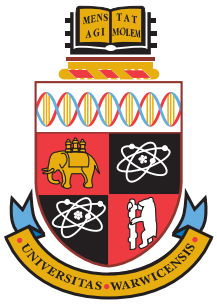


Extracting Circadian Clock Information From A Single Time Point Assay

by

Denise F. Vlachou

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Supervisors: Professor David A. Rand and Professor Francis A. Lévi

MOAC Doctoral Training Centre

April 2018

Abstract

A working internal circadian clock allows a healthy organism to keep time in order to anticipate transitions between night and day, allowing the temporal optimisation and control of internal processes. The internal circadian clock is regulated by a set of core genes that form a tightly coupled oscillator system. These oscillators are autonomous and robust to noise, but can be slowly reset by external signals that are processed by the master clock in the brain.

In this thesis we explore the robustness of a tightly coupled oscillator model of the circadian clock, and show that its deterministic and stochastic forms are both significantly robust to noise. Using a simple linear algebra approach to rhythmicity detection, we show that a small set of circadian clock genes are rhythmic and synchronised in mouse tissues, and rhythmic and synchronised in a group of human individuals. These sets of tightly regulated, robust oscillators, are genes that we use to define the expected behaviour of a healthy circadian clock. We use these "time fingerprints" to design a model, dubbed "Time-Teller", that can be used to tell the time from single time point samples of mouse or human transcriptome.

The dysfunction of the molecular circadian clock is implicated in several major diseases and there is significant evidence that disrupted circadian rhythm is a hallmark of many cancers. Convincing results showing the dysfunction of the circadian clock in solid tumours is lacking due to the difficulties of studying circadian rhythms in tumours within living mammals. Instead of developing biological assays to study this, we take advantage of the design of Time-Teller, using its underlying features to build a metric, Θ, that indicates dysfunction of the circadian clock. We use Time-Teller to explore the clock function of samples from existing, publicly available tumour transcriptome data.

Although multiple algorithms have been published with the aims of "time-telling" using transcriptome data, none of them have been reported to be able to tell the times of single samples, or provide metrics of clock dysfunction in single samples. Time-Teller is presented in this thesis as an algorithm that both tells the time of a single time-point sample, and provides a measure of clock function for that sample.

In a case study, we use the clock function metric, Θ, as a retrospective prognostic marker for breast cancer using data from a completed clinical trial. Θ is shown to correlate with many prognostic markers of breast cancer, and we show how Θ could also be a predictive marker for treatment efficacy and patient survival.

Acknowledgements

Firstly, I would like to thank my supervisor Prof David Rand for his incredible guidance throughout this PhD project. I would like to thank him for his generous time in reading all of my early drafts, and how he encouraged me to produce the highest standards of work that I could achieve.

It has been a great experience to work with my supervisor Prof. Francis Lévi, who is a pioneer in the field of chronobiology, and whose ideas formed the motivation for the basis of this thesis. I would like to thank Francis and the rest of the chronobiology group at Warwick, for accepting me as the mathematician outsider in their lab group, which has resulted in me developing the cross-discipline working skills which have formed the basis of my scientific career. Additionally, I'd like to thank Dr Giorgos Minas for his help with using his pcLNA algorithm code, and Dr Sylvie Giacchetti for her guidance and contribution of ideas for the REMAGUS dataset.

The methods in this thesis were developed using the gene atlas data published by Zhang *et al.* I would like acknowledge and thank the group that produced this incredibly valuable data set. The human time course data used in this thesis was shared with us by Dr Georg Bjarnason, in a generous and fruitful collaboration. This incredible and unique dataset has made all of the human results in this thesis possible.

I'd like to thank my family, friends, and GSK colleagues for supporting me over the past few months whilst I have been working and writing, especially Maria, Ziedo, and Elena for proof reading.

I would like to thank everyone who was part of my time at MOAC; my fellow students and friends, Naomi, Christina, and Nikola, and everyone in the systems biology PhD office. Finally, I'd like to thank Hugo, who has been my teacher, mentor, friend, and primary source of distraction for the past 5 years.

This work was funded by the EPSRC through the MOAC doctoral training centre.

Declaration

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy and is presented in accordance with the specified regulations. It has been composed by myself and has not been submitted in any previous application for any degree.

Contents

Abbreviations

Affy	Affymetrix (microarray)
AU	Arbitrary units
CCN	Core clock network
CT	Circadian time (free running)
E-box	Evening box (promoter motif)
ECCN	Extended core clock network
EFS	Disease free survival
ER+/-	Estrogen receptor positive/negative
D-box	Day time box (promoter motif)
DFS	Disease Free Survival
FDR	False discovery rate
FIM	Fisher information matrix
GEO	Gene expression omnibus (`https://www.ncbi.nlm.nih.gov/geo/`)
GUI	Graphical user interface
HER+/-	Herceptin positive/negative
LD	12 hr Light-Dark cycle conditions (a.k.a. entrained)
LL	Constant light conditions. (a.k.a. free running)
KD	Knock-down
KO	Knock-out
LNA	Linear noise approximation
lPC	Local principal component
MLE	Maximum likelihood estimate
MTA	Material transfer agreement
MTTM	Molecular time table method
ODE	Ordinary differential equation
OS	Overall survival
PC	Principal component
PCA	Principal component analysis
pcLNA	phase corrected linear noise approximation
pCR	Pathological complete response
PLSR	Partial least squares regression
PR+/-	Progesterone receptor positive/negative
RRE	Ror and Rev-Erb element (promoter motif)
SPC	Sparse principal component
SVD	Singular value decomposition
TN	Triple negative (tumour)
WT	Wild Type
ZT	Zeitgeber time (entrained with light-dark cycles)

List of Figures

List of Tables

Motivation

The methods in this PhD project were developed towards the aim of a single task. The task was to determine whether the circadian clock gene expression was different to "healthy" circadian gene expression in a single transcriptome from a tumour biopsy. This thesis contains the work done to establish what "normal" circadian clock gene expression is, why and how we expect a determined behaviour for different individuals and tissues, and how we can measure if a single biopsy sample has a "healthy" clock.

Notes to reader

This thesis will present novel models and methods of data analysis. All data used in this thesis is publicly available and is cited using the original studies and GEO accession numbers. The only exception to this is the Bjarnason human timecourse data which was shared under an MTA agreement with Dr Georg Bjarnason, of Sunnybrook Research Institute, Toronto, Canada. This data is unpublished at the time of writing of this thesis.

This thesis is an interdisciplinary study using mathematics and statistics to understand complex biological and clinical data. It is assumed that the reader has a good grounding in mathematics, statistics, and basic genetics.

These works are organised so that each chapter builds on the results and conclusions of each previous chapter. We begin with an introduction to circadian rhythms, and each subsequent chapter will contain its own introduction, and relevant literature will be reviewed in the appropriate chapters.

Chapter 1

Introduction to the Circadian Clock

Internal oscillators have been observed and hypothesised for hundreds of years. Jean-Jacques d'Ortous de Mairan is known to be one of the first chronobiologists. In 1779, he noticed that mimosa plants would move their leaves in a 24 hour rhythm, in synchrony with the light. He then observed that the plant continued this behaviour even in constant darkness, indicating some sort of internal timing mechanism that was not driven by a reaction to light [25].

The term "circadian" is derived from the Greek "circa" (approximately) and "dian" (day), and has been used in scientific literature for the study of daily rhythms since its coining in the 1950s [26]. The "circadian clock" is now used as a collective term to describe the mechanisms that have evolved so that an organism can anticipate and synchronise to, the light-dark cycles that result from the 24-hour rotation of the Earth [27]. The field of circadian rhythms has now grown to the point of huge scientific and even public interest.

1.1 Popular perception of the circadian clock

The terms *circadian rhythm* or *circadian clock* are generally understood to concern the sleep-wake cycle and routine of an individual as these physiological and behavioural events can be observationally tracked. People often label themselves as "night owls" or "larks" depending on their body's natural preference for times of sleep and optimal performance. People are likely to understand their body's natural rhythms and plan their daily activities to get the most out of their day. The harnessing of one's natural circadian rhythms in order to improve one's health and quality of life is becoming increasingly popular. For example, many studies reported in the media concern the association between timings of food intake and weight gain or diabetes[1]. An internet search on the effects of meal timings will bring up countless articles claiming, for example, that eating in the evenings is detrimental to weight loss [28, 29, 30].

[1] A good example can be found at http://www.bbc.co.uk/news/health-27161671

Another popular way that people attempt to take control of their internal circadian rhythm, is to limit exposure to light in the evenings. In a healthy individual, melatonin secretion occurs late in the evening to induce sleep [31]. Blue light is known to suppress melatonin [32] and LEDs used in some light bulbs, TVs, computers, smart phones, and tablets, have peak emission in the blue light range. This has led to studies assessing what our modern lifestyles are doing to melatonin rhythms [32]. As a result of these studies, modern phones and computer screens are capable of going into "night mode", where they effectively filter out blue light in the evening hours.

Chronic disruption of normal rhythms of melatonin are attributed to psychiatric diseases such as sleep disorders or depression [33, 34, 35]. Melatonin supplements and daytime blue light therapies are possible treatments for these disorders, which attempt to reinstate the proper rhythms of melatonin. Furthermore, the use of melatonin tablets by the general public as sleeping tablets has become fashionable and its use is increasing [36].

The circadian clock has a role in nearly all major aspects of physiology, including sleep, metabolism, hormone secretions, body temperature, blood pressure, excretions, immune function, healing, and more [37, 38]. A representation of the organisation of some of the physiological and behavioural circadian rhythms are summarised by figure 1.1. This figure shows that there are times throughout the day where certain physiological factors are at their highest and lowest, where possible timings for highs and lows are shown for body temperature, blood pressure, alertness, sleep, melatonin, etc. Note that the timings on this figure are an example of a hypothetical daily rhythm, and not a strict population representation.

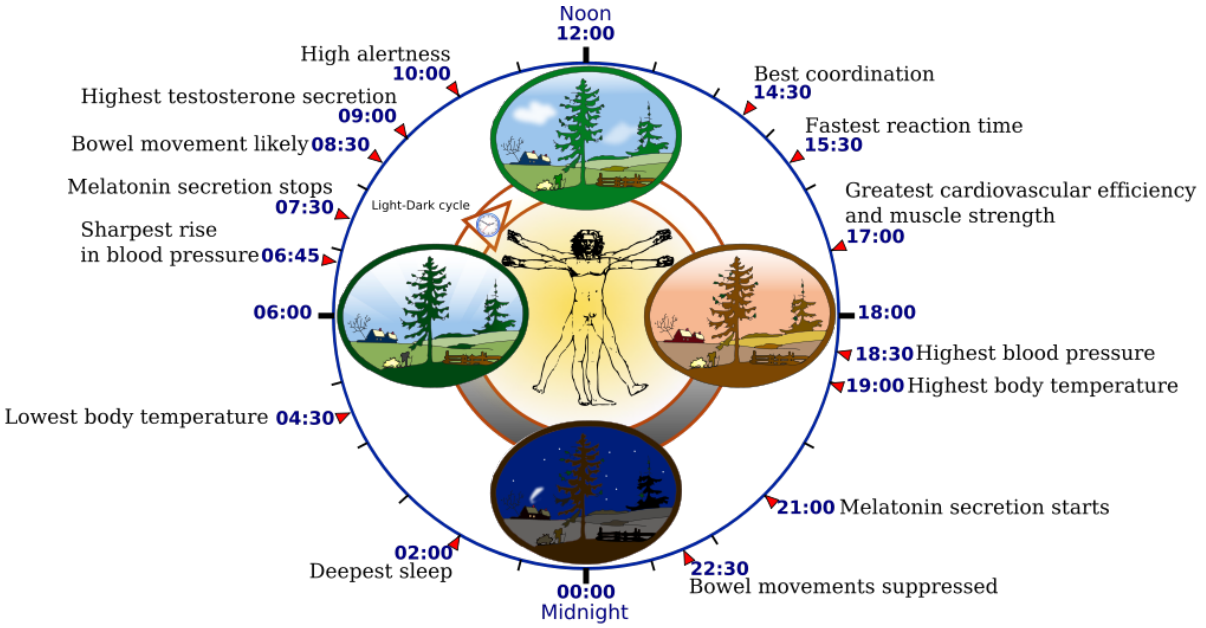


Figure 1.1: **Sketch representing possible timings for the biological circadian clock.** From [1].

A lesser known aspect of the circadian clock is an underlying genetic component, and it is considerably more difficult to observe and understand it. This thesis will concern the genetic component of the circadian clock, which is known as the "molecular circadian clock".

1.2 The molecular circadian clock

Internal molecular oscillators were first observed in *Drosophila* in the 70s, and were shown to be a certain group of genes that are periodically expressed over 24 hours [39]. In 2017, the Nobel prize in Physiology or Medicine was awarded to Jeffrey C. Hall, Michael Rosbash and Michael W.Young for their discoveries of the molecular mechanisms controlling the circadian clock in *Drosophila* [40].

There exists a molecular circadian clock of 24-hour periodically oscillating genes and proteins inside (almost) every cell of our bodies, and inside the cells of most living organisms on the earth [41]. The study of molecular rhythms is an increasingly popular and promising field of research. This is because these 24-hour rhythmic genes and proteins have been shown to regulate many biological processes such as the cell cycle, metabolism, apoptosis, immune function, and wound healing [37, 38]. As a result, the dysfunction of the molecular circadian clock is associated with many diseases, just as the dysfunction of sleep and hormone patterns are.

The study of the connection between the circadian clock and other cellular processes promises to answer questions on diseases and open up new opportunities for treatments. As mentioned above, a sleep disorder can be treated with melatonin supplements and blue light to artificially re-instate the hormone rhythms and hopefully reinstate the "normal" physiological rhythms [42]. Using this same concept, we can try to treat disease by re-instating molecular rhythms, or taking advantage of them in some way to optimise treatments. The practise of dosing drugs based on time of day is termed *Chronotherapy*, and has been practised by Professor Francis Lévi and his teams for the past 20 years, to use optimum time of day to dose chemotherapy to cancer patients [43]. This will be further discussed in Chapter 6.

In order to exploit molecular circadian rhythms so that we can develop new treatments for diseases, we first must understand how the *functioning* circadian clock works. However, the molecular circadian clock is highly complex, so requires novel and multiplex experimental and analytical approaches to decode its dynamics. Mathematical modelling has been a very useful tool to explore the database of knowledge that biological experiments has built. This will be discussed in chapter 2.

When we refer to the molecular circadian clock, we refer to a collection of mRNA transcripts that oscillate in abundance over 24 hours. The molecular circadian clock

is a complex gene network of multiple feedback loops involving transcription processes, post-translational modifications of proteins, protein-protein interactions, chromatin modification, and more [44]. Most of the core network of cycling genes are translated into transcription factors that regulate and control the transcription of many output genes. While a majority of molecular circadian research has focused on transcriptional mechanisms, translational and post translational mechanisms also have significant roles in circadian regulation. For example, there is evidence that there are some constitutively expressed genes that have cycling proteins [45]. Due to the lack of existing datasets and generally less understanding of circadian proteins this thesis focuses only on mRNA rhythms. It is possible that similar methods will be applicable to datasets containing information on timecourses of protein abundances, when future technologies are better able to quantify proteins.

1.3 The role of the circadian clock

Molecular circadian clocks are a feature of the internal workings of all kingdoms of life [9]. The molecular components of the circadian clock between different kingdoms are not conserved, suggesting that the clocks must have evolved independently [27]. A summary of some of the most studied circadian clocks and their evolutionary relationships is shown in figure 1.2. Human and murine clocks are the most studied mammalian clocks and have very similar components. *Drosophila* and zebrafish share more than 70-75% of their genes with humans and have clocks that have some similar components to the mammalian clock, such as Pers and Crys. *Arabidopsis thaliana* is a plant that has been extensively studied so its well documented genome makes it a very good model organism in which to study circadian clocks. Although the plant clock genes are completely different to the mammalian clock genes, the architecture is very similar. Consistent architectural features of each kingdom's molecular circadian clock are complex auto-regulatory networks, multiple feedback loops, and genetic redundancies [46]. The fact that independent clocks have evolved with similar architectures across kingdoms suggests that a robust and functional internal circadian rhythm is a non-optional requirement of successfully surviving on the Earth (this will be further discussed in chapter 2). It follows that a dysfunctional circadian clock has been associated with many pathologies such as cardiovascular and inflammatory diseases, depression, and cancer. Circadian rhythm disruption in cancer, and more specifically breast cancer, will be the focus of chapter 6.

Definition of a circadian rhythm

Circadian rhythms have an approximately 24 hour period, are endogenous, self-sustained and persisting in the absence of any environmental cues (such as the light/dark cycle), and

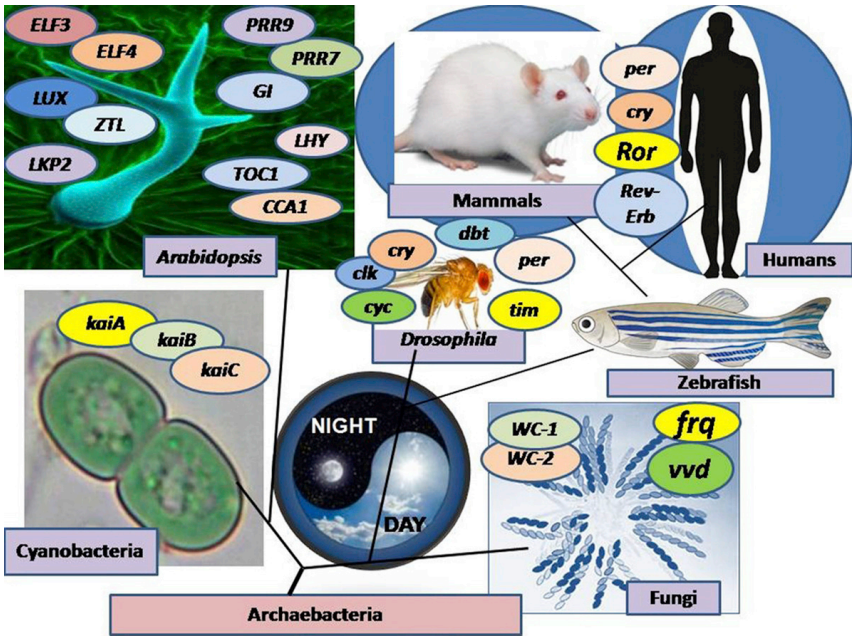


Figure 1.2: **Model organisms most commonly used in the study of circadian clocks**. From [2], figure shows different sets of clock genes across kingdoms.

are entrainable to periodic forcing such as forcing from light, temperature [47]. Zeitgeber time (ZT) describes the number of hours after lights have come on (for example the number of hours since sunrise). Circadian time (CT) describes the number of hours of free-running time where no forcing is present (for example the number of hours since the last time conditions changed).

1.4 The mammalian molecular circadian clock

Many review articles have been published with the aim of summarising current knowledge of the mammalian molecular circadian clock [48, 49, 50, 6, 41, 51, 4]. The genes that make up the core mechanism of the mammalian circadian clock are not easy to define. The subset of circadian genes that are core to the autonomous circadian oscillator changes depending on the criteria used to define what "core" is. Consequently, every study reports a slightly different set of core clock genes. However, through years of study and collation, around 20 genes have been repeatedly observed to be circadian in mammals [52]. Studies in bioinformatics that utilise data mining techniques have attempted to collate this vast array of information in order to determine what the true core clock genes are [53, 3]. An example of this is shown in figure 1.3, where Lehmann *et al.* [3] studied an extensive collection of databases and literature. Using these results, they found 14 genes that they call the core clock network (CCN) and 28 genes that they call the extended core clock network (ECCN). The red boxed genes in figure 1.3 are those that (we find) the literature more frequently reports as those closely associated with the circadian clock [4].

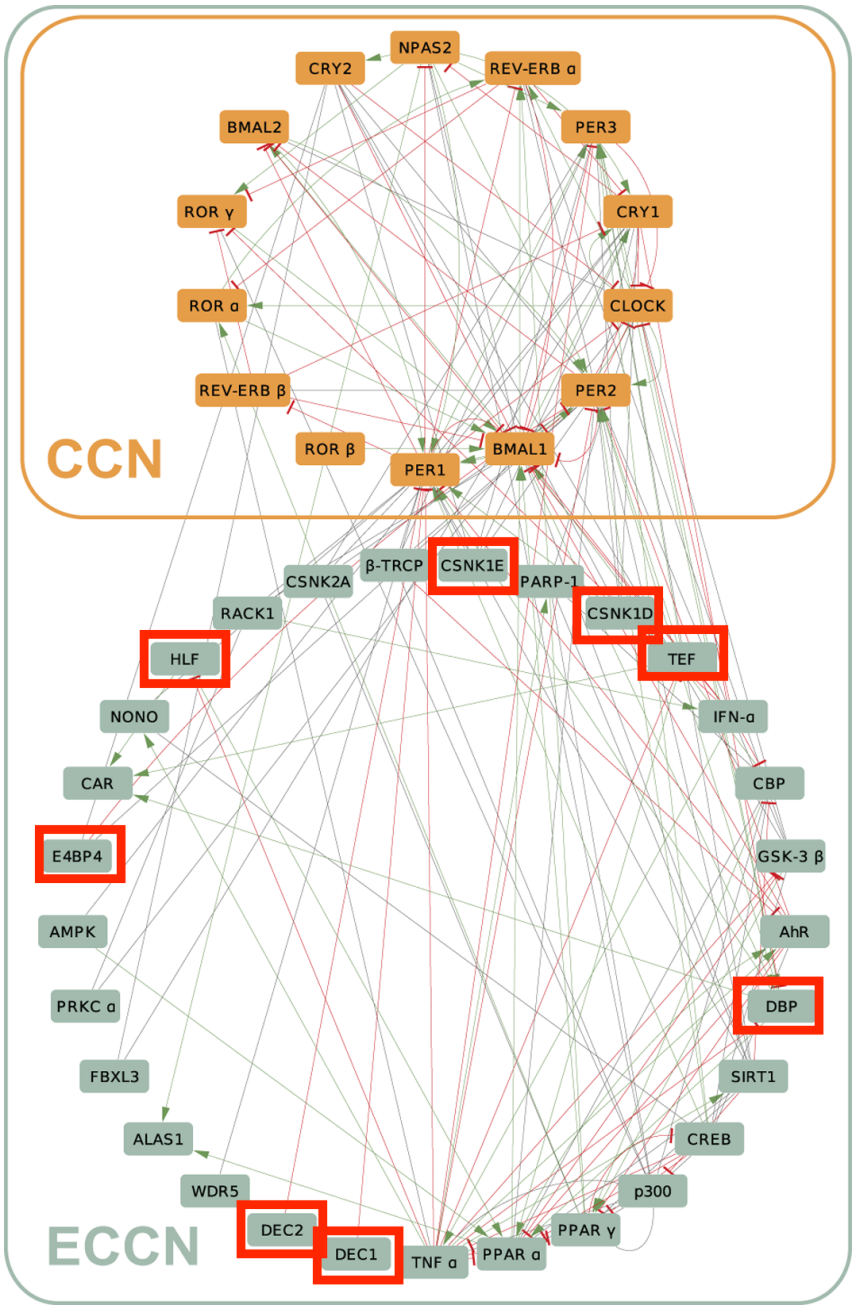


Figure 1.3: **The human core clock network (CCN) and the extended core clock network (ECCN)**. Edited from [3]. The CCN (orange) contains the core-clock elements. Genes in red boxes are those that are often associated with the circadian clock.

