4-1W	A4, A8, A11, C6, D3	H1-3W	A6, A8, B3, B4, C3, D2, D3, D7
C3-2M	\	F4-1M	E8, G8	H1-3M	E1, E3, E7, E8, E12, F5, G3, G5, G6, G7, G12, H2, H3, H4, H5, H6, H11, H12
C3-3W	A3, A8, B11, C12	F4-4W	D7, D11	H1-11W	\
C3-3M	E8, F2, F3, F5, F8, G3, G4, H8, H10,	F4-4M	\	H1-11M	E3, E8, F8, F9, G9, H6, H9, H10

Appendix 2: Definition of dot product and inner product.

Definition 9.1.1. For $x, y \in \mathbb{R}^n$ the dot product of x, y denoted $x \cdot y$ is defined by [5], [61]:

$$x \cdot y = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n$$

Where $x = (x_1, \ldots, x_n)$ and $y = (y_1, \ldots, y_n)$

Definition 9.1.2. A inner product is a generlization of a dot product. For a real vector space an inner product $\langle \cdot, \cdot \rangle$ satisfies the following properties. Let u,v and w be vectore and α be a scalar then [5], [61]:

- $\langle u + v, w \rangle = \langle u, w \rangle + \langle v, w \rangle$
- $\langle \alpha v, w \rangle = \alpha \langle v, w \rangle$
- $\langle v, w \rangle = \langle w, v \rangle$
- $\langle v, v \rangle \geq 0$ and equal if and only if $v = 0$

Appendix 3: The normal distribution. The probability density of the normal distribution denoted ϕ is as follows [50]:

$$\phi\left(x|\mu, \sigma\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{2}$$

where μ is the expectation, σ is the standard deviation and σ^2 is the variance.

References

[1] Abdi, H. and Williams, L. (2010). Principal component analysis. WIREs Comp Stat, 2(4), pp.433-459.

[2] Adams, R. and Fournier, J. (2003). Sobolev spaces. 2nd ed. Amsterdam: Academc Press, pp.59-77.

[3] Akhiezer, N. and Glazman, I. (1993). Theory of linear operators in Hilbert space. New York: Dover Publications.

[4] Alabadi, D. (2001). Reciprocal Regulation Between TOC1 and LHY/CCA1 Within the Arabidopsis Circadian Clock. Science, 293(5531), pp.880-883.

[5] Allenby, R. (1995). Linear algebra. Amsterdam: Elsevier.

[6] Anwer, M., Boikoglou, E., Herrero, E., Hallstein, M., Davis, A., Velikkakam James, G., Nagy, F. and Davis, S. (2014). Natural variation reveals that intracellular distribution of ELF3 protein is associated with function in the circadian clock. eLife, 3.

[7] Balasubramanian, S., Schwartz, C., Singh, A., Warthmann, N., Kim, M., Maloof, J., Loudet, O., Trainer, G., Dabi, T., Borevitz, J., Chory, J. and Weigel, D. (2009). QTL Mapping in New Arabidopsis thaliana Advanced Intercross-Recombinant Inbred Lines. PLoS ONE, 4(2), p.e4318.

[8] Bartos, B., 2017. Design and Analysis of Time Series Experiments. Oxford University Press.

[9] Bates, D. and DebRoy, S. (2004). Linear mixed models and penalized least squares. Journal of Multivariate Analysis, 91(1), pp.1-17.

[10] Bell-Pedersen, D., Cassone, V., Earnest, D., Golden, S., Hardin, P., Thomas, T. and Zoran, M. (2005). Circadian rhythms from multiple oscillators: lessons from diverse organisms. Nat Rev Genet, 6(7), pp.544-556.

[11] Benhenni, K., Ferraty, F., Rachdi, M. and Vieu, P. (2007). Local smoothing regression with functional data. Computational Statistics, 22(3), pp.353-369.

[12] Bertsekas, D. (1982). Constrained optimization and Lagrange multiplier methods. New York: Academic Press.