There is a huge amount of biological complexity in circadian networks. The molecular machinery that underlies the circadian clock consists of multiple transcriptional/translational feedback loops that are coordinated by specific sequences of rhythmic binding [6]. The evolution of the circadian clock has resulted in many redundancies, so the loss of function of a single clock gene generally does not result in arrhythmic phenotype.

The circadian clock is regulated by at least three clock-controlled DNA elements in the promoter site of the genes, known as the morning-time (E-box (canonical and non-canonical)), day-time (D-box), and night-time (Rev-Erb/ROR binding element or RRE)

elements [54]. The phase of circadian gene expression is generally correlated with the presence of these motifs in the gene promoter site. The translated proteins in turn bind to these specific promoter elements to activate or inhibit expression in a very specific way, which results in a 24 hour loop. The number of promoter sites changes for each clock gene, and most clock genes have a combination of motifs. Currently, there is no comprehensive study investigating these motifs in all clock genes. Table 1.1 indicates the presence of a motif for which evidence exists [20, 55], and figure 1.5 shows a schematic representation of the core clock gene's relationships by their promoter sites. A summary of known binding motifs could inform a powerful mathematical model that takes into account the correct combinations of promoters on each site. Korencic *et al.* [55] have published a simple model of this type using 5 clock genes in a delay differential equation model.

A simplified summary of the core circadian clock genes is presented, however the reader should note that the real level of biological complexity is expected to be much greater than is summarised here.

Core clock genes

Figure 1.3 represents a good summary of the core clock genes as it reflects most of the literature. The CCN genes in this figure, the ECCN genes in red boxes, and 2 additional genes are presented here as the core clock genes.

There is only one known clock gene that is essential for the circadian clock to maintain rhythm, and that is Bmal1 (Brain And Muscle ARNT-Like 1)[2]. There is evidence that Bmal1 is also a translation factor, as well as a transcription factor [45], which could explain its importance. Bmal2 is a paralog[3] of Bmal1 and is often considered to be important to the circadian clock function [3]. However, evidence is not consistent [56], and it is clear that the role of Bmal2 is far less important than the role of Bmal1. Bmal1 is in the family of basic-helix-loop-helix (BHLH) proteins, as are Clock (Circadian Locomotor output cycles kaput) and Npas2 (Neuronal PAS Domain Protein 2) [57]. CLOCK[4] and NPAS2 proteins form complexes with BMAL1. Both BML:CLK and BML:NPS complexes bind to E-box motifs to drive the expression of many other core clock genes that contain canonical E-boxes and non canonical E-boxes. A double knock-out of both Clock and Npas2 results in an arrhythmic phenotype [57]. The BML:CLK and BML:NPS complexes are involved in at least 3 major feedback loops; the period and cryptochrome negative loop, the Rev-Erb negative loop, and the Ror positive loop.

BML:CLK and BML:NPS activate the expression of the period genes Per1, Per2, and Per3, and the cryptochrome genes Cry1 and Cry2 via their E-box elements. They consequently form PER:CRY complexes which interact with BML:CLK and BML:NPS

[2]aka Arntl (Aryl Hydrocarbon Receptor Nuclear Translocator Like)
[3]Paralogs are genes related by duplication within a genome.
[4]Clock is rhythmic in mice, but not often observed to be rhythmic in humans.

complexes, blocking their activity [7]. The degradation of PER proteins by casein kinases $CKI\epsilon$ and $CKI\delta$, along with the shuttling of the proteins and complexes in and out of the nucleus, are known to be important post translational timing mechanisms [49].

The BML:CLK and BML:NPS complexes also activate transcription of Rev-Erbα and Rev-Erbβ. REVERBα and REVERBβ bind to RRE promoters on Bmal1, Npas2 and Clock, inhibiting their expression [49]. Ror (Retinoic Acid-related orphan receptor) (α, β & γ) genes are similarly activated, but ROR proteins bind to RRE elements and activate the expression of Bmal1, Clock, and Npas1.

The gene Ciart (Circadian Associated Repressor Of Transcription, aka Chrono/Gm129) is a relatively new addition to the set of core genes driving circadian clock dynamics. Ciart expression is also driven by BHLH complexes and CIART blocks the activity of the BHLH complexes. No review articles yet mention Ciart as a gene associated to the core circadian clock, but recent studies have identified Ciart's important role [58, 59, 21][5].

Figure 1.4 shows a simple representation of these 4 core feedback mechanisms. It is expected that the circadian proteins will also regulate many other 'clock controlled genes'.

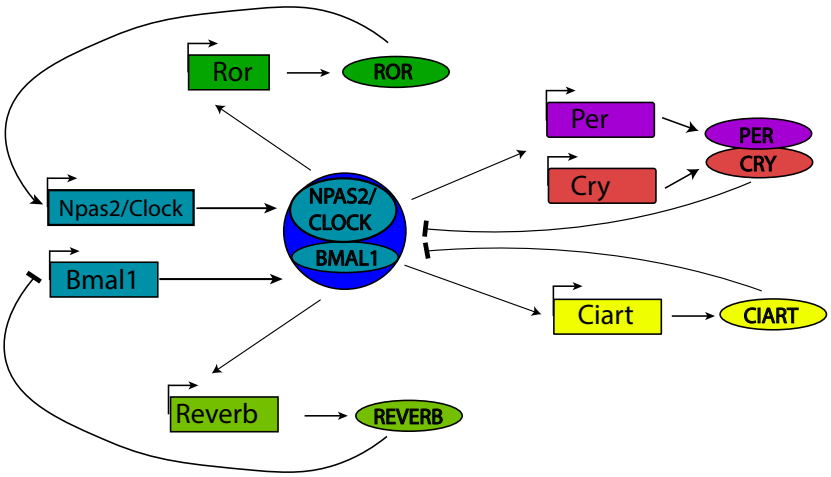


Figure 1.4: **Simple schematic showing the basic feedback loops that are likely to be the central time keeping mechanisms of the circadian clock.** Transcription of Pers, Crys, Rors, Rev-Erbs and Ciart are all activated by BML complexes via their Ebox promoter regions. PER:CRY complexes and CIART interfere with BML complex activity, resulting in negative feedback loops. REVERB binds to RRE sites and inhibits Bmal1, Npas2 and Clock transcription, and ROR binds to RRE sites and activates them.

Dec1 and Dec2 are also genes in the BHLH family [60] and are reported to be associated with the finer tunings and robustness of the circadian clock via interactions with E-boxes [61]. Among other BML:CLK and BML:NPS regulated transcripts are 4 PAR-bZIP genes, Dbp, Tef, Hlf, and Nfil3 that bind to D-box motifs in the regulatory sequences of other

[5]This thesis will also present significant evidence for Ciart's role as a core clock gene in both mouse and human data.

clock controlled genes [62, 63]. These genes are thought to be important for downstream regulation of other genes. They may have some involvement in the central oscillation mechanism, for example DBP has been observed to activate Period genes [62] but their involvement in the core time keeping mechanism is not widely reported.

Wee1 is a gene encoding a kinase that is responsible for regulating the CYCLIN B1/CDC2 complex and subsequent entry into mitosis [64]. The Wee1 promoter contains E-box motifs and so its expression is driven by the BHLH complexes. Wee1 is important when studying the link between the circadian clock and the cell cycle mechanism [65, 66].

Timeless is a very important circadian gene in *Drosophila*, and although exists in mammals, its role in the mammalian clock is still debated [67]. TIMELESS has been shown to regulate replication termination and cell cycle progression, and to have some significant post-translational interaction with PER-CRY complexes. Although Timeless is reported as a core clock gene in many studies, its mRNA does not show significant mammalian rhythms so will not be discussed in this thesis.

The core clock genes that have been discussed here are summarised in table 1.1, and partially in figure 1.5. Neither Ciart or Wee1 were identified by the review by Lehmann *et al.* [3] or Ukai *et al.* [4], but are included in table 1.1 due to recent evidence of their importance [58, 21, 59, 68, 65]. Bmal2 is included in this table but is italicised to highlight that there is limited evidence to suggest that it has an important role as a circadian clock gene[6]. The reader should note the alternative names for the genes; annotation sets often use these interchangeably and so it has been necessary to use all names in the writing of this thesis.

1.5 The master clock

The suprachiasmatic nucleus (SCN) is a region in the hypothalamus and is known as the master clock. The SCN coordinates all peripheral oscillators so that a synchronised rhythm is maintained at the organism level [69]. The SCN processes signals from light and other circadian inputs and directs the secretion of glucocorticoid hormones such as melatonin and cortisol [70]. The underlying neural and humoral mechanisms for the synchronisation of peripheral circadian clocks are still to be elucidated [44]. The SCN is the pacemaker of the circadian clock and its received inputs can set the phase of all peripheral clocks in the body (after sufficient time). For example, when humans cross time zones, the sudden shift to a new LD system results in the individual experiencing jet lag. The body eventually re-entrains to the new cycles of inputs [71]. Although input signals can reset the circadian clock, they are not necessary for driving the rhythms. Studies

[6]Bmal2's presence in this set may be a consequence of the text mining methods used.

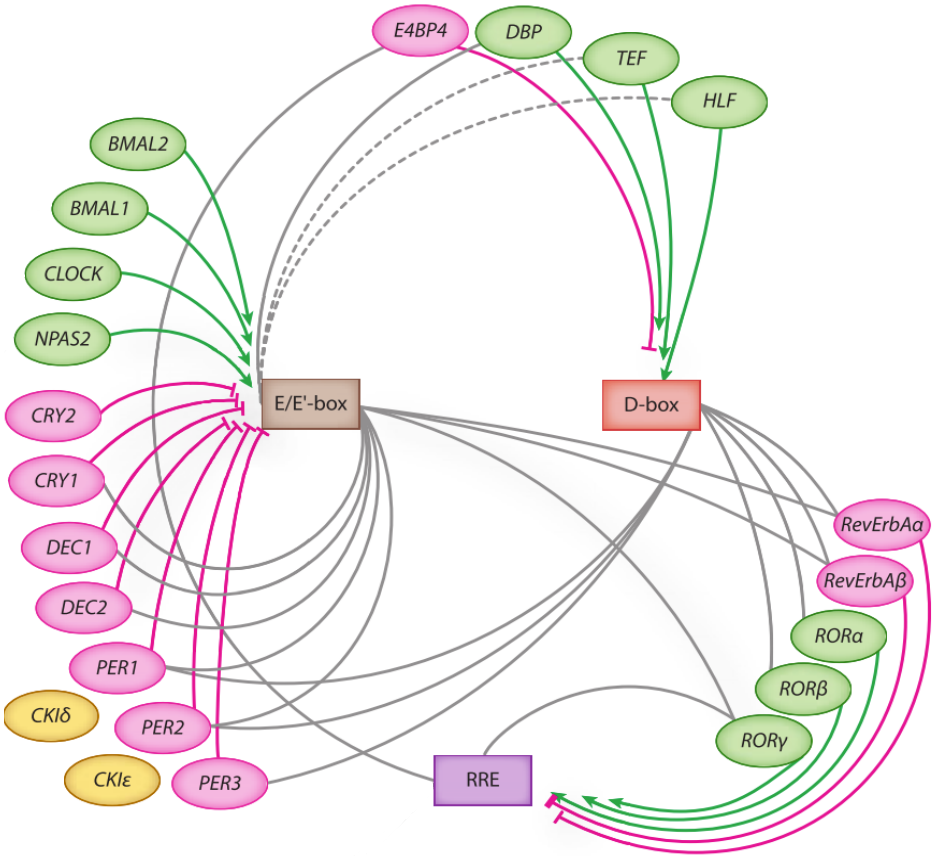


Figure 1.5: **Schematic a representation of the transcriptional network of the mammalian circadian clock.** Edited from [4]. E/E'-boxes are located on the noncoding regions of Per1, Per2, Cry1, Dbp, Rorγ, Rev-Erbα, Rev-Erbβ, Dec1, and Dec2. D-boxes are located on Per1, Per2, Per3, Rev-Erbα, Rev-Erbβ, Rorα, and Rorβ. RREs are located on those of six genes Bmal1, Clock, Npas2, Cry1, E4bp4, and Rorγ .

have shown that mice in constant light or dark conditions (and with no social cues) will continue to exhibit circadian behaviour [72].

1.5.1 Summary

As represented in figure 1.6, the circadian clock is made up of three major components:

- inputs that convey time information from the environment to the internal clock via the SCN,
- outputs that mediate clock-controlled behaviours and physiological processes,
- and a core clockwork capable of autonomous oscillation, which is present in peripheral oscillators in all cells of the organism. The core clock genes are synchronised in peripheral organs but subsets of clock controlled genes are tissue dependent [73].

Family Name	Gene Name	Alternatives	Receptors	Regulates
Basic	Bmal1	Arntl	RRE	E
Helix-Loop-Helix	Clock		RRE	E
	Npas2		RRE	E
	Bmal2	*Arntl2*		
	Dec1	Bhlhe40	E	E, **D**
	Dec2	Bhlhe41	E	E, **D**
Rev-Erbs	Rev-Erbα	Nr1d1	E, D, **RRE**	RRE
	Rev-Erbβ	Nr1d2	E, D	RRE
Period	Per1		E, D	
	Per2		E ,*D*	
	Per3		D	
Cryptochrome	Cry1		E, **RRE**	
	Cry2			
Orphan Nuclear	Rorα	Rora	D	RRE
hormone receptor	Rorβ	Rorb	D	RRE
	Rorγ	Rorc	E, RRE	RRE
PAR leucine zipper	Dbp		E, **RRE**	D
	Tef		*E*	D
	Hfl		*E*	D
	Nfil3	E4bp4	RRE	D
Casein Kinases	CK1δ	Csnk1d		
	CK1ε	Csnk1e		
	*Ciart	Chrono, Gm129	E	
Tyrosine kinase	*Wee1	E		

Table 1.1: **Table showing the established central clock genes, their alternative names, promoter motifs, and promoter motifs to which they have been found to bind.** This information is likely to not be complete or exhaustive. *These genes were not reported to be part of the circadian clock genes in Lehmann *et al.* (figure 1.3), but recent evidence has led to their addition to this set. Motifs in bold are those that are not included in figure 1.5, but have been reported in other literature [20, 21, 22, 23, 24]. Motifs in italics are those reported in Ukai et al [4], but these promoter sites are not commonly reported in other literature. Green motifs indicate activation and red motifs indicate repression.

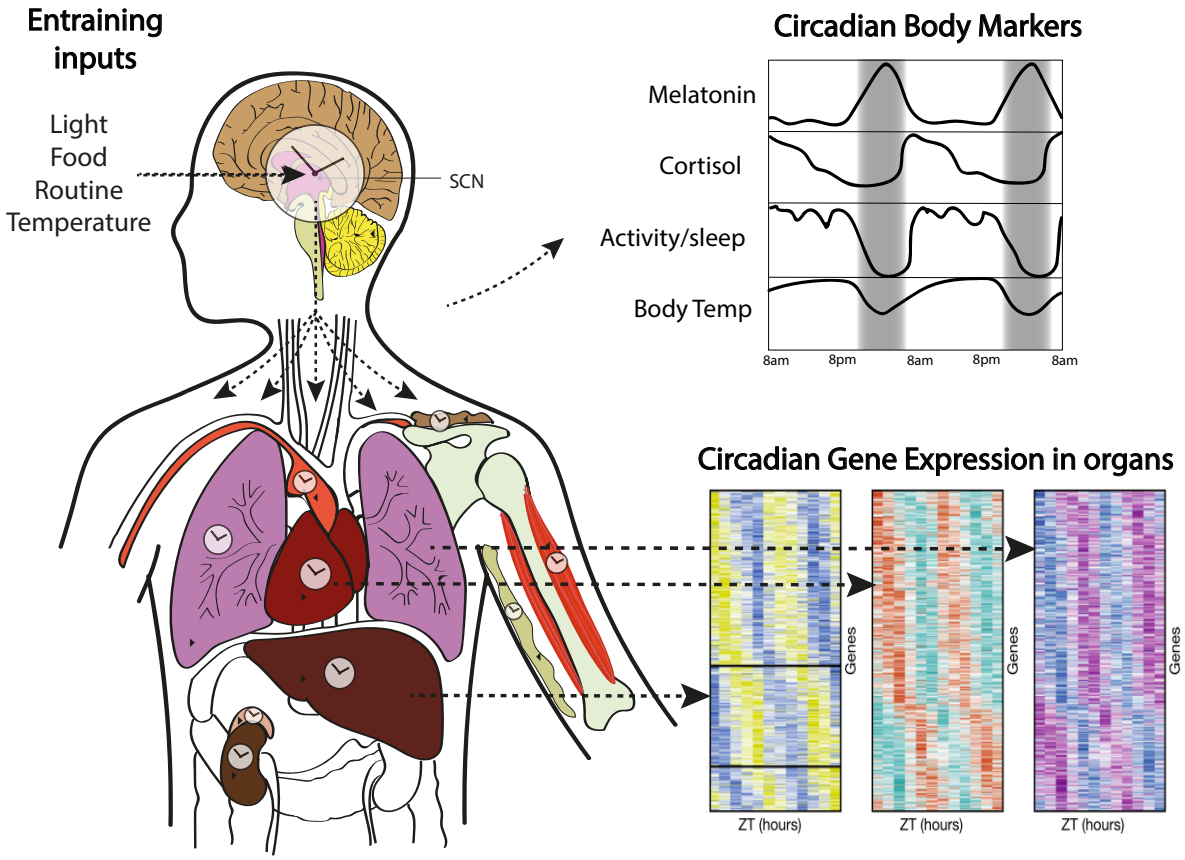


Figure 1.6: **Sketch that summarises the three major components of circadian clocks**. Circadian inputs entrain the clock via the SCN. Circadian body markers are physiological markers that can be tracked through activity/sleep and body temperature, and melatonin and cortisol levels in the blood [5]. There are many genes that show circadian behaviour in tissues. Some genes have organ specific circadian behaviour[6].

1.5.2 Use of mouse transcriptome to inform human disease

This thesis has required the use of timecourse transcriptome data, i.e. data from samples taken at fixed intervals over a time period of a day or two. Human tissue timecourse samples are rare due to the ethical and practical difficulties in biopsying a piece of human tissue every 1-4 hours over a full day. There does exist full genome timecourse human data from blood samples [14], but as will be discussed in chapter 4, it uses a custom microarray making it incompatible for comparison to other independent datasets. This work will present rare timecourse data from live human oral mucosa (from a standard human microarray GeneChip) which is currently unpublished and was shared with us for the purposes of this thesis. However, the availability of time-labelled human tissue transcriptome data is still comparatively limited to that of mouse data. In fact, except for the human timecourse data that we acquired through collaboration, we could not find any healthy human transcriptome data that was labelled with the time of day of tissue sampling. The use of time labelled mouse transcriptome data to validate our algorithm design was essential. Large sets of high resolution data are available from mouse experiments, with data from multiple organs and from genetically identical mice in different environments.

Mouse models are often used as model organisms to study human biology due to the genetic similarities. However, the networks linking genes to regulatory processes and disease are likely to differ between the two species [74]. Consequently, although both mouse and human data are used in this work, mouse and human datasets will at no point be directly compared; the mouse data is used only to validate any assumptions that we will make with the human model.

Furthermore, all (real) data used in this work originates from living tissue. Although *in vitro* models are useful for many purposes, circadian rhythms can be manipulated and influenced massively by artificial culture methods. It is likely that signals from the SCN and humoral signalling across the whole body is very important for the proper functioning of the mammalian circadian clock. As we aim to develop an algorithm to be potentially useful in the clinic, we will only use data that originated from live tissue samples.

1.6 Computational challenges

Modern high throughput technologies in areas such as genomics, transcriptomics, proteomics and metabolomics are generating huge amounts of data, and providing greater scope for our study and understanding of the circadian clock. The complexity and size of these circadian datasets pose a computational challenge. The standard data analysis techniques for differential gene expression, which are primarily based on comparisons of gene expression correlations (e.g. gene set enrichment analysis) can rarely be used to

generate useful metrics in circadian data. If one imagines attempting to measure the correlation of a cosine wave with a sine wave, standard correlation analyses will provide metrics indicating no relationship between the variables.

Computational methods in rhythmicity detection of timecourse data have been central to circadian data analysis. The phase of a 24 hour periodic measurement (the highest peak of a sine-like wave) is a vital metric, where the absolute value of the gene expressions are often less meaningful. Multiple methods of rhythmicity analysis have been developed for this purpose and will be discussed in chapter 3.

Timecourse data is essential for the generation of useful circadian metrics using rhythmicity analysis. This requires long, expensive and complex experiments where identical protocols are carried out every 1-6 hours, across at least a full 24 hour period, needing consistency in carefully controlled environments. Over the past few years, it has been a scientific and mathematical challenge to find time fingerprints in transcription so that algorithms can be developed that can tell the time from less time points, and even single transcriptomes. A literature review of the studies that have attempted to do this will be presented in chapter 4.

This thesis uses innovative and novel methods to demonstrate that the robust and predictable nature of the core circadian clock facilitates deduction of a time signature from just a single sample. Time-Teller will be introduced as an algorithm that can tell, from one single sample, if the circadian clock is functional, and if it is, at what time the sample was taken.

Time-Teller is a data and mathematics driven algorithm, and although biological context allows us to understand results and derive comparisons with expected biology, the algorithmic design is completely unbiased by the information thus far presented. The reader will observe agreement between the set of training genes used to train Time-Teller, and the genes presented above as the core clock genes (table 1.1). This is a result of the reproducibility of circadian gene expression, which is a consequence of the robustness of circadian clock. This will be discussed in the following chapter.

1.7 Structure

This thesis is organised so that each chapter builds on the results and conclusions of each previous chapter. Each chapter contains its own introduction, literature review (where relevant), and conclusion section. The chapters are organised as follows:

2. **The Robustness of the Circadian Clock.** Circadian clock robustness will be explored using open-source tools and a mathematical model of the circadian clock. An equivalent stochastic model is developed and a similar analysis is performed.

3. **Timecourse Transcriptome Analysis.** - This chapter will introduce the mouse, human, and *in silico* timecourse datasets used to train Time-Teller. A novel method in circadian gene identification is presented.

4. **Time-Teller** - This chapter will review existing time-telling methods and present the novel Time-Teller algorithm. Time-Teller is validated in a "leave-one out" sense with the human, mouse and *in silico* training data.

5. **A Metric of Clock Dysfunction** - This chapter will present a novel metric, Θ, that represents circadian clock function. This is tested via *in silico* knock-downs, and independent mouse and human data sets.

6. **Circadian Clock Dysfunction in Human Cancer** - This chapter will review evidence for clock dysfunction in cancer. Θ is measured for cancer data sets and we explore how Θ might be used as a novel prognostic tool in breast cancer, in order to inform personalised treatment strategies.

7. **Discussion** - This chapter will briefly summarise findings and discuss suggestions and opportunities for further work.

Chapter 2

The Robustness of the Circadian Clock

Various models of differential equations have been published that attempt to describe the underlying dynamics of the molecular circadian clock [44]. Although all of the detailed mechanisms that govern gene expression are not fully understood, mathematical models incorporating some experimental findings are of great help when trying to understand complex systems like the circadian clock. Models can help us to discover new characteristics of complex systems, and to test any hypotheses that cannot easily be tested experimentally. This chapter will focus on the analysis of the dynamics of one particular mammalian circadian clock model, referred to in this thesis as the Relogio model [7].

Circadian clocks have complex structures with multiple interconnected feedback loops; the models that describe them have a high level of complexity. Investigating changes in parameters of these models needs a much more robust method than a trial-and-error approach. Based on theory presented in three publications by Rand *et al.* [75, 76, 77], and using a consequent Matlab GUI called PeTTsy [78], a formal analysis of parameter sensitivity of the Relogio model is presented in this chapter. This allows insights into robustness and flexibility of the model under parameter perturbations. Following this, the deterministic Relogio model is adapted to a stochastic model, and a parallel analysis is performed using methods in Minas *et al.* [8].

2.1 Models of gene expression

Mathematical models of temporal gene expression are designed based on reaction schemes derived from experimental evidence. Often there is a large amount of published literature on the reaction scheme of interest, the experimental designs and results are complex, and sometimes the literature can even report contradictory outcomes. When modelling gene expression, transcriptome data is informative to show how gene expression changes over

time, under different conditions, or when the genes themselves are edited. Experimental results inform us as to which transcription factors bind to which promoter sites, via (for example) ChIP-seq or surface plasmon resonance [79]. The stability of proteins and mRNA can be estimated using labelled genes and proteins and measuring florescence, for instance. Protein abundance is much harder to measure, but can be measured using techniques such as fluorescence imaging, or mass spectrometry. Data from mutant experiments such as that from a gene knock-down or over-expression, or RNAi experiments, provide additional insights into dynamics.

Data mining of the vast array of publications on circadian clock genes, cycling proteins, environmental effects, the effects of gene editing, *etc*, is the first step in circadian clock model design. Data exists for multiple organisms, multiple strains of organism, *in vitro* or *in vivo*, in different environmental conditions, and other factors in experimental design. When designing a reaction scheme for a model, the designer needs to choose a subset of this evidence, and be able to make informed decisions when some evidence contradicts other evidence. Due to such a huge level of complexity and ambiguity, no mathematical model can be a perfect representation of reality, and no circadian clock model can reflect every bit of evidence the scientific community has collected. Even so, we will show here how mathematical models are powerful tools to help us make sense of such vast complexity.

2.1.1 Evolutionary design of circadian clocks

Molecular circadian clocks are a feature of the internal workings of all kingdoms of life [9]. Consistent architectural features of each kingdom's circadian clock are complex auto-regulatory networks, multiple feedback loops, and genetic redundancies [46]. Light is the main input to the circadian clock, but the molecular circadian clock does not need driving forces to maintain oscillation. Mice have been shown to continue to exhibit circadian activity patterns even in constant darkness [72], with period shortening (in constant darkness) or period lengthening (in constant light).

When experiments are carried out on mice, genetically identical, matched gender and age strains are used to minimise experimental variation. We expect that the genes and biology of these "identical" mice have a certain level of synchronisation, as experiments are designed to minimise variation. Humans do not exist in a perfectly routine environment of light dark cycles, temperature cycles, or exact meal times, and humans are certainly not genetically identical to one another. Despite this, the next chapter will show that due to the robustness of the mammalian circadian clock, circadian gene expression is remarkably synchronised amongst human individuals too. In this chapter, we aim to present a mathematical argument for why this occurs.

We do not argue that the circadian clock is completely inflexible, however. Jet Lag is

a physical representation of the body's adaptation to a very large shift of the circadian clock. It may take a few days to phase shift the body a few hours forwards or backwards, but eventually our bodies adapt to new time-zones. We argue that the circadian clock has a balance of robustness and flexibility, in order to maintain robust rhythms whilst adapting to new environments.

Clock design

In theory, a simple unforced oscillator can exist that is made up of only two genes in a negative feedback loop. A stable limit cycle could exist with a 24 hour period, and these two genes could be the time keeper of an entire organism. It would be a crucial job for these two genes; whilst regulating each other, to also regulate all other time-related genes and processes, all in a cycle of 24 hours. It is clear that circadian clocks have evolved to have the complexity to have a contingency plan when things inevitably go wrong.

The number of genes that make up the core of the molecular circadian clock changes across kingdoms, and due to the complex regulatory relationships between genes, it is difficult to even define which genes are the "core" clock genes. The following section outlines (mathematically) why this complexity results in circadian clocks being robust and inflexible. Robust and inflexible refer to the difficulty in changing the behaviour of the circadian clock. Physically this is relating to the body's slow adaptation to an environmental change, molecularly this relates to the consistent gene expression patterns upon stochastic fluctuations, and mathematically this relates to how parameter perturbations in the model do not drastically change the behaviour of the model. A periodic solution can be described (mostly) by its phase, period, and amplitude, but in this section behaviour is only referring to period and phase.

2.1.2 Design principles underlying circadian clocks

It has been famously reported that John von Neumann said

> *With four parameters I can fit an elephant, and with five I can make him wiggle his trunk* [80].

This was meant as an analogy for the dangers of over-fitting a model to data: that even if an initial model set-up is not ideal, by adding just a few more variables and parameters, the model can eventually be fit to any data or show any desired behaviour. This could be true in some simple scenarios: a well constructed four parameter model could be powerful, with perturbations of each parameter having significant effects on the overall behaviour of the model. Here, it is argued that tightly coupled feedback models such as the oscillators considered here, do not have this property. Evidence is discussed for the inflexible character of circadian systems, where the parameters affect the model solution

in a highly correlated way so that the change in the solution has a much lower dimension than the parameters.

2.2 Circadian clock models

There are multiple publications involving the modelling of the mammalian circadian clock; these are summarised in table 2.1. Each model is slightly different, as each model designer has chosen slightly different subsets of the available information. They are all however, primarily based on data from the mouse SCN. The Relogio model will be the only model discussed in this thesis, but a similar analysis could be done for any of the other models[1]. A very good literature review of all models can be found in the recent review article in BMC systems biology by Podkolodnaya *et al.* [44].

Forcing

Forcing represents some external input to the system (e.g. light) and can be built into a model via a periodic term (e.g. via a step function or sine wave) that increases the production rate of a variable quantity with a 24 hour period. Most of models in table 2.1 are unforced; the parameters are fit so the dynamics show desirable behaviours, and the solutions are stable limit cycles with an approximate 24-hour period. These models are free-running, and represent autonomous biological oscillators.

Only Leloup & Goldbeter have attempted to use forcing in the models, as very little is known about how external signals are communicated to our genes; the process of light entering the retina, the signals being processed by the brain, and then being communicated from there to the rest of the body in a way that can control single cell gene expression, is not yet understood. Some studies have attempted to evaluate a molecular change in mammalian clocks when light signals are disturbed [87, 88, 89, 90] but there are mixed conclusions. There is a small amount of evidence that the Per2 gene is involved in light signal processing, and this is how Leloup & Goldbeter incorporate a forcing term into their models. A similar robustness analysis of Leloup & Goldbeter's models is published in [75, 76, 77], with significant conclusions of circadian clock model robustness to parameter perturbation.

2.3 The Relogio Model

Relogio *et al.* [7] designed a set of 19 ordinary differential equations (ODEs) that describe the central mechanism of the molecular circadian clock using two feedback loops. One loop describes the Per/Cry(PC) terms, and the other describes the Rev-Erb/Ror(RBR)

[1]The Korencic delay differential model would have to be linearised first.

Model First Author(s)	Ref	# Vars	# Pars	Description
Relogio	[7]	19	71(76)	ODEs describing dynamics of 3 loops: Per&Cry, Rev-Erb and Ror loops, including terms for mRNA, proteins, and post translational modifications. Free Running.
Korencic	[22]	6	46	Set of delay differential equations, with dynamics based on promoter site interactions. For 6 mRNA only; Bmal1, Per2, Rev-Erbα, Cry1, Dbp, & Rorγ. Free running.
Leloup & Goldbeter	[81][82]	16/19	55/70	ODE describing dynamics of 2 loops: Per&Cry and Rev-Erb loops, including terms for mRNA, proteins, and post translational modifications. Forcing through periodic activation of Per.
Forger & Peskin	[83]	74	36	ODE describing dynamics of 2 loops: Per&Cry and Rev-Erb loops, including terms for mRNA, proteins, and post translational modifications. More terms as paralogs are modelled separately. Free running.
Kim	[84]	74	36	Stochastic version of the Forger & Peskin model
Mirsky	[85]	21	132	ODE describing dynamics of 3 loops: Per&Cry, Rev-Erb and Ror loops, including terms for mRNA and proteins. Paralogs are modelled separately. Free running.
Woller	[86]	16	96	ODEs modelling the circadian clock interaction with metabolic entities. Free running.

Table 2.1: **Summary of mammalian circadian clock models**. Each model has a different design, and set of variables and parameters.

terms. There are 5 mRNA terms for genes Bmal, Cry, Per, Ror and Rev-Erb. Other terms describe protein translation, movement in and out of the nucleus, protein complexing, and phosphorylation. The reaction scheme is shown in figure 2.1. The set of 19 ODEs with their fitted parameter values, are in appendix B.

2.3.1 Relogio *et al.* summary of study

The model design was initiated by a literature search on the behaviour and features of the molecular circadian clock in the mouse SCN. They focused on three main levels of regulation:

- Protein-DNA interactions, where some proteins are transcription factors activating or inhibiting transcription of a gene. For example, presence of REV-ERB in

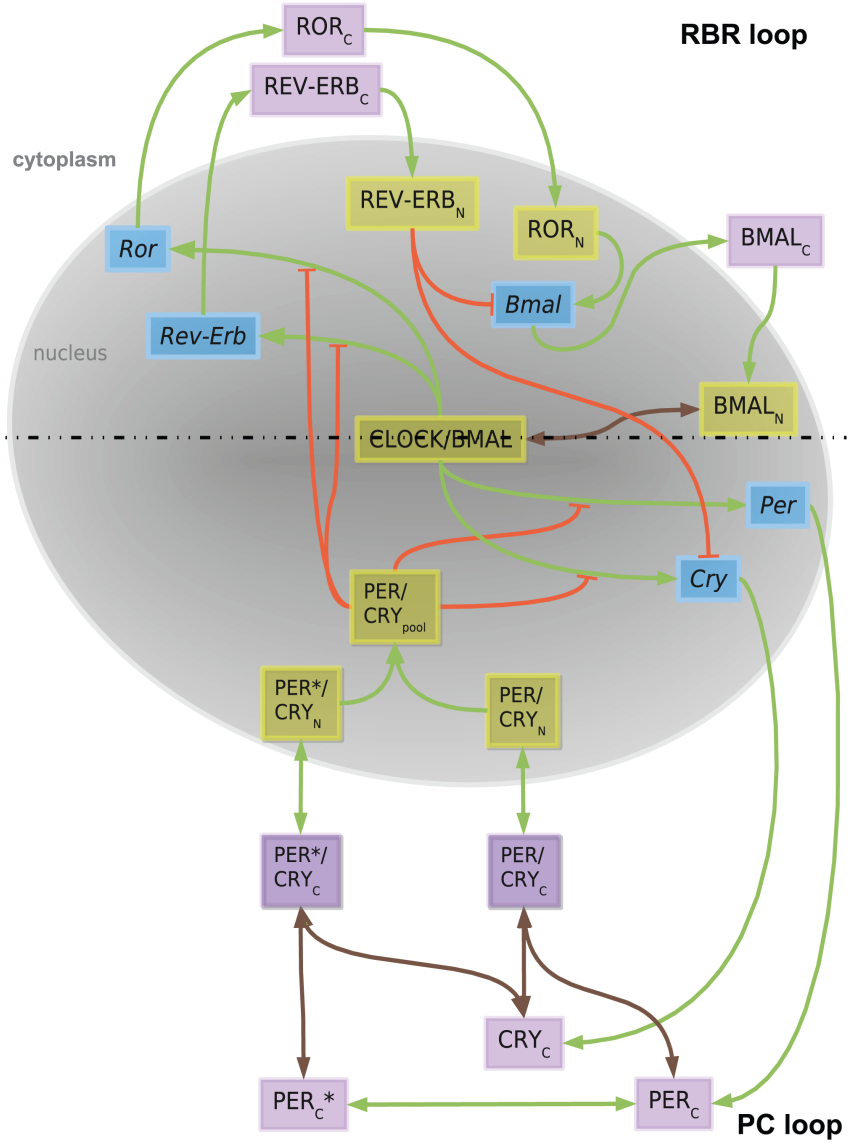


Figure 2.1: **The Relogio model reaction scheme**. Scheme from [7], representing the dynamics of the set of ODEs.

the nucleus inhibits the transcription of Bmal, and the presence of ROR activates it. The mRNA transcription terms were fit using 14 datasets consisting of mostly biochemical data from the mouse SCN with some including mutant phenotypes.

- PER/CRY complex blocking CLOCK/BMAL mediated transcription, which consequently stops/slows transcription of all genes other than Bmal itself.

- Stability of the RNA and proteins (represented by the degradation rates). Phosphorylation of PER is included as it "*contributes to the delay which is necessary to generate a circa 24 hour period.*" mRNA and protein half-life data was accumulated and reported from 7 studies that used mouse fibroblasts and stem cells. There is limited insight into degradation kinetics, so all degradation is applied using linear terms.

The design of the system of equations was based on the law of mass action. The transcription terms are described by Hill functions that were derived using using Michelis-Menten kinetics. Parameters were either fixed at experimentally determined values or fit using LTI (linear-time-invariant) systems theory. Details of this and the fitting process are in the supplementary information of the published study [7].

The authors report that *in silico* knock outs result in similar phenotypes to the experimental data they based their model on. There is some study into the effects of changing parameter values to the phases and period of the model. One of the main results in the study is that changing the Per degradation rate can change the period of the system in a non-monotonic way. They state that the model has enough complexity to be able to offer a reason for unexplained experimental observations. For example, experiments have shown period shortening for both over-expression [91] and under-expression [92] of Per, which does not seem possible at first sight. However, the model also has this behaviour when increasing or decreasing the Per degradation parameter from its fitted value.

There is a vast amount of detail and information in this model. The authors do not claim that it is a precise quantitative model of the circadian clock, but its dynamics capture some of the major feedback loop dynamics in a concise way. Without any editing, and keeping this model exactly as it is published in the supplementay information of [7], we will carry out a formal parameter sensitivity analysis for this model using PeTTsy. First, we discuss the theory underlying PeTTsy.

2.4 Sensitivity Analysis Using PeTTsy

This section is a summary of works in ***Design principles underlying circadian clocks****, by Rand et al. [75, 76, 77], whose content has been designed as the MATLAB GUI, PeTTsy [78].*

We are considering models of the form:

$$\dot{x} = \frac{\mathrm{d}x}{\mathrm{d}t} = f(t, x, k) \tag{2.1}$$

where x are the n state variables and k is a vector of the s parameters of the model ($k_i \geq 0 \; \forall i$). The solution of interest is denoted as

$$x = g(t, k) \tag{2.2}$$

and for all situations in this thesis will represent an attracting stable limit cycle with period T. As the parameters will have different orders of magnitude, it does not make sense to use absolute changes in parameters in the analyses. Relative changes to parameters can

be assessed using $\eta_j = \log k_j$. This means that for small changes δk_i to the parameters, in practise we use $\delta\eta_j = \delta k_j / k_j$

2.4.1 Decomposing the system

The variation of the solution δg produced by parameter perturbation $\delta\eta$ can be described (up to second order terms) as;

$$\delta g(t) = \sum_j \frac{\partial g}{\partial \eta_j}(t)\delta\eta_j + O(||\delta\eta||^2) \tag{2.3}$$

Let M be the linear transformation matrix that maps $\delta\eta \to \sum_j \frac{\partial g}{\partial \eta_j}\delta\eta_j$. Time is restricted to a discrete set of values $t_1, ..., t_N$ so that we only study the vector $(\delta g(t_1), ..., \delta g(t_N))$, and so M has N rows. M has s columns, one for each parameter in the model defined in equation (2.1). We define the discretised derivative matrix M to be

$$M = \begin{pmatrix} \frac{\partial g}{\partial \eta_1}(t_1) & ... & \frac{\partial g}{\partial \eta_s}(t_1) \\ . & . & \\ \frac{\partial g}{\partial \eta_1}(t_N) & ... & \frac{\partial g}{\partial \eta_s}(t_N) \end{pmatrix} \tag{2.4}$$

so that

$$\delta g = \sum_j M \cdot e_j \cdot \delta\eta_j + O(||\delta\eta||^2) \tag{2.5}$$

$$\tag{2.6}$$

Let the singular value decomposition[2] of M be $M = WDU$ where D is a diagonal matrix with entries $\sigma_1 \geq \sigma_2 \cdots \geq \sigma_s$. Substituting the SVD of M for M in (2.5) gives

$$\delta g(t) = \sum_j \left(\sum_i W_{ij}\sigma_i U_i \right) \delta\eta_j + O(||\delta\eta||^2) \tag{2.7}$$

where the U_i are the columns of U.

Letting $S_{ij} = W_{ij}\sigma_i$

$$\delta g(t) = \sum_{i,j} S_{ij} U_i \delta\eta_j + O(||\delta\eta||^2) \tag{2.8}$$

The column vectors $U_i = (U_{i,1}, ..., U_{i,n})^T$ are known as the **sensitivity principal components**, because U is an orthogonal matrix and so the U_i are unit vectors orthogonal to each other. The **sensitivity singular values** $\sigma_1 \geq ... \geq \sigma_s$, represent the dominance

[2]The SVD is discussed in detail in the next chapter and in further detail in appendix A.

of each principal component in the decomposition.

From equation 2.8, and knowing that the U_i are orthogonal directions, we can see that S_{ij} completely determines the size of the effect of small changes $\delta\eta_j$. The matrix $S_{ij} = \sigma_i W_{ij}$ characterises the sensitivity of the system with respect to each parameter, and is known as the **parameter sensitivity spectrum**.

2.4.2 PeTTsy analysis of the Relogio model

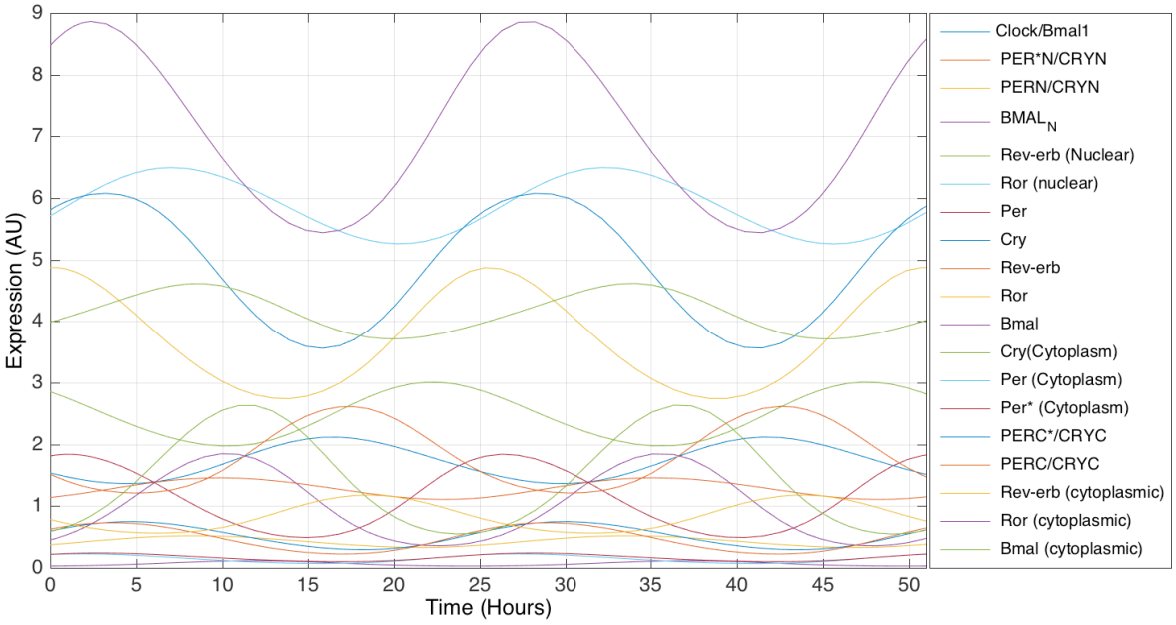


Figure 2.2: **Simulation showing the Relogio model limit cycle solution**. The solution is shown as a timecourse over 52 hours for all 19 variables, and was generated using the ode45 differential equation solver in MATLAB.