[13] Boikoglou, E., Ma, Z., von Korff, M., Davis, A., Nagy, F. and Davis, S. (2011). Environmental Memory from a Circadian Oscillator: The *Arabidopsis thaliana* Clock Differentially Integrates Perception of Photic vs. Thermal Entrainment. Genetics, 189(2), pp.655-664.

[14] Bretzl, H. (1903). Botanische forschungen des Alexanderzuges. Leipzig: B.G. Teubner.

[15] Broman, K., Wu, H., Sen, . and Churchill, G.A. (2003). R/qtl: QTL mapping in experimental crosses. Bioinformatics, 19(7), pp.889-890.

[16] Broman, K. (2005). The Genomes of Recombinant Inbred Lines. Genetics, 169(2), pp.1133-1146.

[17] Broman, K. and Sen, S. (2009). A guide to QTL mapping with R/qtl. 1st ed. Dordrecht: Springer.

[18] Bünning, E. (1973). The physiological clock. London: The English Universities Press.

[19] Castellano, G. and Fanelli, A. (2000). Variable selection using neural-network models. Neurocomputing, 31(1-4), pp.1-13.

[20] Chen, P. and Popovich, P. (2002). Correlation. Thousands Oaks, Calif.: Sage Publications.

[21] Chen, Z. (2013). Statistical Methods for QTL Mapping.

[22] Chui, C., Chan, A. and Liu, S. (1992). An introduction to wavelets. San Diego: Academic Press.

[23] Churchill, G.A. and Doerge, R.W., 1994. Empirical threshold values for quantitative trait mapping. Genetics, 138(3), pp.963-971.

[24] Cohen, J., Cohen, P., West, S. and Aiken, L. (2003). Applied multiple regression/correlation analysis for the behavioral sciences.

[25] Collard, B., Jahufer, M., Brouwer, J. and Pang, E. (2005). An introduction to markers, quantitative trait loci (QTL) mapping and marker-assisted selection for crop improvement: The basic concepts. Euphytica, 142(1-2), pp.169-196.

[26] Covington, M., Panda, S., Liu, X., Strayer, C., Wagner, D. and Kay, S. (2001). ELF3 Modulates Resetting of the Circadian Clock in Arabidopsis. The Plant Cell, 13(6), p.1305.

[27] Craven, P. and Wahba, G. (1978). Smoothing noisy data with spline functions. Numer. Math., 31(4), pp.377-403.

[28] Cubbon, S., Antonio, C., Wilson, J. and Thomas-Oates, J. (2009). Metabolomic applications of HILIC-LC-MS. Mass Spectrom. Rev., 29(5), pp.671-684.

[29] Cureton, E. and Mulaik, S. (1975). The weighted varimax rotation and the promax rotation. Psychometrika, 40(2), pp.183-195.

[30] Daan, S. and Pittendrigh, C. (1976). A Functional analysis of circadian pacemakers in nocturnal rodents. Journal of Comparative Physiology ? A, 106(3), pp.253-266.

[31] Darlington, T. (1998). Closing the Circadian Loop: CLOCK-Induced Transcription of Its Own Inhibitors per and tim. Science, 280(5369), pp.1599-1603.

[32] Darrah, C. (2006). Analysis of Phase of LUCIFERASE Expression Reveals Novel Circadian Quantitative Trait Loci in Arabidopsis. PLANT PHYSIOLOGY, 140(4), pp.1464-1474.

[33] Dierckx, P. (1982). Algorithms for smoothing data with periodic and parametric splines. Computer Graphics and Image Processing, 20(2), pp.171-184.

[34] Dilruba, S. (na). QTL mapping of Flowering Time in a Recombinant Inbred line population of Arabidopsis thaliana harbouring the hsp90 mutation. Masters thesis, University of Bonn.

[35] Dunlap, J. (1999). Molecular Bases for Circadian Clocks. Cell, 96(2), pp.271-290.

[36] Dunlap, J., Loros, J. and DeCoursey, P. (2004). Chronobiology. Sunderland, Mass.: Sinauer Associates.

[37] Eastment, H. and Krzanowski, W. (1982). Cross-Validatory Choice of the Number of Components from a Principal Component Analysis. Technometrics, 24(1), p.73.