The stable limit cycle of the set of ODEs designed by Relogio *et al.* [7] is shown as a 52 hour timecourse in figure 2.2. It should be noted here that the precise amplitude is meaningless and so is expressed in arbitrary units (AU)[3]. The important characteristics of the solution are shape, period, and phase. The model was analysed with PeTTsy, which calculated the period of the oscillations to be 25.2 hours.

Parameter sensitivity

The sensitivity singular values are plotted in figure 2.3 and show an exponentially decreasing behaviour. This shows that there is a very constrained behaviour to the dynamics of this model as the parameters are changed. The first four principal components of the Relogio ODEs account for $\sim 96\%$ of the variation in the model, the first accounting for

[3]By making appropriate changes in the parameters, we can usually quite easily scale any solution $g_i(t)$ to a desired amplitude.

$\sim$ 66 %. This shows the low dimension of possible solutions the system can have after parameter perturbation. This means that the order of phases does not change in response to small changes, and when the order of phases does change, there are only 2 or 3 likely combinations.

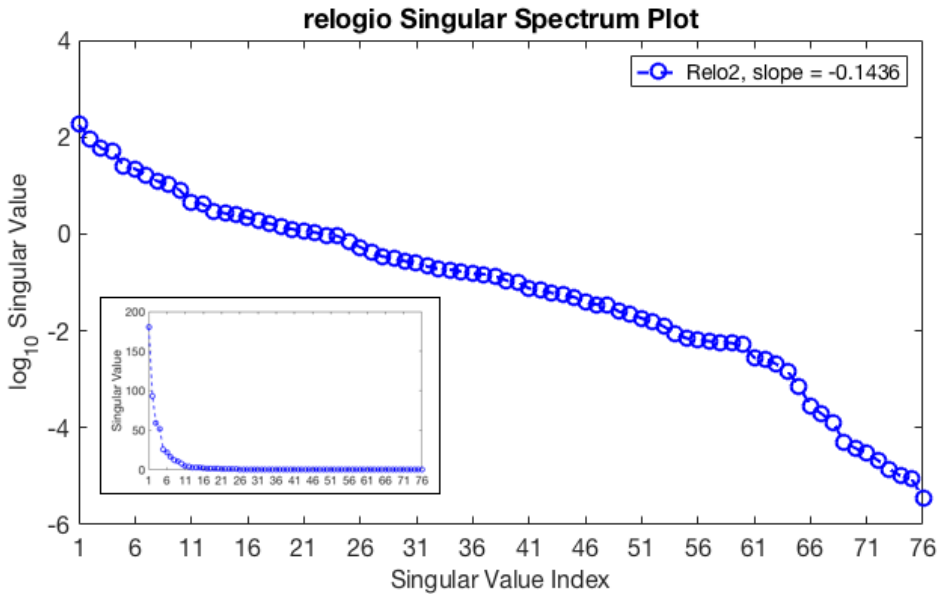


Figure 2.3: **Plot showing the exponentially decreasing Sensitivity Singular Values for the Relogio Model**. Plot is shown in log scale, with smaller plot in linear scale to show that only 4 singular values have significant value.

Additionally, we use the parameter sensitivity spectrum to show that just a small subset of the 76 parameters can push the model into these directions. The parameter sensitivity spectrum is shown for the 20 most sensitive parameters, for the first four principal components in figure 2.4.

This evidence shows that the behaviour of this solution to the Relogio model is very much determined by the model structure, and this cannot be changed by small perturbation to the parameter values.

ODE model analysis summary

This analysis has shown that the influential directions in the parameter space of the Relogio model are much less than its total dimension. It is only practically possible to move the solutions in relatively few directions using the parameters.

This section provided evidence for the hypothesis presented in [75], that circadian clocks are ultimately robust and inflexible to small changes. The following section will question whether this holds for the Relogio model's dynamics when they are modelled stochastically.

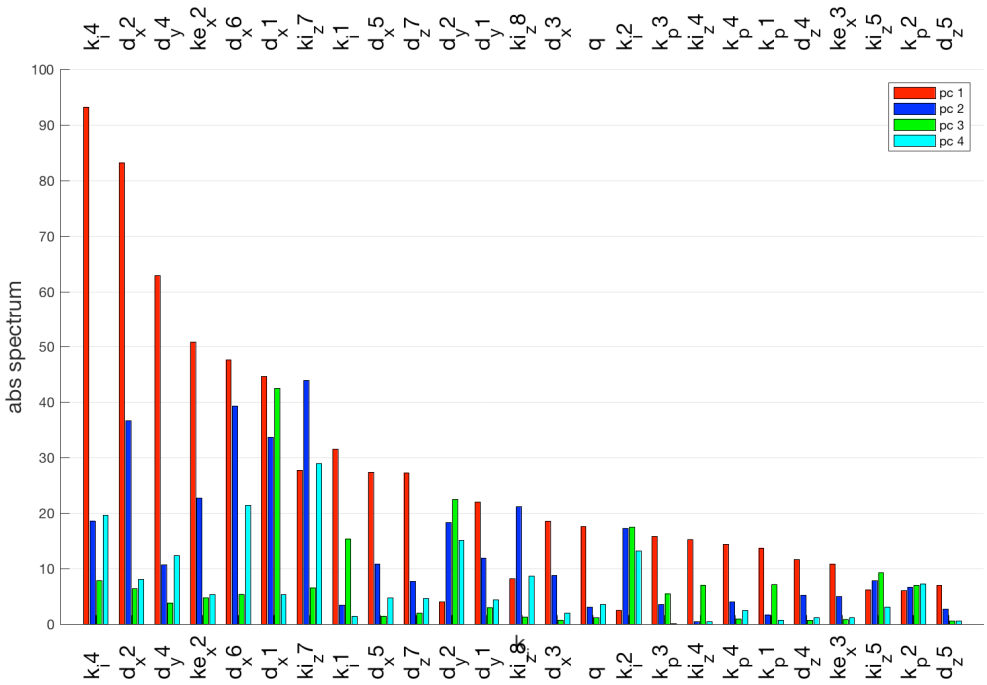


Figure 2.4: **Bar plot showing the absolute parameter sensitivity spectrum**. The y-axis represents corresponding values of entries in S_{ij}, representing the contribution each parameter j has to principal component i (see (2.8)).

2.5 Stochastic Modelling

Deterministic models do not incorporate random noise and internal fluctuations present in nature; deterministic models are an ideal, expected behaviour when a system is very large. However, individual cells have a discrete number of molecules and proteins, and internal fluctuations in the human body are very much a reality. Stochastic models allow the incorporation of noise and randomness to a model. In the deterministic approach taken in the previous sections, both time and variables (a measure of concentration) were positive continuous numbers. In the stochastic models, time is still continuous, but the variables are discrete integers, and represent number of molecules as opposed to concentration.

2.5.1 Introduction to stochastic models

Stochastic models of molecular dynamics are usually described in terms of reactions. The amounts of the molecular components are described by a state vector, $Y(t) = (Y_1(t), ..., Y_n(t))^T$ where $Y_i(t)$ $i = 1, ..., n$ denotes the number of molecules of each species at time t. These molecules undergo a number of possible reactions (e.g. transcription, translation, degradation) where the reaction of index j changes $Y(t)$ to $Y(t)+\nu_j$, $\nu_j \in \mathbb{R}^n$. The vectors ν_j are called stoichiometric vectors. Each reaction occurs randomly at a rate $w_j(Y(t))$ (often called the propensity functions), which is a function of $Y(t)$.

It is common in studying stochastic systems to introduce a system size Ω. This parameter occurs in the intensities of the reactions $w_j(Y(t))$ and controls molecular numbers. For cellular systems a natural choice is to use molar concentrations and therefore regard Ω as Avogadro's number in the appropriate molar units (e.g. nM^{-1}) multiplied by the volume of the reacting solution in appropriate units (e.g. in litres (L)). The use of the system size Ω is also important because it allows the vector of concentrations $Y(t)/\Omega$ to be defined for the stochastic system. It is the dynamical system in terms of these concentrations that have a well-defined deterministic limit as Ω tends to infinity given by the differential equations studied in the sections above.

In a stochastic model, there are no longer exact solutions to the model, as with a deterministic model, but there are individual trajectories that are Markov in nature (i.e. memoryless, and the current state is all the future state depends on, and not any past state). Exact trajectories can be simulated using the Gillespie algorithm [93] (also known as the stochastic simulation algorithm). Distributions of the solutions can then be approximated from simulated data to summarise the model. Generating enough trajectories to accurately predict these probability distributions is however, an extremely computationally expensive task, so is not ideal for calculating multiple trajectories for large systems.

2.5.2 Stochastic circadian models

There have been few studies that have attempted to look at stochastic models of circadian dynamics, and the majority concern plant [94], drosophila [95], or other non-mammalian rhythms. Gonze *et al.* [96] modelled a simple oscillator (based on the *Drosophila* circadian clock) to assess the robustness of circadian oscillations with respect to molecular noise. They found that robust circadian oscillations can occur with a limited number of mRNA and protein molecules. A stochastic mammalian model of circadian rhythms was published by Forger *et al.* [83], where they adapted their previous 74 equation ODE model to a stochastic model. They report interesting results such as observing that some variables in the stochastic model can oscillate even when they do not in the ODE model, and that events happening on the promoter (proteins binding to the promoter site of clock genes) must happen very quickly for 24 hour rhythms to be possible. They attempt some robustness analysis on this model, and state that the robustness of the circadian clock appears to increase as more molecules are present, or more frequent promoter interactions occur. They show that gene duplication (e.g. redundancy between Per1 and Per2) increases the robustness by providing more promoters with the transcription factors can interact with.
Instead of attempting to replicate this existing stochastic model of mammalian circadian rhythms, we have created a stochastic version of the Relogio model [7], which is based on the deterministic Relogio model and incorporates a system size parameter, Ω. This adaptation was done based on methods in [96]. We consider it has units L/nM and a value in the range 100-1000 in accords with typical cell size estimates [8].
The individual trajectories were calculated in MATLAB, using an implementation of the Gillespie algorithm, shown below.

Gillespie Algorithm

For a stochastic model with
n variables in state vector $\mathbf{Y}(t) = (Y_1(t), ..., Y_n(t))$,
P parameters that are scaled by Ω if they have units of concentration,
M reaction rates in vector $\mathbf{w}(t) = (w_1(\mathbf{Y}(t)), ..., w_M(\mathbf{Y}(t)))$,
a stoichiometry matrix, $S^{(n\times M)}$ made up of M vectors $\nu_{\mathbf{j}} \in \mathbb{R}^n$ (the stoichiometric vectors)
the algorithm to simulate one trajectory is;

set running time T_{final} ;
set system size Ω;
set initial conditions $\mathbf{Y_0}$ (must be whole numbers) ;
set time $t = 0$;
while $t < T_{final}$ **do**
  generate 2 random numbers $0 \leq r_1, r_2 \leq 1$;
  calculate reaction rates $\mathbf{w}$ based on current $\mathbf{Y}$ state;
  calculate $w_{sum} = \sum_{i=1}^{M} w_i$;
  finding the index j of the first reaction that satisfies $r_1 < \sum_{i=1}^{j}(\frac{w_i}{w_{sum}})$;
  update the state vector to reflect the j^{th} reaction happening: $\mathbf{Y} = \mathbf{Y} + \nu_j$;
  update time $t = t + \frac{\log(1/r_2)}{w_{sum}}$.
end

Algorithm 1: The Gillespie algorithm, adapted from [97].

2.5.3 The stochastic Relogio model

The stochastic Relogio model has 44 reactions using 19 variables and 72 parameters, where parameters with units of concentration are scaled with system size parameter accordingly. As $\Omega \to \infty$, the stochastic solution converges to the ODE solutions. The 44 reaction rates for the stochastic Relogio model are shown in appendix B. Using $\Omega = 500$, and the same initial condition on the limit cycle for all simulations, 5 trajectories of the BMAL1/CLOCK complex variable were simulated and are shown in red in figure 2.5. The scaled ODE trajectory is shown in black for comparison. It is apparent that the stochastic model loses synchrony from the ODE model over time. Without any forcing in the model, the period is variable about the ODE period of 25.2 hours, and there is an increasing loss of synchrony amongst independent simulations. It is important to recognise here that, although the variable phases of ODE and stochastic models do not match, the relationships between variables stay the same. Figure 2.6 shows the same simulations as in figure 2.5 but for 10 periods. CLOCK/BMAL1 is plotted against another variable, PER*N/CRYN, again with the deterministic limit cycle in black. The trajectories all follow the elliptic shape of the limit cycle, and appear to be normally distributed about the deterministic mean, which is an important observation for the following section.
The first difficulty in evaluating these distributions around the deterministic mean is to simulate enough trajectories in order to generate a distribution. This could be solved by an algorithm called the linear noise approximation (LNA), however the second difficulty is that the times that each trajectory passes a transversal plane to the deterministic mean drifts due to the stochastic nature of the simulation (i.e. periodic stochastic models lose synchronicity of trajectories over time). The pc-LNA is an algorithm that aims to overcome these problems.

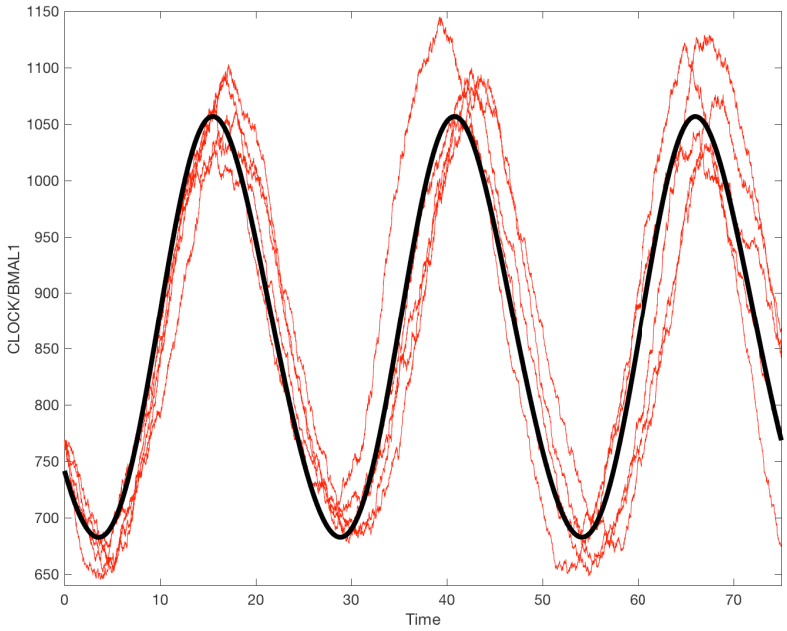


Figure 2.5: **Plot showing timecourse limit cycle solution and stochastic trajectories of the Relogio model.** Plots shown are for the variable representing the BMAL/CLOCK complex. The limit cycle solution is shown in black, and 5 stochastic simulations in red, for $\Omega = 500$. The ODE model solution was scaled for comparison.

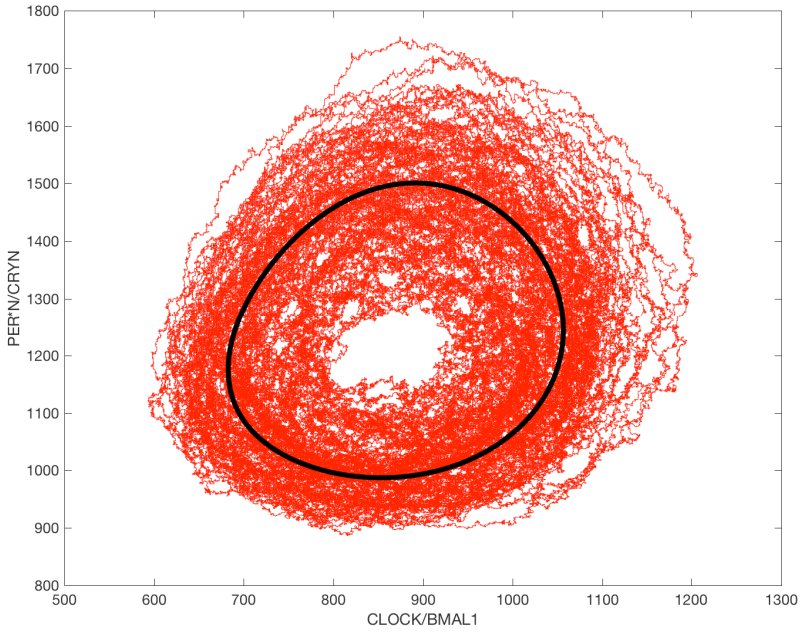


Figure 2.6: **Plot showing 2D limit cycle solution and stochastic trajectories of the Relogio model.** Plots shown are for the variable representing the BMAL/CLOCK and PERN/CRYN complexes. The limit cycle solution is shown in black, and 5 stochastic simulations in red, for $\Omega = 500$. The ODE model solution was scaled for comparison.

2.5.4 Linear noise approximation

The LNA allows approximation to analytical solutions of desired terms, unlike the Gillespie algorithm. It is commonly used as an alternative algorithm to calculate trajectories for stochastic systems, allowing further analysis to be done [8]. The linear noise approximation (LNA) expresses the stochastic solution $Y(t)$ in terms of a stochastic variable $\xi(t)$ using the following Ansatz;

$$X(t) = \frac{Y(t)}{\Omega} = x(t) + \frac{\xi(t)}{\sqrt{\Omega}} \tag{2.9}$$

where $x(t)$ is the deterministic solution (which has units of concentration).

We are interested in the periodic solutions of $\dot{x} = F(x)$, so that $x(t) = g(t)$ for $0 \leq t \leq \tau$, and $x(t)$ is a stable limit cycle of period τ.

Let $J(x)$ be the $n \times n$ Jacobian of $F(x)$ (i.e $(J(x))_{ij}$ is $\partial F_i/\partial x_j$ evaluated at x).
Let $C(s,t)$ be the family of $n \times n$ fundamental matrices which are the solutions of :

$$\frac{\mathrm{d}}{\mathrm{d}t}C(s,t) = J(g(t))C(s,t) \tag{2.10}$$

with the properties

- $C(s,s) = I$
- $C(t_1,t)C(t_0,t_1) = C(t_0,t)$ for all $t_0 \leq t_1 \leq t$
- $C(0,t)C(0,s)^{-1} = C(s,t)$

The evolution of the stochastic variable $\xi(t)$ is given by:

$$\xi(t) = C(t_0,t)\xi(t_0) + \eta(t_0,t), \quad t > t_0 \tag{2.11}$$

where $\eta(t_0,t) \approx MVN(0, V(t_0,t))$ and $MVN(\mu,\Sigma)$ represents the multivariate normal with mean μ and covariance matrix Σ. Here, the covariance matrix is

$$V(t_0,t) = \int_{t_0}^{T} C(s,t)E(s)E(s)^T C(s,t)^T \mathrm{d}s \tag{2.12}$$

and $E(s) = \sqrt{SW(x(s))}$ is the matrix product of the stoichiometry matrix and the diagonal matrix of reaction rates $u_j(x(s))$.

The key result about the LNA is that if we fix $t > t_0$, then as $\Omega \to \infty$ the true distribution of ξ converges to the distribution given by the LNA [98].

2.5.5 Transversal distributions

For each point x on the limit cycle γ we define our transversal to be an $(n-1)$ dimensional hyperplane $\mathbb{S}_x$ normal to γ at $x \in \gamma$. We use the sketch in figure 2.7 to explain the transversal distributions. $X(t)$ is a stochastic trajectory of the oscillatory system, with initial condition $X(0)$ at time t_0, and is represented by the blue and red trajectories. $G_N(X(t))$ represents a point on the limit cycle γ, which is shown as the black curve in figure 2.7. The segment connecting $X(t)$ to $G_N(X(t))$ is orthogonal to the tangent to γ at $G_N(X(t))$, i.e. $X(t)$ lies in the hyperplane orthogonal to this tangent.

$Q_x^{(r)}$ is the rth intersection of a trajectory to a given transversal section S_x to the limit cycle, shown as the "target" in the sketch. Each intersection $Q_x^{(r)}$, is shown as the black crosses on the transversal plane, for each period of the red trajectory.

2.5.6 pcLNA

The Gillespie algorithm produces exact trajectories of the stochastic system but the computing power and time needed is immense for the size of the systems we want to study. In order to do further analysis, it is necessary to use a more approximate approach. As shown above, the LNA [99] is fast and analytically tractable, but Minas *et al.* have shown that for free running oscillators, when the LNA is used, the variance of $\xi(t)$ grows without bound as t increases. Minas and Rand [8] have developed the phase-corrected LNA (pcLNA), which overcomes this limitation. Using this analytically tractable pcLNA, we are able to calculate and analyse the probability distributions of phase states at the transversals. As the transversal distributions are MVN, we are then able to compute a stochastic sensitivity analysis of the system, using Fisher information.

2.5.7 pcLNA implementation

We discuss the pcLNA briefly here, but the full explanation can be found in [8]. The approach is to amend the LNA ansatz in (2.9) so that we can correct time;

$$X(t) = g(s) + \frac{\kappa(t)}{\sqrt{\Omega}} \tag{2.13}$$

where $g(s_i) = G_N(X(t))$ is used to reset time from t to s and $G_N(X(t))$ represents a point on the limit cycle γ.

Now, in the LNA, for free running oscillators, while the variance of $\xi(t)$ grows without bound as t increases, the variance of $\kappa(t)$ is uniformly bounded. The pcLNA algorithm iteratively uses the LNA in steps of Δt to move from state $X(s_{i-1})$ to the new state $X(s_{i-1} + \Delta t) = X_i$. After each step, the phase of the system is corrected such that

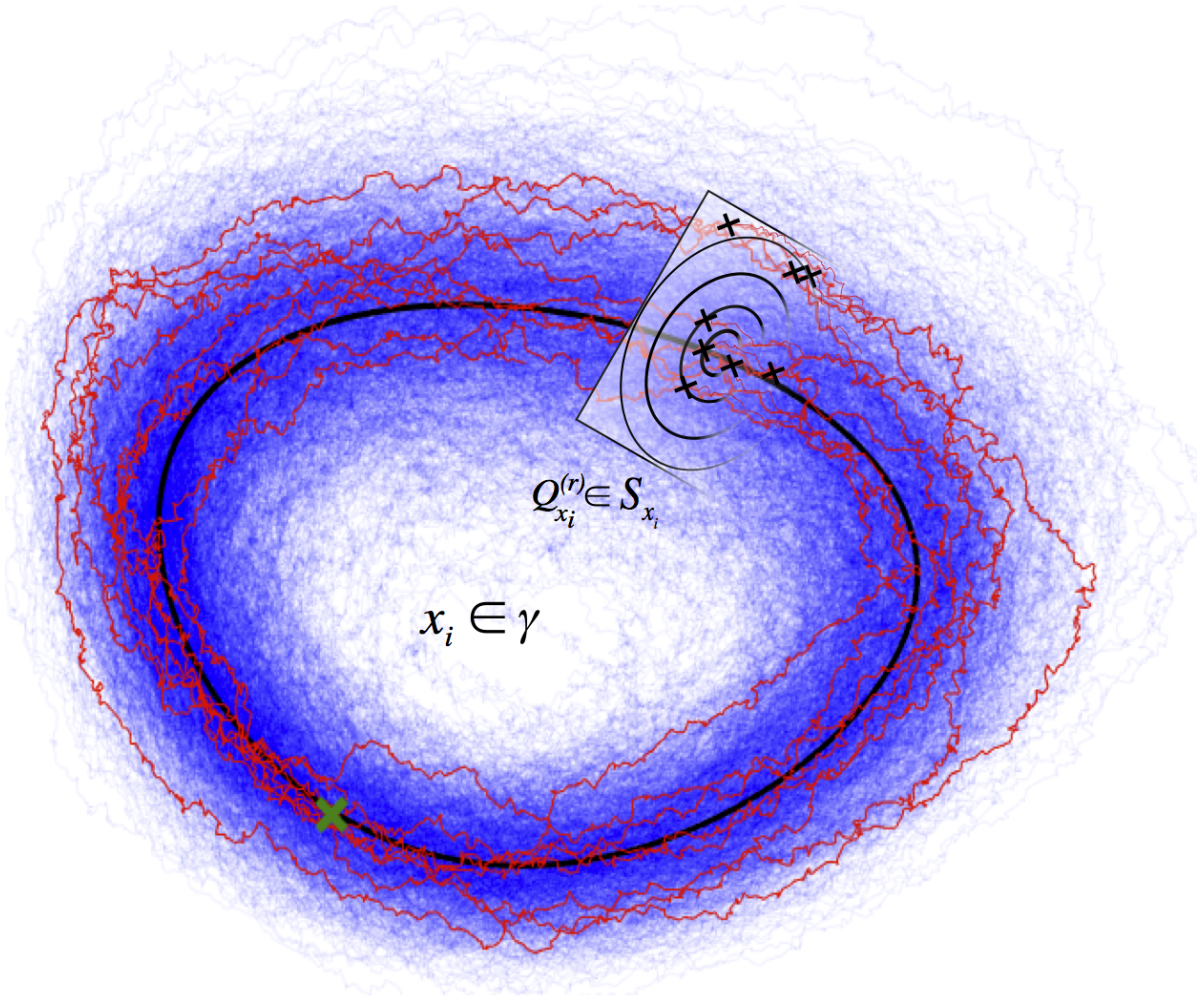


Figure 2.7: **Sketch showing 1000 Gillespie simulations for 2 variables and 8 periods of the stochastic Relogio model.** Trajectories $X(t)$ in blue, are normally distributed about the limit cycle, $g(t)$, in black. One trajectory of 8 periods is highlighted in red. S_{x_i} is the transversal to the limit cycle where the red trajectory intersects at black crosses, $Q^r_{x_i}$. The distribution of Q on S is a multivariate Gaussian distribution of dimesion $n-1$. We are interested in how this changes as parameters change.

$g(s_i) = G_N(X_i)$ so that

$$\xi(s_{i-1} + \Delta t) = \Omega^{1/2}(X_i - g(s_{i-1} + \Delta t)) \tag{2.14}$$

is replaced by $$\kappa(s_i) = \Omega^{1/2}(X_i - g(s_i))) \tag{2.15}$$

where the $\kappa(s_i)$ are MVN distributed.

In summary, the steps of the algorithm are;

1. Choose a time-step size $\delta\tau > 0$

2. Set initial condition $\kappa(s_0)$ and $X_0 = g(s_0) + \Omega^{-1/2}\kappa(s_0)$

3. For iteration $i = 1, 2, ...$
   (a) sample $\xi(s_{i-1} + \delta\tau)$ from MVN$(C_i\kappa(s_{i-1}), V_i)$,(b) Compute $X_i = g(s_{i-1} + \delta\tau) + \Omega^{-1/2}\xi(s_{i-1} + \delta\tau)$
   (c) set s_i to be such that $G_N(X_i) = g(s_i)$ and $\kappa(t_i) = \Omega^{1/2}(X_i - g(s_i))$

This is summarised in the sketch in figure 2.8.

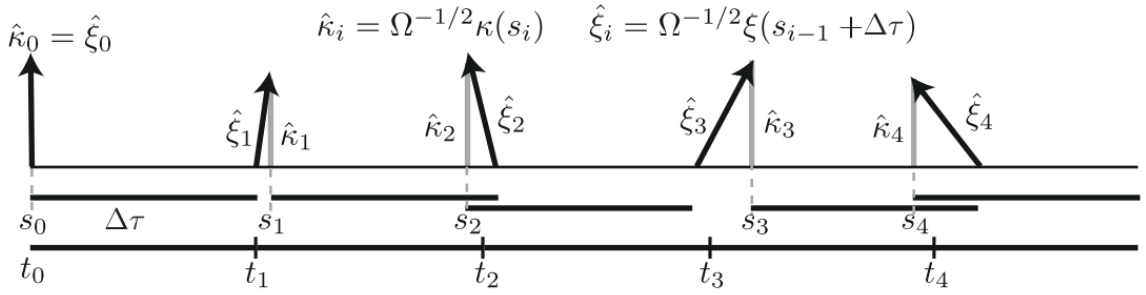


Figure 2.8: **Sketch showing the main step in the pcLNA algorithm.** From [8], The solid horizontal bars below the horizontal axis are all of length $\Delta\tau$, the basic time step of the algorithm. The black arrows show $\hat{\xi}$ and the grey arrows $\hat{\kappa}$.

2.5.8 Fisher information matrix

The Fisher Information is a measure of the amount of information that a variable X contains about a parameter θ. If $\ell = \log P(X, \theta)$, where P is the probability distribution, the Fisher Information Matrix (FIM) $\mathcal{I} = \mathcal{I}_{ij}$ is a square matrix, where i and j are the ith and jth entry of parameter vector θ;

$$\mathcal{I}_{ij} = E\left(\frac{\partial \ell}{\delta\theta_i}\frac{\partial \ell}{\delta\theta_j}\right) \tag{2.16}$$

Now consider q phase states of the limit cycle $x_i = g(t_i), i = 1, ...q$ on γ where $0 \leq t_1 < t_2 < ... < t_q < \tau$. If $X(t)$ is a stochastic trajectory we consider how it meets the transversal sections at the x_i as t increases.

Let $Q_{x_i}^{(r)}$ be the rth intersection of the transversal to x_i, so that multiple transversals are possible for multiple periods of each trajectory. If we let $\underline{Q} = Q_{x_1}^{(1)}, ..., Q_{x_q}^{(1)}, Q_{x_1}^{(m)}, ..., Q_{x_q}^{(m)}$, then we are interested in the MVN distributions at these transversals;

$$P(\underline{Q}|X(t)) = P(Q_{x_1}^{(1)}, ..., Q_{x_w}^{(m)}|X(t)) \tag{2.17}$$

These distributions allow us to anaytically compute the Fisher Information matrix in order to perform parameter sensitivity analysis on the stochastic Relogio model.

As P in the above scenario is an MVN with mean $\mu = \mu(\theta)$ and covariance $\Sigma = \Sigma(\theta)$ then

$$\mathcal{I}_{ij} = \frac{\partial \mu^T}{\partial \theta_i} \Sigma^{-1} \frac{\partial \mu^T}{\partial \theta_j} + \frac{1}{2} tr \left(\Sigma^{-1} \frac{\partial \Sigma^T}{\partial \theta_i} \Sigma^{-1} \frac{\partial \Sigma^T}{\partial \theta_j} \right) \tag{2.18}$$

As μ and Σ are calculable, so is the Fisher information matrix. The FIM measures the sensitivity of P to a change in parameters, where the eigenvalues of the FIM are a measure of this sensitivity, and the eigenvectors characterize the parameters that contribute to this sensitivity (see [8]).

2.5.9 Sensitivity analysis on the stochastic Relogio model

The pcLNA algorithm was used to generate enough trajectories so that we could calculate the Fisher Information Matrix for the stochastic Relogio model, using $\Omega = 1000$. Figure 2.9 shows that the eigenvalues of the FIM decay exponentially, with the first eigenvalue being significantly higher than the others. This means that the influential directions in the parameter space of the system are much less than its total dimension. This suggests that this stochastic circadian model is very robust to parameter perturbations.

Additionally, figure 2.10 shows that the subset of the parameters that can push the model into these directions. k_{i2}, k_{i1}, and k_{i4} are the Cry, Per, and Ror inhibition rates, respectively. These are the parameters that can push the system into the direction of the first principal component. This is in agreement with the findings of Forger et al [100], when they observe that the speed that a transcription factor can bind to a promoter site has significant effects on the model's dynamics.

Both the ODE model and stochastic model show very few possible behaviours that the models can present upon parameter perturbation. Seven of the top nine most sensitive parameters were in agreement in both the stochastic and ODE model analyses. These are d_{x1}, d_{x2}, d_{x6}, ki_{z7}, ke_{x2}, k_{i1}, and k_{i4}. This suggests that circadian gene expression has the same characteristics of robustness even when noise and stochasticity are taken into account.

The parameters k_{i4}, d_{x6}, and ki_{z7} determine the Ror inhibition, RORN degradation, and RORC translation rates. This is in agreement with conclusions in Relogio *et al.* [7] where their results suggested an important role of the RBR loop on the clock system.

d_{y4} is very sensitive in the ODE model, but not in the stochastic model. Conversely k_{i2} is very sensitive in the stochastic model, but not as much in the ODE model. As the Relogio model was not designed to model the effects of stochastic noise, we do not dwell too much on the literal representation of these parameters, but acknowledge that the difference in the analyses may show additional insights to these models that we would not otherwise have known.

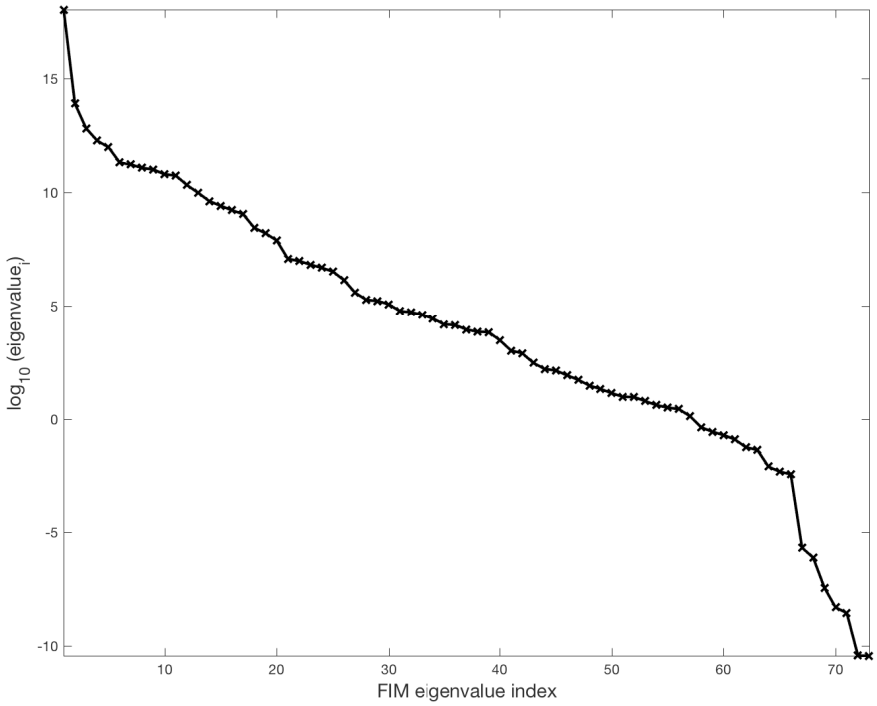


Figure 2.9: **Plot of the eigenvalues of the FIM for the stochastic Relogio model.** Eigenvalues rapidly decrease, indicating few directions that the model solution can be pushed in.

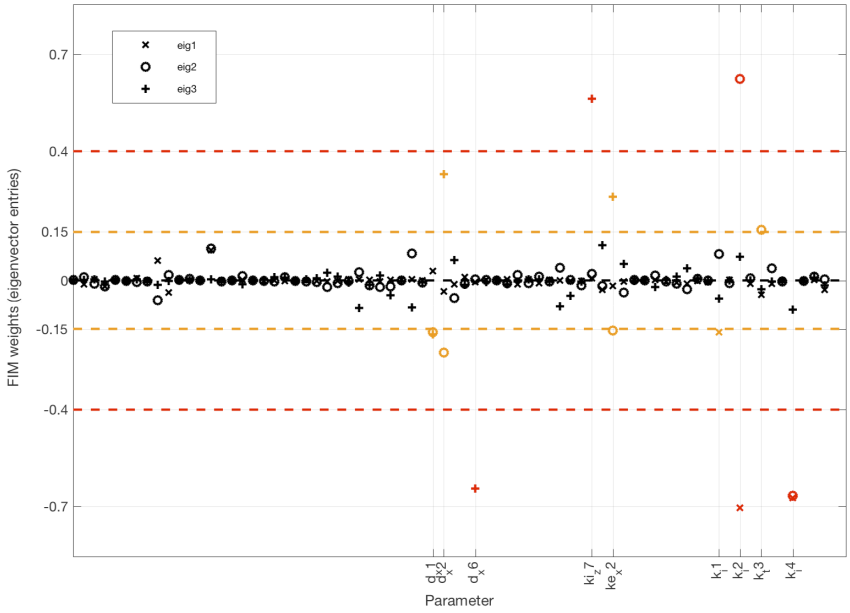


Figure 2.10: **Plot of the eigenvectors of the FIM for the stochastic Relogio model.** These are the weights for sensitivity of the parameter for the stochastic Relogio model.

2.6 Summary of chapter

This chapter has explored the robustness of the circadian clock using an ODE model and an equivalent stochastic model. The necessity and probable evolutionary design for this robustness was discussed before *in silico* evidence was provided.

The phase sensitivity to parameter perturbation was measured for both the deterministic and stochastic Relogio model. The results show that both the ODE and stochastic Relogio model are very robust to parameter perturbation, and only a small subset of parameter changes would have significant effects on the behaviour of the model. The sensitive parameters were mostly in agreement suggesting that circadian gene expression has the same characteristics of robustness even when noise and stochasticity is taken into account.

In tightly coupled circadian oscillators, phase orderings are relatively unchanged by small perturbations to the system. This *in silico* evidence now allows us to approach real circadian data, with reason for expectation of a certain level of robust and synchronised behaviour amongst independent samples. This will be now explored in the next chapter.

Chapter 3

Timecourse Transcriptome Analysis

This chapter describes and analyses what will make up the training data for Time-Teller models that will be introduced in the next chapter. The mathematical analyses of the previous chapter provides reason to believe that we will see a certain level of robustness amongst real circadian timecourse data, so that such models would be possible to train. Current methods in circadian data analysis are almost all about fitting ∼24 hour rhythmic curves to the data in order to measure oscillation robustness. A novel mathematical approach is taken in this chapter to analyse synchronicity, and used in combination with standard approaches. The purpose of this synchronicity analysis is to choose a set of reliable training genes for the Time-Teller model.

The online availability of published time-course transcriptome data from GEO and the circadian rhythms labs has made this project possible. One extremely valuable unpublished human dataset was acquired through an MTA agreement with Dr Georg Bjarnason of Sunnybrook Research Institute, Canada.

Altogether 15 unique and independent datasets have been used in this thesis, and this chapter summarises them, where emphasis is given to the main timecourse datasets. Only mouse and human datasets are used in this thesis. Equivalent studies and data exist for circadian genes in organisms such as drosophila or zebra fish, on which a similar analysis could be done, if there were reason to do so.

3.1 Timecourse transcriptome data collection

RNA-seq and microarrays are the main technologies used to quantify whole genome expression. RNA-seq provides single base pair resolution data. The raw data consists of millions of sequences of "ACGT"s with read quality metrics that need alignment to a reference genome, QC, and normalisation, before any quantification can be done. RNA-seq experiments need input parameters such as depth of sequencing prior to the experiment, complicating the design [101]. The single base pair resolution provides gene quantification whilst providing additional information that allows, for example, measurement of

alternate spliceforms.

RNA-seq data is huge, complex and has historically been far more expensive than microarray analysis. For studies that require multiple samples, RNA-seq has not yet been a major contender, and there is a lack of timecourse RNA-seq data available. It is anticipated in the next few years, larger scale RNA-seq experiments will be realistic. The methods in this thesis will be readily adaptable to timecourse RNA-seq data. To show this, a short subsection in this chapter provides an analysis of the existing RNA-seq timecourse mouse data, and shows the comparability of microarray and RNA-seq data.

The majority of available timecourse transcriptome data exists from microarrays, as in the past they have been easier and cheaper for probing the transcriptomes of multiple samples. Probes have been designed that allow unambiguous identification of genes, where a single numerical signal from a probe represents that gene's expression. The probes have a high level of specificity, so different probes can even identify different spliceforms of the genes. Various GeneChips are ready-made for humans and mice, available to buy from various companies. New GeneChips are made and sold as knowledge and technology develops.

The data produced from a microarray is semi-quantitative, and does not measure gene expression in an absolute manner. Microarrays can provide information into changes in gene expression in a controlled environment, when samples are compared to each other. To clarify; a microarray cannot absolutely tell if there are more transcripts of gene X than of gene Y in a single assay, but can tell us if there are more or less of X in assay 1 compared to assay 2. The numbers that a microarray produces are functions of both transcript number and of binding affinity of a transcript to a probe, amongst other factors [102]. As different GeneChips use slightly different designs, effectively weighting these contributing factors differently, direct comparisons between different GeneChips are not viable. Mappings between GeneChips are available and possible, but would add more variation and complexity to an analysis.

3.1.1 Microarrays

A number of commercial technology platforms offer microarray technologies, the biggest four including Affymetrix, Illumina, Agilent and Nimblegen. The Affymetrix gene chip system is the original (and most widely used) system, and is used exclusively in this thesis. Affymetrix gene chip technology is not discussed here, but a good explanation can be found in [102].

The largest set of mouse timecourse data available uses GeneChip Affymetrix MoGene 1.0 ST. The human timecourse data from the MTA was quantified with GeneChip Affymetrix HG U133 2.0. These GeneChip labels are only relevant to thesis when we try to compare datasets, as different GeneChips are not easily compared. In this thesis, only

microarray data of these GeneChips will be used. An example will be shown in this section to explain the difficulties of comparing different GeneChips and different technologies.

Microarray normalisation algorithms

The result of a microarray is essentially an image of spots of different intensities. The goal of microarray processing procedures is to get a single value of expression for each probe from these intensities. The most popular algorithms to do this are all built into the Bioconductor package in R. Raw microarray data is in the .CEL file format, which is handled by the affy package as an AffyBatch object. The gene-level intensities stored in both the resulting ExpressionSet object can produce probe level gene intensities.

The most popular microarray normalisation algorithm is robust multiarray averaging (RMA) which performs background correction, normalization, and summarization in a modular way [103]. The summarization step fits a parametric model that accounts for probe effects, assumed to be fixed across arrays, and improves outlier detection. Residuals, obtained from the fitted model, permit the creation of useful quality metrics.

GeneChip RMA (gcRMA) is an extension of the RMA that is able to use the sequence-specific probe affinities of the GeneChip probes to attain more accurate gene expression values. gcRMA also takes into account MM (mismatch) data, which means it also takes into account non-specific binding. gcRMA results in a percentage of the data being omitted, if it does not meet a set quality standard. This does not suit this project, as it would result in different sized data sets or a gene of interest being omitted.