[38] Everitt, B. and Hothorn, T. (2011). An introduction to applied multivariate analysis with R. New York: Springer.

[39] Falconer, D. and Mackay, T. (1996). Introduction to quantitative genetics. 1st ed. Essex, England: Longman.

[40] Folland, G. (1992). Fourier analysis and its applications. Pacific Grove, Calif.: Wadsworth and Brooks/Cole Advanced Books and Software.

[41] Gasser, T. and Wang, K. (1997). Alignment of curves by dynamic time warping. The Annals of Statistics, 25(3), pp.1251-1276.

[42] Golub, G. and von Matt, U. (1997). Generalized Cross-Validation for Large-Scale Problems. Journal of Computational and Graphical Statistics, 6(1), p.1.

[43] Hajasz, P. (1996). Sobolev spaces on an arbitrary metric space. Potential Analysis, 5(4), pp.403-415.

[44] Haley, C. S. and Knott, S. A. (1992). A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69. pp.315324.

[45] Halmos, P. (1957). Introduction to Hilbert space and the theory of spectral multiplicity. New York: Chelsea Pub. Co.

[46] Hamilton, J.D. (1994). Time series analysis (Vol. 2). Princeton: Princeton university press, pp.43-72.

[47] Hanano, S., Domagalska, M., Nagy, F. and Davis, S. (2006). Multiple phytohormones influence distinct parameters of the plant circadian clock. Genes to Cells, 11(12), pp.1381-1392.

[48] Hargreaves, J., Knight, M., Pitchford, J. and Davis, S. (2017). Clustering nonstationary circadian rhythms using locally stationary wavelet representations. arXiv :1607.08827

[49] Harmer, S., Hogenesch, J., Straume, M., Chang, H., Han, B., Zhu, T., Wang, X., Kreps, J. and Kay, S. (2000). Orchestrated Transcription of Key Pathways in Arabidopsis by the Circadian Clock. Science, 290(5499), pp.2110-2113.

[50] Hartkemeier, H. (1968). Introduction to applied statistical analysis. Belmont, Calif.: Dickenson Pub. Co.

[51] Hastie, T. and Tibshirani, R. (1990). Generalized additive models. London: Chapman and Hall.

[52] Heintzen, C., Nater, M., Apel, K. and Staiger, D. (1997). AtGRP7, a nuclear RNA-binding protein as a component of a circadian-regulated negative feedback loop in Arabidopsis thaliana. Proceedings of the National Academy of Sciences, 94(16), pp.8515-8520.

[53] Hill, J. (1757). The sleep of plants. London: Printed for R Baldwin.

[54] Hooker, G.(2010). Functional Data Analysis A Short Course. Cornell University.

[55] Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24(6), pp.417-441.

[56] Huang, W., Perez-Garcia, P., Pokhilko, A., Millar, A., Antoshechkin, I., Riechmann, J. and Mas, P. (2012). Mapping the Core of the Arabidopsis Circadian Clock Defines the Network Structure of the Oscillator. Science, 336(6077), pp.75-79.

[57] James, G. (2007). Curve alignment by moments. The Annals of Applied Statistics, 1(2), pp.480-501.

[58] James, G., Witten, D., Hastie, T. and Tibshirani, R. (2013). An Introduction to Statistical Learning. 6th ed. New York: Springer Science+Business Media, pp.29-40.

[59] Jolliffe, I. (2002). Principal component analysis. New York: Springer.

[60] Kaul S, Koo HL, Jenkins J, et al (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408 (6814), p.796-815.

[61] Kaye, R. and Wilson, R. (2003). Linear algebra. Oxford: Oxford Univ. Pr.

[62] Kearsey, M. (1998). The principles of QTL analysis (a minimal mathematics approach). Journal of Experimental Botany, 49(327), pp.1619-1623.

[63] Kendall, M. (1980). Multivariate analysis. 2nd ed. New York: MacMillan, pp.13-31.

[64] Kerr, R., McLachlan, G. and Henshall, J. (2005). Use of the EM algorithm to detect QTL affecting multiple-traits in an across half-sib family analysis. Genetics Selection Evolution, 37(1), pp.83-103.