Many other normalisation techniques exist but will not be mentioned here as the RMA is sufficient. However, a problem with the RMA, when handling multiple datasets, is that normalization and summarization require all arrays to be analyzed simultaneously in order not to introduce a batch bias. This study uses dozens of datasets, and it is not viable to run the algorithm on so much data at once, and when working incrementally with new batches of arrays.

fRMA

McCall *et al.* [104] developed the frozen robust multi-array analysis (fRMA) algorithm, which allows individual and independent analysis of microarrays while retaining the advantages of multi-array preprocessing methods such as the RMA. fRMA is aligned to the aims of this PhD project as it allows independent and reproducible normalisation of samples, whether they are independent single samples or part of a batch of samples.

The fRMA package is compatible with all of the affy package commands in bioconductor. The goal of fRMA is to obtain reliable gene-level intensities from the raw microarray data, just as the RMA, but in addition to the raw data, the fRMA algorithm requires a number of frozen parameter vectors. Among these are the reference distribution to which

the data are normalized and the probe-effect estimates. McCall *et al.* have computed these frozen parameters for many popular Affymetrix platforms. The data for each of these platforms is stored in an R package called frmavecs which is built into bioconductor. By default, the frma function attempts to load the appropriate data package for the input data object.

Frozen parameters have already been calculated (by the authors of fRMA) for all GeneChips used in this thesis. All data used in this study was downloaded as raw .CEL files and processed with the fRMA algorithm, unless stated otherwise. An example R-script for the normalisation process is shown in appendix C.

3.2 Summary of datasets

The data used in this thesis is summarised in table 3.1 for mice, and table 3.2 for humans. All mice are C57BL/6 strain, and all human datasets use the Affymetrix Human chips U133 2.0 Plus.

Name of Data Set	**Accession Number**	**Date**	**Type**	**Platform**	**Tissue**	**Description**
Zhang	GSE54652 [73]	2014	Microarray + RNAseq	MoGene 1.0 ST arrays	12 tissues	2hrs for 48 hrs. 1 week LD, then DD. Food & water *ad libitum*. Pooled
Hughes Liver	GSE11923 [10]	2009	Microarray	Affymetrix Mouse Genome 430 2.0 Array	Liver	12/12 LD. Some restricted feeding
Fang	GSE59460 [105]	2014	Microarray	MoGene 1.0 ST	Liver	WT C57BL/6 and Rev-Erb KO. Mice entrained to LD cycle before being euthanised at ZT10
Barclay	GSE33381 [106]	2012	Microarray	MoGene 1.0 ST	Liver and White Fat	Mice entrained to LD cycle for 1 week. Samples collected at CT1, CT7, CT13, CT19 for normal routine and sleep stressed mice from both tissues.
LeMartletot	GSE35789 [107]	2012	Microarray	MoGene 1.0 ST	Liver	Pooled RNA from the whole liver of 5 mice. Samples were collected at times ZT2, ZT6, ZT10, ZT14, ZT18, ZT22, ZT2(+24). Mice were entrained to 12 hour LD cycles for 2 weeks prior.

Table 3.1: **Table summarising the mouse datasets used in this thesis.**

Name of Data Set	Accession Number Reference	Date	Tissue	Category	Description
Bjarnason	unpublished	unpublished	Oral Mucosa	Healthy Timeourse	10 patients samples every 4 hours for 1 day
Sri/UK OM	GSE51010 [108]	2015	Oral Mucosa	Healthy - cross continent	5 UK samples, 3 Sri Lankan Samples. Non smokers.
smoking OM	GSE17913 [109]	2010	Oral Mucosa	Healthy - Non smoker & smoker	40 current smokers and 40 age-and-gender matched never smokers.
Autopsy Data	GSE3526 [110]	2006	Various	Post-mortem (Previously healthy)	Samples taken from 10 individuals up to a maximum of 8.5 hours post-portem. All individuals died due to sudden events and had no known chronic diseases.
Feng	GSE30784 [111]	2008	Oral Mucosa	Healthy, Dysplasic, Cancer	167 OSCC, 17 dysplasia and 45 normal oral tissues from different donors
Richardson	GSE7904 [112]	2006	Breast	Cancer &Healthy	42 Tumour samples and 7 healthy
REMAGUS	GSE26639 [113]	2011	Breast	Cancer	226 Tumour samples with prognostic factors and survival

Table 3.2: **Table of Human Microarray Datasets used in this thesis.** All datasets use Affymetrix HGU133 2.0 plus GeneChip

3.3 Mouse Timecourse

The dataset that inspired the design of Time-Teller was published in November 2014 by John Hoganesch's group in Philadelphia, USA [9]. It is a large-scale, high time resolution, mouse time-course transcriptome dataset, and is referred to in this thesis as the Zhang data. This section starts with a literature review of that study[1], and then novel analysis of the data is presented. The focus will be on the microarray data, with a short section to discuss the RNA-seq data. Additionally, a preceding dataset to the Zhang data, published by Hughes *et al.* [10] is discussed in this section. This timecourse data of mouse liver uses a different GeneChip (Affymetrix MoGene 430 2.0), is compared to the liver timecourse in the Zhang dataset. This is to both show how comparable independent data sets can be, whilst also highlighting the issues faced when comparing datasets.

3.3.1 A circadian gene expression atlas in mammals

Zhang *et al.* [73] quantified the transcriptomes of 12 mouse tissues every 2 hours over 48 hours with microarrays, and every 6 hours using RNA-seq. The tissues sampled were from the adrenal gland, aorta, brainstem, brown fat, cerebellum, heart, hypothalamus, kidney, liver, lung, skeletal tissue, and white fat of the mice.

Experimental setup

6 wk-old male C57/BL6 mice were entrained to a 12h:12h LD schedule for 1 week, and then released into constant darkness (CT0). The mice were provided food and water *ad libitum*. From CT18 post-release, 3 mice were killed every 2 h, for 48 hours (until CT64), and specimens from 12 organs were snap frozen in liquid nitrogen. This is shown in the sketch in figure 3.1.

Microarray methods

RNA was extracted and pooled for 3 mice for each tissue and time point. RNA abundances were quantified using Affymetrix MoGene 1.0 ST arrays using the standard manufacturer's protocol. The results of this form the .CEL files that are used in this thesis, which are accessible via GEO accession number GSE54652.

In their analysis, Zhang *et al.* used the RMA algorithm to normalise the data using the Affymetrix Expression Console Software. As standard, GeneChips were matched to genes and filtered for protein coding status, resulting in 19,788 genes forming the background set. They performed oscillation detection using the R package JTK_CYCLE to fit time series data and detect oscillations on transcripts, using a cutoff significance of $q<0.05$ (5% FDR).

[1] The non-coding RNA sections of the study are omitted from this review.

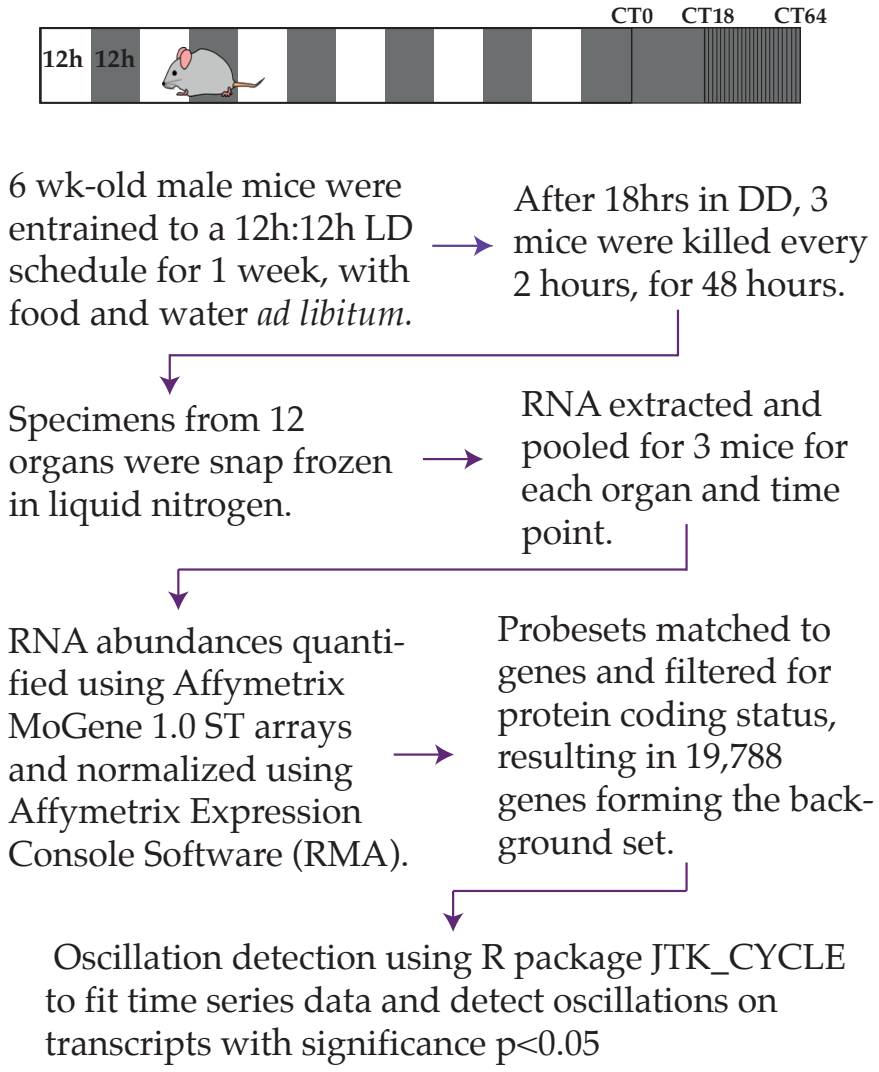


Figure 3.1: **Sketch of the experimental work-flow that produced the Zhang data.**

RNA-seq methods

RNA samples from CT22 - CT64, every 6 hours were pooled. These were converted into Illumina sequencing libraries using Illumina TruSeq Stranded mRNA HT sample Prep kit, and the manufacturer's protocol was used. Libraries were pooled into groups of 6 and sequenced in one Illumina HiSeq 2000 lane by using the 100bp paired end chemistry (16 lanes total). Fastq files containing raw RNA-seq reads were aligned to the mouse genome (mm9/NCBI37) using STAR (default parameters). RNA quantification was performed with HTSeq, in Stranded Mode (default parameters). Protein coding genes quantified using ENSEMBL annotation. Quantification values were normalised using DESeq2. These final values are what is used in this thesis.

Zhang rhythmicity analysis results

Zhang *et al.* [73] define a circadian gene to be one that has a significant ∼24 hour period. They detected rhythms, and assigned phases to each rhythmic gene using JTK_CYCLE, a non-parametric algorithm published by Hughes et al. [114] as an efficient way of detecting rhythmic components in genome-scale datasets. JTK_CYCLE will be discussed further in section 3.3.2, but for now can be understood to be an algorithm that gives a confidence measure that a timecourse is periodic, and provides estimations of the period and phase of transcripts that have a false discovery rate (FDR) $q < 0.05$.

Zhang *et al.* report that the liver has the most circadian genes with 3,186. The three brain regions had the fewest circadian genes, with the hypothalmus having the least with 642. They hypothesise some reasons for this; the sampling of specific brain regions is difficult, the brain has a heterogeneous mixture of cell types expressing different sets of genes, or the different cell types could be out of phase with each other.

A major finding of this study is that 43% of protein coding genes were detected to have a circadian rhythm in at least one of the 12 sampled organs. They also conclude that most circadian expression is organ-specific. By extrapolating the number of circadian genes across organs, they estimate that around 55% of all protein coding genes are circadian in at least one organ in the whole mouse.

They found that 1400 genes were phase-shifted with respect to themselves by at least 6 hours between 2 organs, with 131 genes completely anti-phased. Due to this they draw the conclusion that these are "clock controlled genes" with organ specific phases.

Their analysis found 10 genes that oscillated (q< 0.05) in all 12 tissues: *Bmal1 (Arntl)*, *Dbp*, *Rev-Erbα (Nr1d1)*, *Rev-Erbβ (Nr1d2)*, *Per1*, *Per2*, *Per3*, *Usp2*, *Tsc22d3*, and *Tspan4*. Most of these genes are well known to be part of the central oscillator system.

Zhang *et al.* found that the core clock genes oscillated with the peak phases of a given gene falling within 3h of each other across all organs. Figure 3.2 illustrates the synchronised expression profiles of some of the core clock genes across all tissues. They also included *Cry1* and *Cry2* in their "core clock gene" set. This appears to be due to the expectation that these genes are part of the core clock, and not because of the results of their data analysis. Figure 3.2 also shows some possible interactions amongst these genes.

Comments are not made on the expression synchrony of the genes that are found to oscillate in all organs, but are not thought to be part of the core clock setup, these being *Usp2*, *Tsc22d3*, and *Tspan4*. USP2 plays a significant role in modulating proinflammatory cytokine induction, and has been reported to be required for TNFα induced NF-κB signalling [115]. As Usp2 mRNA presents highly circadian and synchronised circadian rhythms, it could be an interesting candidate to study for the interaction between the cir-

cadian regulation of inflammatory pathways. TSC22D3 is also involved in the regulation of inflammatory pathways.

TSPAN4 is a cell surface protein that regulates cell motility [14]. Although it has often been observed to have circadian behaviour, there is a lack of literature on the circadian expression of Tspan4.

The heatmap in figure 3.2 shows that Arntl and Clock are approximately anti-phased to the other genes. Zhang *et al.* reported that the expression of oscillating genes peaks during transcriptional "rush hours" before dawn and dusk.

Zhang *et al.* report that the majority of best-selling drugs directly target the products of rhythmic genes, that these drugs have short half lives, and most relevantly to this thesis, that they may benefit from timed dosage.

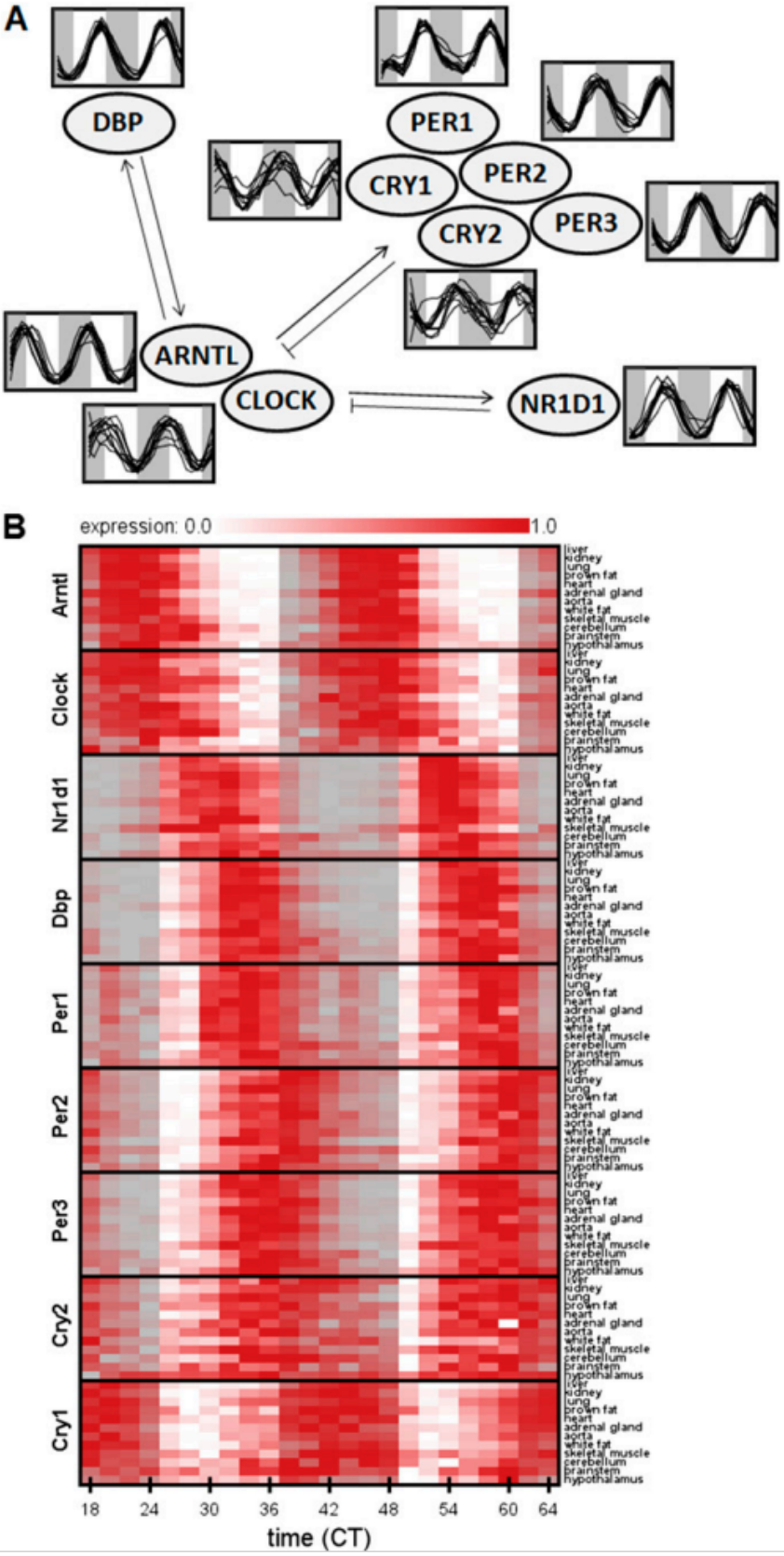


Figure 3.2: **Expression of core circadian oscillator genes across organs.** From [9], the figure emphasises the synchronised expression of this set of circadian genes. (A) Expression of each gene in all organs normalised and superimposed. Arrows indicate likely gene interactions. (B) Heatmap representation of normalised expression from A.

3.3.2 Novel rhythmicity and synchronicity analysis of zhang data

Data preparation

In order to carry out a new analysis, the raw data was downloaded from NCBI GEO in the form of 288 .CEL files. The bioconductor package in R was used to perform fRMA normalisation of protein coding genes, and annotate them with gene names. After fRMA processing, the gene expression values are expressed in $\log_2$, and values are in the range 2-14. Figure 3.3 shows the distribution of gene expression values for all 288 sets, summarising 35,556 probes (includes all probes, not just those for protein coding genes). The similar distribution amongst samples indicates that the fRMA normalisation was successful.

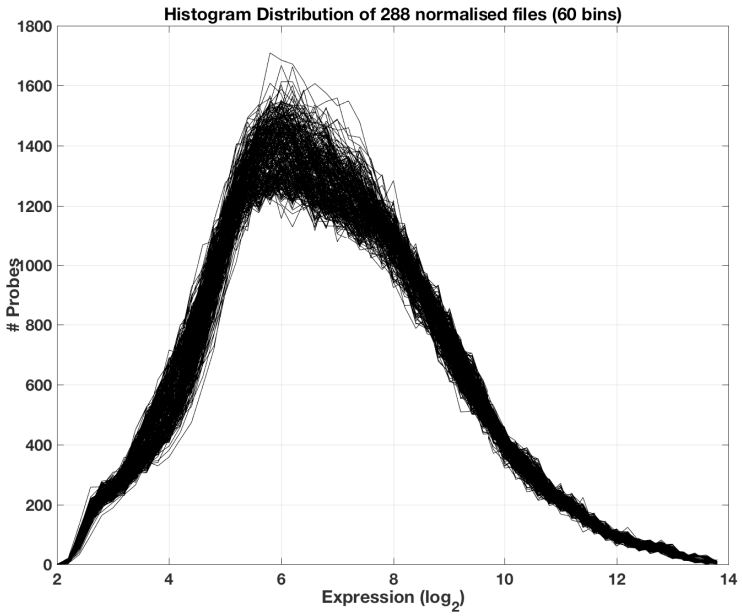


Figure 3.3: **Histogram showing the distribution of gene expression values after fRMA analysis of the Zhang data.** The agreement in expression shows that the fRMA properly normalised the samples.

The resulting data is imported from R into MATLAB and reorganised as 3 dimensional structure of 35,556 probes, 12 organs, and 24 time points.

Existing rhythmicity detection algorithms

Rhythmicity detection is an important data analysis method for circadian data, in order to identify cycling transcripts. Here we discuss the most popular methods of circadian rhythm detection in transcriptome data. A good review is written is by Wu eta al [116]. Note that all rhythmicity detection methods (except for JTK_CYCLE) are applied to

normalised data.

The most conceptually basic method of rhythmicity detection is **COSINOR** analysis. COSINOR regression is simply a least squares regression for fitting a fixed phase length sine wave to sparse data points from a fixed (short) time interval. The algorithm provides a measure of confidence of the fit, but prior information is needed - usually that we expect a 24hr rhythm [117].

COSOPT is a similar method to COSINOR, except that instead of an iterative regression fit, COSOPT measures the goodness-of-fit between experimental data and a series of cosine curves of varying phases and period lengths [118]. p-values (known as pMMC-β values) are calculated by scrambling the experimental data and re-fitting it to cosine curves in order to determine the probability that the observed data matches a cosine curve by chance alone [119].

JTK_CYCLE is a non-parametric test procedure that was designed specifically to detect cycling transcripts [114]. In addition to providing optimal phase, amplitude, and period estimates for each transcript, JTK also outputs permutation-based p-values and Benjamini-Hochberg q-values. The authors claim that compared to other cycling tests, JTK has advantages in the statistical power of its p-value assignments, improved resistance to outliers, as well as relative computational efficiency. JTK_CYCLE is publicly available to download as an R package[2]. All reported JTK values in this thesis have been produced using this package.

A good review of other rhythmicity detection algorithms for genome data is found in Wu *et al* [116]. JTK_CYCLE and cosine fitting algorithms work well to detect rhythmicity in sufficiently large transcriptome timecourse datasets. However, JTK_CYCLE needs a sufficient number of data points in an array (multiple points per time for low resolution datasets, or sufficiently high resolution within one period). The choice of rhythmicity detection method should, and usually are, based on the specific goal and experimental design. Each method has different strengths, speeds, and conditions on the data to ensure adequate performance.

Zhang's rhythmicity detection

In their study, Zhang *et al* [9] define a circadian gene to be any gene that achieves a JTK false discovery rate less than 0.05 ($q < 0.05$) in at least one organ. They calculate that the liver has 3,186 genes that satisfy this threshold, which means that they label 16% of

[2] `https://github.com/mfcovington/jtk-cycle`

all liver genes as circadian. We suggest that this binary way of classifying "circadian" $q < 0.05$, and "not circadian" $q > 0.05$ results in some lost information. For example, 676 of these liver genes satisfy the criteria that $q < 0.0005$, so there is probably a lot of important information in the very small FDR values that is not being used. The very small FDR values are calculated for data that look almost like perfect sine waves (see figure 3.2), and we suggest that this should be recognised as far more significant than the genes which might show some noisy periodic behaviour.

Zhang *et al.* attempt to answer the following questions using their data:

- How many genes oscillate with 24 hour rhythms for each organ? This is answered in their study using numbers of genes that satisfy the FDR threshold $q < 0.05$.
- What genes oscillate with 24 hour rhythm in all organs? This answered in their study using the FDR threshold, and combining "hits".
- What genes oscillate with 24 hour rhythm in all organs with similar phase? This answered in their study using JTK cut-off values, combining "hits" and comparing variation of calculated phases.

This final question could be answered in a much more robust way, if a different type of analysis were used. Here we present a novel analysis that allows the answering of the questions:

- What genes have **synchronised** expression across organs?
- What genes have the most synchronised behaviour across organs, with a 24 hour period?

To explain the motivation for this type of analysis, we look at the phases calculated by Zhang *et al.* with JTK_CYCLE. 13 of the core clock genes are represented on the circular clock sketch in figure 3.4. The number around the outside represents the CT time, the mean phase of each clock gene across the 12 organs is shown with a bold coloured line, and each gene is represented by a different colour. The agreement of these phases are obvious from the figure, and this is an extremely significant result. We suggest that the synchronisation of the circadian rhythms across organs should be the hypothesis to be tested in the analysis, and not just an interesting coincidental result.

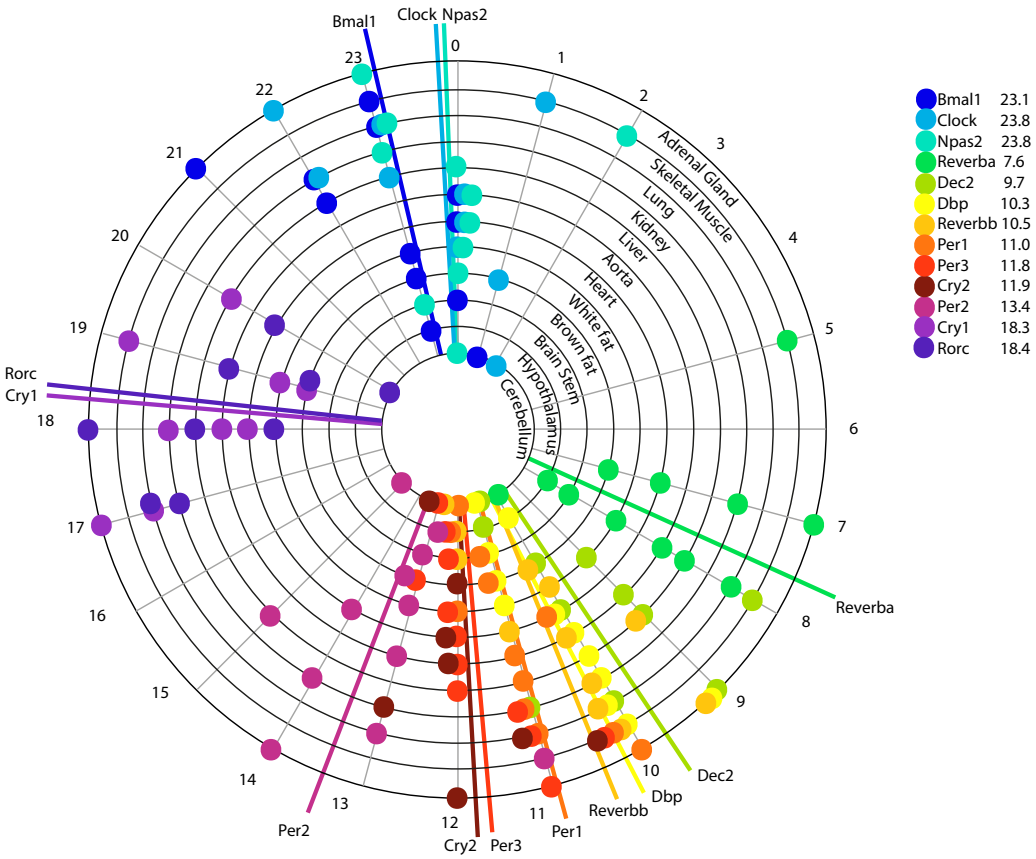


Figure 3.4: **Circular plot showing mean phase of 13 core clock genes.** The data used was for the 12 organs in the Zhang dataset. The clusters of transcriptional activity, and the synchronised phases of the genes across organs, are apparent. The phases of Bmal1, Clock, and Npas2 occur around ZT23-24 (equivalent to 11pm-midnight). These are genes with RRE motifs in their promoter regions. The phases of both Rorγ and Cry1 are later than the transcriptional rush hour that contains the other circadian genes.

Synchronicity analysis using linear algebra

The singular value decomposition (SVD) is a matrix decomposition technique, which is a precursor step to the well known principle component analysis (PCA). It allows the identification of the axes of maximum variance (the eigenvectors or principal component axes), and the calculation of how much of the variation in the data can be explained by each principal component (by the singular values).

Here we give a brief explanation of SVD, but a more detailed summary is given in appendix A, where SVD's association to PCA is also explained.

The single value decomposition of the $m \times n$ matrix A, for m observations and and n features, and rank r, is defined as:

$$A = U\Sigma V^T = \mathbf{u_1}\sigma_1 \mathbf{v_1}^T + ... + \mathbf{u_r}\sigma_r \mathbf{v_r}^T \tag{3.1}$$

where

- $\mathbf{u}_i$ are the columns of $U^{(m\times m)}$ and are the orthonormal eigenvectors of the covariance matrix AA^T and the left singular vectors of A
- $\mathbf{v}_i$ are the columns of $V^{(n\times n)}$ and are the orthonormal eigenvectors of the covariance matrix A^TA and the right singular vectors of A
- $\sigma_i = \sqrt{\lambda_i} = \|A\mathbf{v_i}\|$ are the lengths of the vectors $A\mathbf{v_i}$
- $\Sigma^{(m\times n)}$ is made up of a diagonal matrix of σ_i's (where $\sigma_1 > \sigma_2 > ... > \sigma_r > 0$) in its upper left position, with 0's everywhere else.

$\mathbf{v}_i$ are known as the **principal components** and σ_i are known as the **singular values**.

Normalisation

A crucial step to prepare the data for SVD is to normalise A so that each observation has equal weighting in the finding of the principal components. In the circadian data, this ensures that the different amplitudes of oscillation for different organs do not introduce a bias to the results. To normalise, each of the 35,556 probes is organised into a matrix, A, of S rows (#organs), and T columns (#timepoints). Each row is then normalised so it has a mean of 0 and a standard deviation of 1.

SVD of Zhang data

After fRMA processing, the resulting 35,556 probe values for eight[3] organs at 24 time-points were saved in a .txt file in R and imported into MATLAB. The data was then structured into 35,556 matrices of size 24 (time) by 8 (organ), one for each probe, and normalised as above. SVD was performed on each matrix using MATLAB's inbuilt SVD algorithm. In this case, the principal components correspond to time-series with the first PC providing the dominant temporal shape found across the various organs and the other principal components describing how this is modified across organs in a graded way. It is important to note that SVD is computationally inexpensive and this SVD takes less than 4 seconds to complete for all data[4]. Each matrix (representing each probe), is then represented by singular values and principal components. The measure for how much the first principal component $\mathbf{v}_i$ represents all of the data is calculated with

$$\% = \frac{\sigma_1^2}{\sum_{i=1}^{r} \sigma_i^2} \times 100 \tag{3.2}$$

The genes with the top 20 largest % variance explained by the first PC are shown in (non-normalised) timecourse expression in figure 3.5. The reader should remember when looking at the timecourses in this figure, that these results have been generated by nothing but simple linear algebra, and no cycling detection algorithms have been used. Despite this, 14 of the 20 top synchronised probes are known to be related to the circadian clock. Amongst these genes are Bmal1, Npas2, Per1, Per2, Per3, Rev-Erbα, Rev-Erbβ, Dbp, Ciart, Wee1, Tef, and Nfil3.

Probe 10556487 is highly circadian, but is unlabelled. This appears to be a unique protein coding gene, which can be labelled with BC150970, A630005I04Rik or RIKEN cDNA A630005I04 gene, but does not have functional annotations. This gene contains 4 canonical E-box regions "CAGCTG", which may explain its tight circadian control. Further investigation into this gene may provide more insight into the mouse circadian core clock mechanism.

The robustness of rhythms detected in this study results from the combination of robustness of the clock and also from the experimental and sampling protocol, which is optimised to reveal circadian rhythms, so is unlikely to reveal harmonic rhythms. Four of the synchronised probes do not have 24-hour periods, and one might speculate that these are identified 8 hour harmonics, but it could also be the result of something like fluctuating conditions in the lab during the experiment (although this is just speculation). Three of these (10584580, 1058457 and 10584576) are probes for the Hspa8 gene. Hspa8 is a gene for a heat shock protein, and here it shows very pronounced synchronised rhythms.

[3]Brain and Whitefat data are excluded as the noise to amplitude ratios are higher. Exclusion of this data is ok as it was pre-nomalised by fRMA so does not affect the batch statistics.

[4]4 seconds on a laptop with 16GB RAM and 2.5 GHz Intel Core processor.

Similarly, Erdr1 10608711 is a protein associated with stress.

Nonetheless, it is very interesting to see such obvious synchronicity in the expression of genes across multiple organs in mice over time, circadian or otherwise.

Combination analysis

Each rhythmicity detection algorithm provides a slightly different set of "rhythmic" probes because they use different techniques. If we want to really understand the data, using a few methods in combination can only reinforce our conclusions and learnings from the data. There is no reason to choose just one, and justify the choice, as many studies do.

We re-analysed the data with JTK_CYCLE. Zhang *et al.* used JTK_CYCLE on linear values, whereas this reanalysis uses log_2 values. Such transformations are often useful in statistical data analysis. The difference in using log_2 data is that higher expressed clock genes need to have high amplitude rhythm to be significant, but lower expressed genes with a rhythm need less amplitude to be significant. This decision was made as only phase and period, and not amplitude, are the characteristics we are interested in. An example for why this decision was made, is shown in figure 3.6. The Gm129 (also known as Chrono or Ciart) gene has two probes that show obvious circadian behaviour, but due to low expression values, the JTK FDR value is not significant[5].

To run the JTK algorithm, the parameters were set at 10-14 for estimated period (as data is 2 hour resolution), and time step was set to 2 (hours). Other than this, all parameters are kept as the default values, as suggested in the user manual published with the JTK_CYCLE algorithm. Figure 3.7 shows the (geometric mean) JTK FDR plotted against the % variance explained by the first principal component, for each probe. An arbitrary cut off of the top 50 genes for each metric has been used for labelling purposes, and the "NA" or duplicate gene names are omitted in the top-right and bottom left quadrants. 26 of the top 50 genes are common amongst these two metrics. The bottom left quadrant represents genes that are 24 hour rhythmic, but not in synchrony across organs. The top right quadrant are the genes that are synchronised in expression, but not in a 24 hour rhythmic pattern.

Minor harmonics are not detected by JTK_CYCLE and are treated as noise. Due to this, Per1 (with a small second harmonic) is not reported by JTK to be significantly rhythmic, but it is still quite obviously circadian (figure 3.5 shows this). Per1 is reported as "oscillating in all organs" in the study, but reanalysis (and also their own published values on the circa website) shows FDR rates >0.05 in multiple organs.

[5]This information can be found on the web database `http://circadb.hogeneschlab.org/`.

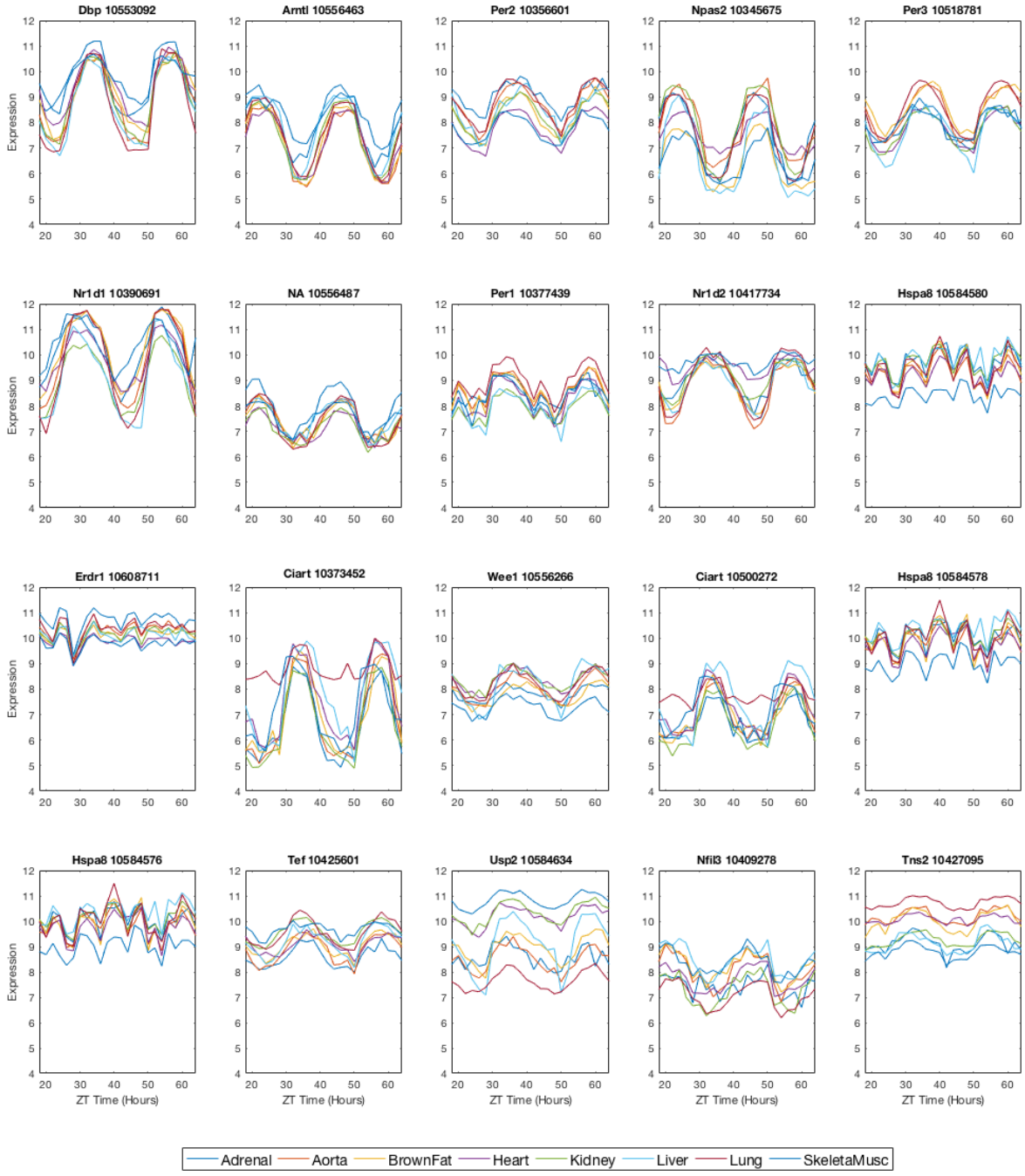
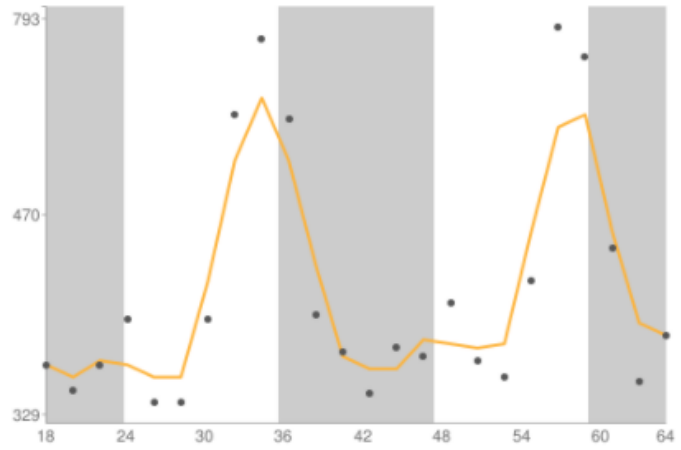


Figure 3.5: **Timecourse expression of the probes in the Zhang data with the top 20 highest % variance explained by the 1st PC**. 16 of the 20 probes are obviously circadian.



Experiment Mouse 1.OST Lung (Affymetrix)

UCSC RNAseq RNAseq_NM_001033302

Probeset 10373452

Links Wikipedia HomoloGene

Symbol Gm129

Unigene Mm.331218

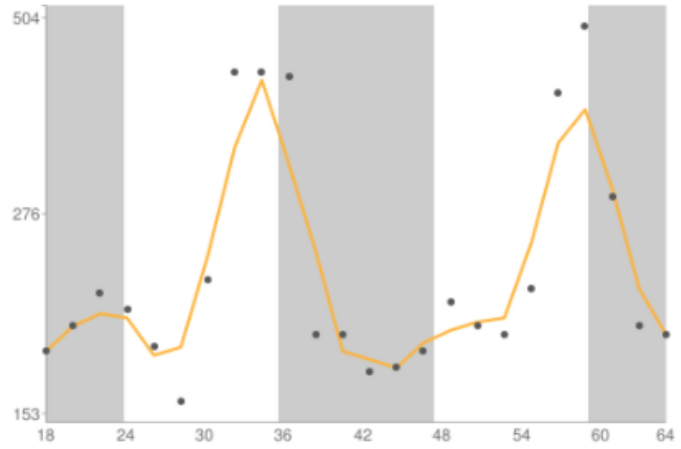
RefSeq Protein *None*

RefSeq DNA NM_001033302

Description predicted gene 129

	p-Value	q-Value	period	phase
JTK	0.0162831	0.139711	24.0	10.0
Lomb Scargle	0.0285	0.513346	22.291	10.1
DeLichtenberg	0.7291	1.0	24.0	*NA*

back to the top



Experiment Mouse 1.OST Lung (Affymetrix)

UCSC RNAseq RNAseq_NM_001033302

Probeset 10500272

Links Wikipedia HomoloGene

Symbol Gm129

Unigene Mm.331218

RefSeq Protein *None*

RefSeq DNA NM_001033302

Description predicted gene 129

	p-Value	q-Value	period	phase
JTK	0.0228067	0.17694	26.0	8.0
Lomb Scargle	0.032	0.527733	22.291	9.7
DeLichtenberg	0.5505	1.0	24.0	*NA*

back to the top

Figure 3.6: **Screenshot of a circadb query of gm129 in the lung, for the Zhang data.** The q-values are insignificant even though the 24 hour rhythm is quite clear.

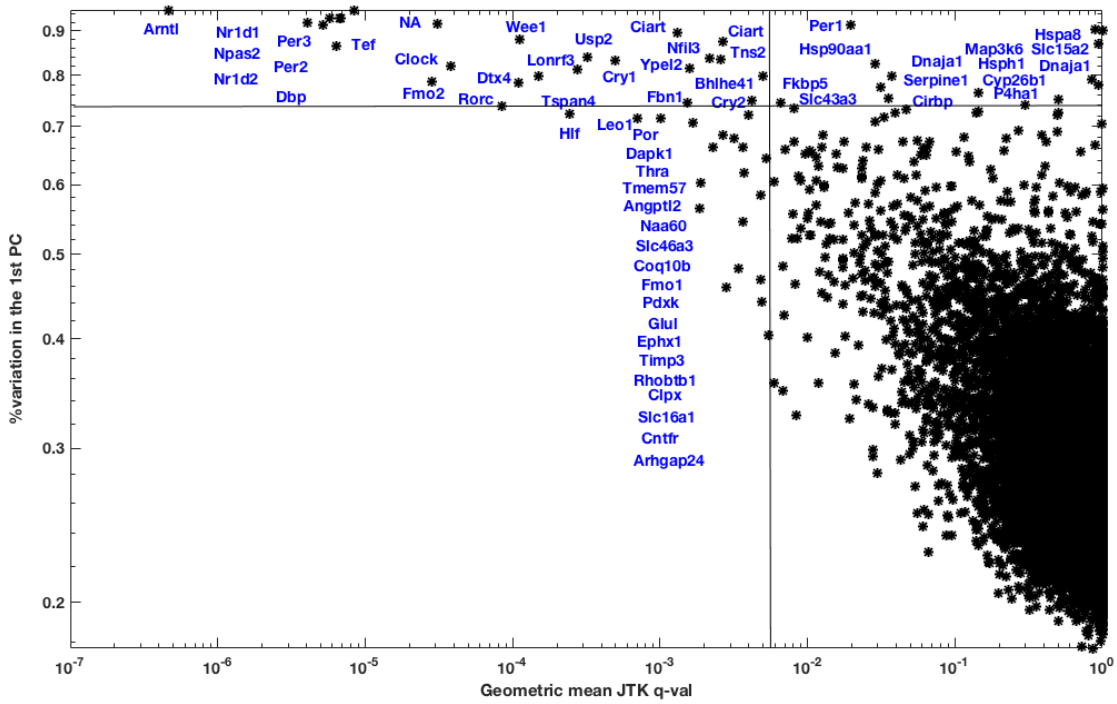


Figure 3.7: **Scatter plot of singular values vs geometric mean JTK values for the Zhang data.** The top 50 probes for each metric are presented, where 26 of the 50 are shared.