[65] Kerwin, R., Jimenez-Gomez, J., Fulop, D., Harmer, S., Maloof, J. and Kliebenstein, D. (2011). Network Quantitative Trait Loci Mapping of Circadian Clock Outputs Identifies Metabolic Pathway-to-Clock Linkages inArabidopsis. The Plant Cell, 23(2), pp.471-485.

[66] Kleppner, D. and Kolenkow, R. (2010). An introduction to mechanics. Cambridge, U.K: Cambridge University Press.

[67] Kneip, A. and Gasser, T. (1992). Statistical Tools to Analyze Data Representing a Sample of Curves. The Annals of Statistics, 20(3), pp.1266-1305.

[68] Kneip, A. and Ramsay, J. (2008). Combining Registration and Fitting for Functional Models. Journal of the American Statistical Association, 103(483), pp.1155-1165.

[69] Koornneef, M., Alonso-Blanco, C. and Vreugdenhil, D., 2004. Naturally occurring genetic variation in Arabidopsis thaliana. Annu. Rev. Plant Biol., 55, pp.141-172.

[70] Kowalski, S., Lan, T., Feldmann, K. and Paterson, A. (1994). QTL mapping of naturally-occurring variation in flowering time of Arabidopsis thaliana. MGG Molecular and General Genetics, 245(5).

[71] Krishna, P. and Gloor, G. (2001). The Hsp90 family of proteins in Arabidopsis thaliana. Cell Stress and Chaperones, 6(3), pp.238-246.

[72] Krzanowski, W. and Marriott, F. (1994). Multivariate analysis. London: E. Arnold.

[73] Krzyko, M. and Waszak, . (2013). Canonical correlation analysis for functional data. Biometrical Letters, 50(2).

[74] Lamesch, P., Berardini, T., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D., Garcia-Hernandez, M., Karthikeyan, A., Lee, C., Nelson, W.,

Ploetz, L., Singh, S., Wensel, A. and Huala, E. (2011). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Research, 40(D1), pp.D1202-D1210.

[75] Lay, D. (2006). Linear algebra and its applications. Boston: Pearson/Addison-Wesley.

[76] Levitin, D.J., Nuzzo, R.L., Vines, B.W. and Ramsay, J.O. (2007). Introduction to functional data analysis. Canadian Psychology, 48(3), p.135.

[77] Liu, X. and Mller, H. (2004). Functional Convex Averaging and Synchronization for Time-Warped Random Curves. Journal of the American Statistical Association, 99(467), pp.687-699.

[78] Lock, S. (2016). Functional Data Analysis of the CCR2 Circadian Rhythm in *Arabidopsis thaliana*. BSc. University of York.

[79] Lynch, M. and Walsh, B. (1998). Genetics and analysis of quantitative traits. 1st ed. Sunderland, Mass.: Sinauer.

[80] Mackay, T., Stone, E. and Ayroles, J. (2009). The genetics of quantitative traits: challenges and prospects. Nature Reviews Genetics, 10(8), pp.565-577.

[81] McClung, C. (2006). Plant Circadian Rhythms. The Plant Cell, 18(4), pp.792-803.

[82] McLachlan, G. and Krishnan, T. (2008). The EM algorithm and extensions. Hoboken, N.J: Wiley-Interscience.

[83] Meinke, D.W., Cherry, J.M., Dean, C., Rounsley, S.D. and Koornneef, M., 1998. *Arabidopsis thaliana*: a model plant for genome analysis. Science, 282(5389), pp.662-682.

[84] Meyer, C. (2000). Matrix analysis and applied linear algebra. Philadelphia, PA: Society for Industrial and Applied Mathematics.

[85] Millar, A. (1992). A Novel Circadian Phenotype Based on Firefly Luciferase Expression in Transgenic Plants. THE PLANT CELL ONLINE, 4(9), pp.1075-1087.

[86] Millar, A., Carre, I., Strayer, C., Chua, N. and Kay, S. (1995). Circadian clock mutants in Arabidopsis identified by luciferase imaging. Science, 267(5201), pp.1161-1163.

[87] Millar, A. (1999). Biological clocks in Arabidopsis thaliana. New Phytologist, 141(2), pp.175-197.

[88] Millar, A. (2003). Input signals to the plant circadian clock. Journal of Experimental Botany, 55(395), pp.277-283.

[89] Minors, D. and Waterhouse, J. (1988). Mathematical and statistical analysis of circadian rhythms. Psychoneuroendocrinology, 13(6), pp.443-464.

[90] Murtaugh, P. (2009). Performance of several variable-selection methods applied to real ecological data. Ecology Letters, 12(10), pp.1061-1068.

[91] Nadeau, J.H. and Frankel, W.N., 2000. The roads from phenotypic variation to gene discovery: mutagenesis versus QTLs. Nature genetics, 25(4), p.381.

[92] Neal, R. and Hinton, G. (1998). A View of the EM Algorithm that Justifies Incremental Sparse, and other Variant. Learning in graphical models, 89, pp.355-368.

[93] Nelson, W., Tong, Y.L., Lee, J.K. and Halberg, F., 1979. Methods for cosinor-rhythmometry. Chronobiologia, 6(4), p.305.

[94] Nohales, M. and Kay, S. (2016). Molecular mechanisms at the core of the plant circadian oscillator. Nature Structural and Molecular Biology, 23(12), pp.1061-1069.

[95] Ott, J. (1999). Analysis of human genetic linkage. 3rd ed. Baltimore: Johns Hopkins University Press.

[96] Palmer, J. (1977). Growth rhythms and the history of the earth's rotation. Earth-Science Reviews, 13(3), pp.289-290.

[97] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. Philosophical Magazine Series 6, 2(11), pp.559-572.

[98] Peduzzi, P., Hardy, R. and Holford, T. (1980). A Stepwise Variable Selection Procedure for Nonlinear Regression Models. Biometrics, 36(3), p.511.

[99] Percival, D.B. and Walden, A.T., 2006. Wavelet methods for time series analysis (Vol. 4). Cambridge university press.

[100] Piepho, H.P. (2001). A quick method for computing approximate thresholds for quantitative trait loci detection. Genetics, 157(1), pp.425-432.

[101] Queitsch, C., Sangster, T. and Lindquist, S. (2002). Hsp90 as a capacitor of phenotypic variation. Nature, 417(6889), pp.618-624.

[102] R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: https://www.R-project.org/.

[103] Rajpal, V., Rao, S. and Raina, S. (2016). Molecular breeding for sustainable crop improvement. 2nd ed. Switzerland: Springer, pp.31-59.

[104] Ramsay, J. and Dalzell, C.J. (1991). Some tools for functional data analysis. Journal of the Royal Statistical Society. Series B (Methodological), pp.539-572.

[105] Ramsay , J. and Silverman, B. (1997). Functional data analysis. New York: Springer.

[106] Ramsay, J., Wickham, H., Graves, S. and Hooker, G. (2014). fda: Functional Data Analysis. R package version 2.4.4. https://CRAN.R-project.org/package=

[107] Ramsay, J., Hooker, G. and Graves, S. (2009). Functional data analysis with R and MATLAB. New York: Springer.

[108] Ramsay, J. and Li, X. (1998). Curve registration. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(2), pp.351-363.

[109] Ramsay, J. (1998). Estimating smooth monotone functions. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 60(2), pp.365-375.

[110] Ramsey, J. and Silverman, B. (2005). Functional Data Analysis. 2nd ed. New York: Springer Science+Business Media, Inc.

[111] Refinetti, R., Cornlissen, G. and Halberg, F. (2007). Procedures for numerical analysis of circadian rhythms. Biological Rhythm Research, 38(4), pp.275-325.

[112] Refinetti, R. (2004). Non-stationary time series and the robustness of circadian rhythms. Journal of Theoretical Biology, 227(4), pp.571-581.

[113] Sangster, T., Bahrami, A., Wilczek, A., Watanabe, E., Schellenberg, K., McLellan, C., Kelley, A., Kong, S., Queitsch, C. and Lindquist, S. (2007). Phenotypic Diversity and Altered Environmental Plasticity in Arabidopsis thaliana with Reduced Hsp90 Levels. PLoS ONE, 2(7), p.e648.

[114] Schning, J., Streitner, C., Page, D., Hennig, S., Uchida, K., Wolf, E., Furuya, M. and Staiger, D. (2007). Auto-regulation of the circadian slave oscillator component AtGRP7 and regulation of its targets is impaired by a single RNA recognition motif point mutation. The Plant Journal, 52(6), pp.1119-1130.

[115] Shao, J. (1993). Linear Model Selection by Cross-validation. Journal of the American Statistical Association, 88(422), pp.486-494.

[116] Shapiro, J. (2007). Composition operators and classical function theory. New York: Springer.

[117] Somers, D. (1999). The Physiology and Molecular Bases of the Plant Circadian Clock. Plant Physiology, 121(1), pp.9-20.

[118] Staiger, D., Shin, J., Johansson, M. and Davis, S. (2013). The circadian clock goes genomic. Genome Biology, 14(6).

[119] Silverman, B. (1996). Smoothed functional principal components analysis by choice of norm. The Annals of Statistics, 24(1), pp.1-24.

[120] Strang, G. (1988). Linear algebra and its applications. San Diego: Harcourt, Brace, Jovanovich, Publishers.

[121] Straume M, Frasier-Cadoret SG, Johnson ML (1991) Least Squares Analysis of Fluorescence Data. In: Lakowicz JR editor. Topics in Fluorescence Spectroscopy, Volume 2: Plenum. pp 117240.

[122] Strayer, C. (2000). Cloning of the Arabidopsis Clock Gene TOC1, an Autoregulatory Response Regulator Homolog. Science, 289(5480), pp.768-771.

[123] The Arabidopsis Genome Initiative, (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature, 408(6814), pp.796-815.

[124] Thomas, G., Weir, M. and Hass, J. (2013). Thomas' Calculus: Early Transcendentals. 13th ed. New York: Pearson, pp.123-295.

[125] Thorne, N., Inglese, J. and Auld, D. (2010). Illuminating Insights into Firefly Luciferase and Other Bioluminescent Reporters Used in Chemical Biology. Chemistry and Biology, 17(6), pp.646-657.

[126] Vitaterna, M.H., Takahashi, J.S. and Turek, F.W., 2001. Overview of circadian rhythms. Alcohol Research and Health, 25(2), pp.85-93.

[127] Viviani, R., Grn, G. and Spitzer, M. (2004). Functional principal component analysis of fMRI data. Human Brain Mapping, 24(2), pp.109-129.

[128] Wang, W., Vinocur, B., Shoseyov, O. and Altman, A., 2004. Role of plant heat-shock proteins and molecular chaperones in the abiotic stress response. Trends in plant science, 9(5), pp.244-252.

[129] Wei, Y., Zhang, N., Ng, M. and Xu, W. (2007). Tikhonov regularization for weighted total least squares problems. Applied Mathematics Letters, 20(1), pp.82-87.

[130] Wilson, J. (2015). Statistical Pattern Recognition. MAT00031H. University of York, Department of Mathematics [Accessed March 2016]

[131] Wood, S. (2000). Modelling and smoothing parameter estimation with multiple quadratic penalties. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 62(2), pp.413-428.

[132] Wu, R., Ma, C. and Casella, G. (2007). Statistical Genetics of Quantitative Traits; Linkage, Maps and QTL. New York: Springer.

[133] Xu, Z., Li, Z., Chen, Y., Chen, M., Li, L. and Ma, Y. (2012). Heat Shock Protein 90 in Plants: Molecular Mechanisms and Roles in Stress Responses. International Journal of Molecular Sciences, 13(12), pp.15706-15723.

[134] Yan, H. (2016). Multivariate Analysis. MAT00021H. University of York, Department of Mathematics

[135] Young, N. (1988). An introduction to Hilbert space. Cambridge: Cambridge University Press.

[136] Zielinski, T., Moore, A., Troup, E., Halliday, K. and Millar, A. (2014). Strengths and Limitations of Period Estimation Methods for Circadian Data. PLoS ONE, 9(5), p.e96462.