A COSINOR analysis was carried out using the MATLAB function cosinor[6] for each normalised organ timecourse (and the geometric mean across organs found). MATLAB's cosinor function returns a false positive metric in terms of p-value.

The geometric mean of these COSINOR false positive p values is plotted against the first singular value in figure 3.8. The first 50 for each metric are labelled, where "NA" or duplicate gene names are omitted in the top-right and bottom left quadrants. 30 of the top probes are shared.

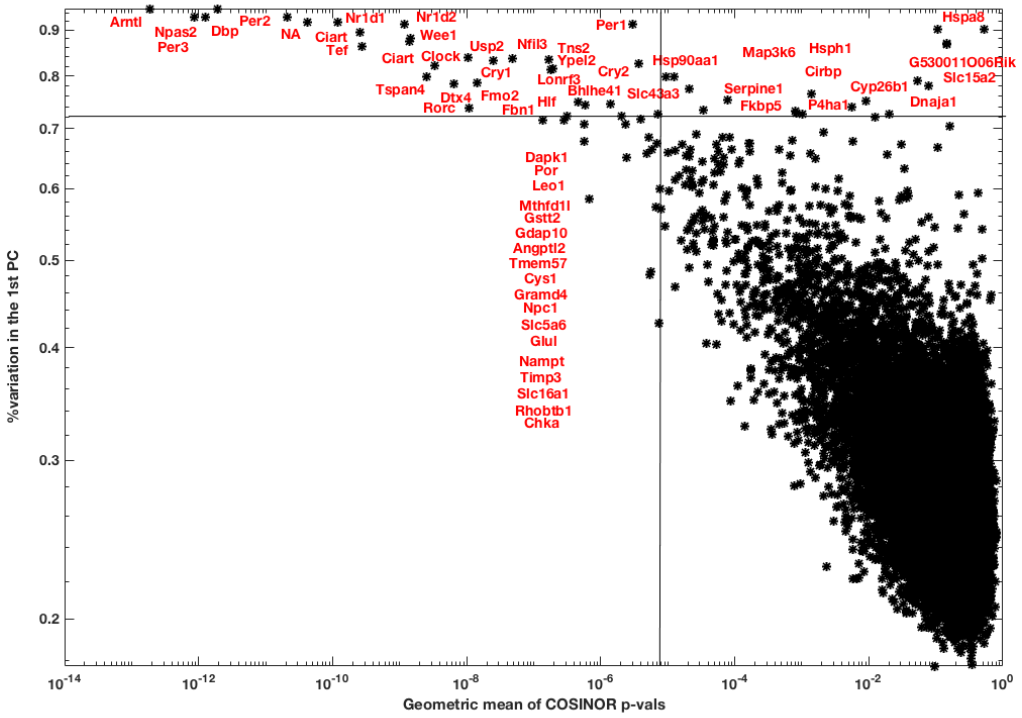


Figure 3.8: **Scatter plot of singular values vs geometric mean cosinor p-values for the Zhang data.** The top 50 probes for each metric are presented, where 30 of the 50 are shared.

The conclusion of these analyses is that no one method is best for the data analysis. If multiple methods, with different strengths are at hand, then performing all methods will ultimately provide most insight into the data. Probes that are ranked highly with some methods and lower with others are often interesting outliers, and not just to be ignored.

JTK_CYCLE can better determine oscillations that are not perfect sine curves than COSINOR can. Using SVD to detect synchronised rhythms could provide false positives as it can highly rank non-circadian synchronised genes. Hence using JTK_CYCLE or COSINOR in parellel helps to identify these false positives, and also provide information on rhythmic, but unsynchronised probes. Hlf, Leo, Dapk1, Por and Tmem57 appear in

[6]Details of cosinor can be found in appendix C.

Genename	Probe	% Var	GeoMean Cosinor	Geomean JTK q vals	Sing rank	Cosinor rank	JTK Rank	Overall rank
Arntl	10556463	0.95	1.90E-13	4.65E-07	49	50	50	149
Npas2	10345675	0.93	8.67E-13	6.77E-06	47	49	45	141
Per3	10518781	0.93	1.25E-12	5.77E-06	46	48	47	141
Dbp	10553092	0.95	1.95E-12	8.48E-06	50	47	43	140
Per2	10356601	0.93	2.12E-11	6.83E-06	48	46	44	138
Nr1d1	10390691	0.92	1.18E-10	4.07E-06	45	44	49	138
Nr1d2	10417734	0.91	1.18E-09	5.15E-06	42	41	48	131
NA	10556487	0.92	4.16E-11	3.10E-05	44	45	41	130
Tef	10425601	0.86	2.72E-10	6.34E-06	34	42	46	122
Wee1	10556266	0.88	1.47E-09	1.13E-04	38	39	37	114
Ciart	10373452	0.90	2.59E-10	1.32E-03	39	43	29	111
Clock	10530733	0.82	3.37E-09	3.79E-05	28	37	40	105
Usp2	10584634	0.84	1.05E-08	3.22E-04	33	35	33	101
Tspan4	10558961	0.80	2.51E-09	1.50E-04	24	38	36	98
Ciart	10500272	0.87	1.39E-09	2.66E-03	37	40	20	97
Fmo2	10359582	0.79	1.44E-08	2.85E-05	21	33	42	96
Dtx4	10466304	0.78	6.36E-09	1.11E-04	20	36	38	94
Cry1	10371400	0.83	2.46E-08	5.00E-04	30	32	32	94
Lonrf3	10599192	0.81	1.79E-07	2.76E-04	26	28	34	88
Nfil3	10409278	0.84	4.90E-08	2.16E-03	32	31	23	86
Rorc	10494023	0.74	1.09E-08	8.48E-05	10	34	39	83
Tns2	10427095	0.83	1.66E-07	2.58E-03	31	29	21	81
Ypel2	10389581	0.82	1.95E-07	1.59E-03	27	27	27	81
Hlf	10389786	0.72	3.17E-07	2.42E-04	3	25	35	63
Leo1	10587211	0.71	1.38E-07	7.07E-04		30	31	61
Fbn1	10487040	0.74	1.42E-06	1.54E-03	13	19	28	60
Per1	10377439	0.91	3.00E-06	1.97E-02	43	15		58
Dapk1	10405693	0.71	2.82E-07	1.01E-03		26	30	56
Bhlhe41	10549276	0.75	4.60E-07	4.22E-03	14	24	11	49
Por	10526363	0.71	5.67E-07	1.67E-03		22	26	48

Table 3.3: **Table of ranked circadian genes, combined results of data analysis methods for circadian behaviour.**

both COSINOR and JTK, but not in singular value ranking. This implies that these genes have different phases in different organs, so are likely to be clock controlled genes that have slightly different activation mechanisms in each organ.

These three metrics allow us to confidently choose a set of probes to make up the training set for the time-telling model. A ranking is done for each method, assigning 1-50 for each probe, and then simply summing these rankings to combine results. This is shown in table 3.3.

Seventeen of these probes match with the genes shown in table 1.1, which was written from a literature search. Genes in this table which do not appear to be highly rhythmic and synchronised in the Zhang data are Rorα, Rorβ, Arntl2, and the Casein kinases, all with insignificant results in all metrics. Cry2 has an overall rank of 34, and is found to be significantly circadian in all metrics, but just falls off the end of table 1.1.

The results of this section suggests that the core elements of circadian clock are very much synchronised across different organs in laboratory mice. This provides further evidence for how robust the behaviour of the core machinery of the circadian clock is, and also that the core genes of the circadian clock are synchronised across tissues.

The top 11 rhythmic and synchronised genes are selected for the training set of the mouse Time-Teller: Arntl (10556463), Npas2 (10345675), Clock (10530733), Nr1d1 (10390691), Nr1d2 (10417734), Per2 (10356601), Per3 (10518781), Ciart (10373452), Dbp (10553092), Tef (10425601), and Wee1 (10556266).

3.3.3 RNA-seq timecourse

The purpose of this section is to show how the RNA-seq data (that will inevitably become the main transcriptome technology in the future) will be able to make up a training set in the same way as the microarray data.

The Zhang [73] RNA-seq data has a 6 hour resolution over 24 hours, for all organs sampled for microarray. Due to the low resolution of this data, it would not be an effective training set for a Time-Teller model. As we have both the RNA-seq and microarray analysis of exactly the same biological samples, we can attempt to compare them. This was not done in Zhang *et al.*'s study. It is an interesting question in itself to compare such data. This data is likely to be very unique in that it consists of 12 organ timecourses where there is both microarray and RNA-seq data from the same samples. There are studies that try to compare microarray and RNA-seq technologies [120], but circadian timecourse data allows a unique type of analysis in that the normalisation of the data can lead to an almost exact match of expression profile.

The RNA-seq data was already processed by Zhang *et al.* [9], and was downloaded from GEO in this post-processed format. The RNA-seq protocol and processing procedure were outlined in the previous section. Ensembl (Biomart) provides a database for comparison of RNA-seq to microarray data. Parameters chosen were *Ensembl Genes 89*, *Mouse genes (GRCm38p5)*, External Attribute *AFFY MoGene 1 0 st v1 probe*, where gene name is also provided. The resulting data file consists of 170,993 rows (92871 unique). Multiple Affymetrix probes can be mapped to the same Ensemble ID, and multiple Ensemble IDs can be mapped to the same Affymetrix probe. All rhythmic genes used here are a one-to-one mapping. In the Affy MoGene 1.0 ST GeneChip, Ciart has 2 probes that are both highly rhythmic. Only 10500272 is used here, mapped to Ensembl ENSMUSG00000038550. The matched RNA-seq and microarray annotations for 10 clock genes are shown in table 3.4.

ENSEMBL ID	Affy ID	Genename
ENSMUSG00000059824	10553092	Dbp
ENSMUSG00000055116	10556463	Arntl
ENSMUSG00000020889	10390691	Nr1d1
ENSMUSG00000021775	10417734	Nr1d2
ENSMUSG00000028957	10518781	Per3
ENSMUSG00000055866	10356601	Per2
ENSMUSG00000020893	10377439	Per1
ENSMUSG00000038550	10500272	Ciart
ENSMUSG00000029238	10530733	Clock
ENSMUSG00000026077	10345675	Npas2

Table 3.4: **Table showing corresponding ENSEMBLE ids for Affymetrix AFFY MoGene 1.0 probes**. 10 circadian genes are shown for example.

The RNA-seq data was transformed to log (base 2), in order to compare with the microarray data. The microarry and RNA-seq data for the 10 genes in table 3.4 are shown in figure 3.9, where each organ is coloured differently. It is clear that the expression using the RNA-seq technology for Nr1d2, Per1, Per2, and Per3 is higher than for the microarray, but the agreement is generally good. The dynamic range of microarrays may be limited compared to RNA-seq, resulting in compression of amplitudes.

Each timecourse was normalised to have a mean of 0 and standard deviation of 1, and this is shown in figure 3.10 where the normalised microarray data are shown with blue lines and normalised RNA-seq data are shown with red stars.

The RNA-seq data appears to have an almost identical profile to the microarray data. Both shapes of data, and the changes in variance over time, appear to match very well. This suggests that when high resolution timecourse RNA-seq data sets are available in the future, Time-Teller can be used in the same way as microarray data.

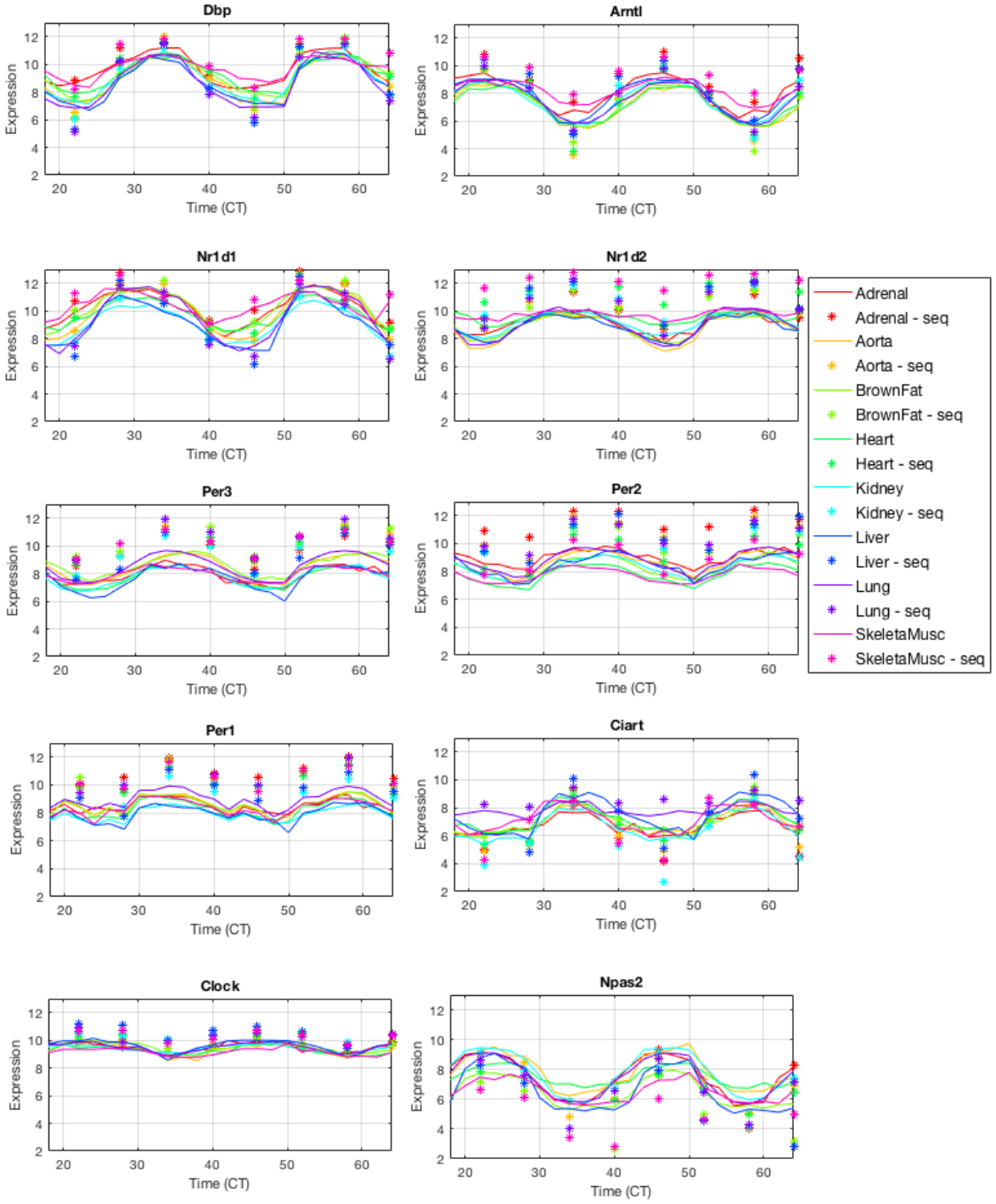


Figure 3.9: **Plots showing microarray and RNA-seq data superimposed.** The line plots show timecourse microarray, and stars stars show timecourse RNA-seq data. Data shown in for the top 10 circadian genes. Data for 8 organs are is shown here (no brain or white fat data).

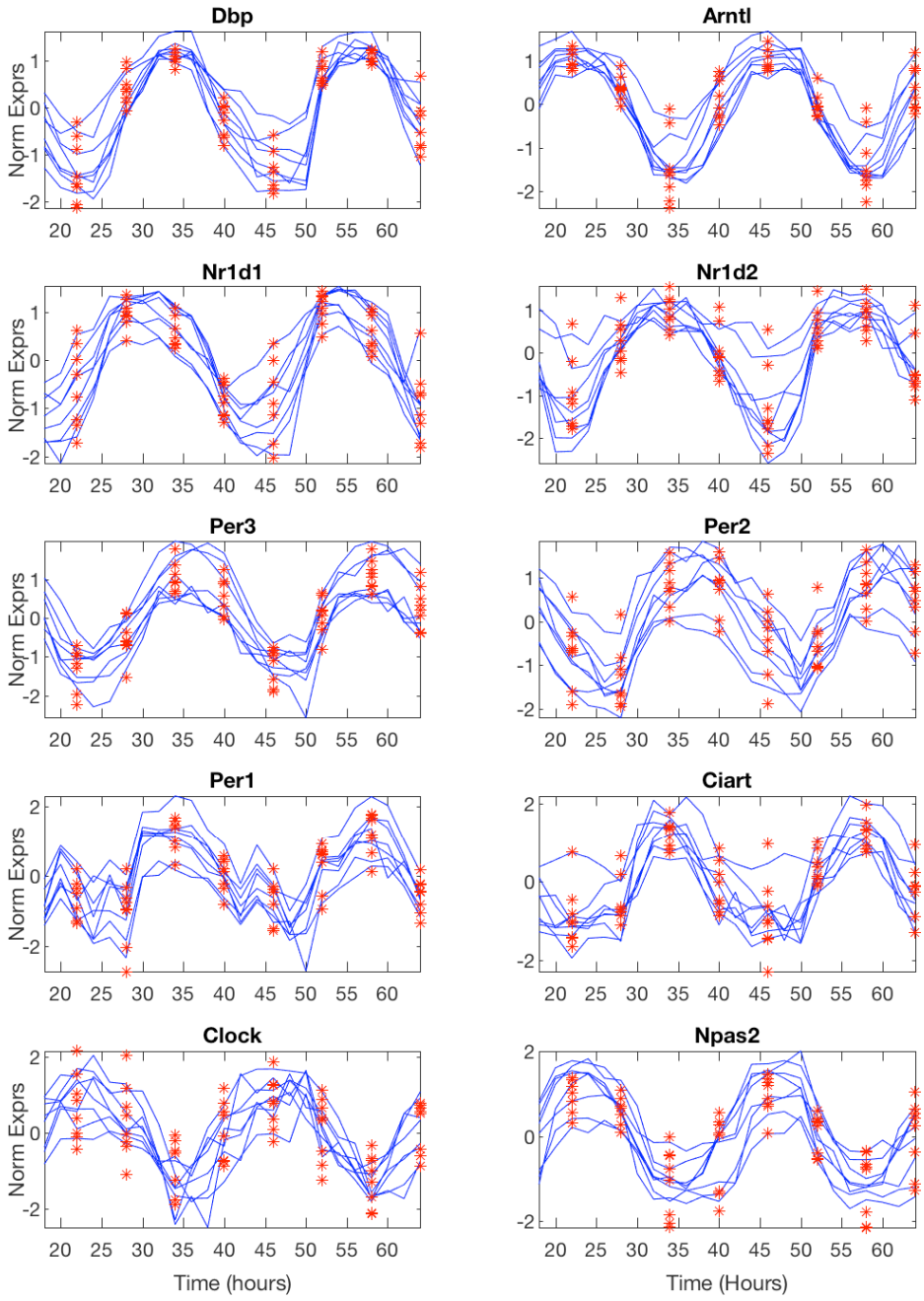


Figure 3.10: **Plots showing normalised microarry and RNA-seq data superimposed.** The blue lines show normalised timecourse microarray, and red stars show normalised timecourse RNA-seq data. Data shown is for the top 10 synchronised circadian genes. Data for 8 organs is shown here (no brain or white fat data).

3.3.4 Comparison of different Affymetrix GeneChips

The purpose of this section is two-fold: the first is to show the reader the difficulties in comparing different GeneChips, and the second is to show how the robustness of the circadian clocks allows the comparison of independent datasets.

Hughes and Hoganesch [10] published a dataset of mouse liver timecourse at 1 hour resolution over 48 hours, in a study preceding the Zhang data by 5 years. The Affymetrix Mouse Genome 430 2.0 GeneChip was used, which is different to the GeneChip of the Zhang data.

Method

The experimental setup of Hughes *et al.*'s study was very similar to that of the Zhang study, except for a 1 hour sampling of only the liver. Mice were entrained to a 12 hour LD environment before being released to constant darkness. Starting 18 hours after the first subjective day (CT18), liver samples from 3-5 mice per time point were collected every hour for 48 hours.

Comparing Hughes and Zhang data

The Affymetrix website (Thermofisher website after 2017) provides plentiful information and data on which probes from each GeneChip best match probes from other GeneChips. The best matches for Affymetrix Mogene 1.0 ST and Affymetrix 420 2.0 are listed in table 3.5 for 12 clock genes.

Genename	Mogene 1.0 ST	Affy 430 2.0
Arntl	10556463	1425099_a_at
Npas2	10345675	1421036_at
Dbp	10553092	1418174_at
Per3	10518781	1460662_at
Nr1d1	10390691	1426464_at
Per2	10356601	1417602_at
Nr1d2	10417734	1416958_at
Tef	10425601	1450184_s_at
Ciart	10373452	1435188_at
Wee1	10556266	1416773_at
Clock	10530733	1418659_at
Usp2	10584634	1417168_a_at

Table 3.5: **Table showing the best matched probes for two different Affymetrix mouse GeneChips for 12 clock genes.**

After fRMA normalisation, the Hughes liver data was plotted on top of the Zhang liver data. The differences in mean intensity and amplitude in most probes is apparent in figure 3.11. Even though the probes are measuring the expression of the same gene, there

will be some differences in signal strength, affinity, etc. It is important to notice that the difference also changes for each probe. The problem that this poses for Time-Teller will become apparent in the next chapter.

It is, however, clear that the shapes of the expression profiles are very similar. This is remarkably clear in the normalised data in figure 3.11. Not only are the phases aligned, but the shapes of expression are almost perfectly matched. It must be emphasised here that these data sets were created around 5 years apart. The same LD schedule was used, and same strain of mouse, but there was a slight difference in feeding amongst many other minor differences in experiment work flow. This similarity in waveform across heterogeneous datasets has been noticed before in plant clock data sets [121]. The reproducibility of circadian expression profiles provides even more evidence towards the robustness of the circadian clock, and that there is a robust and reproducible time fingerprint in the transcriptome.

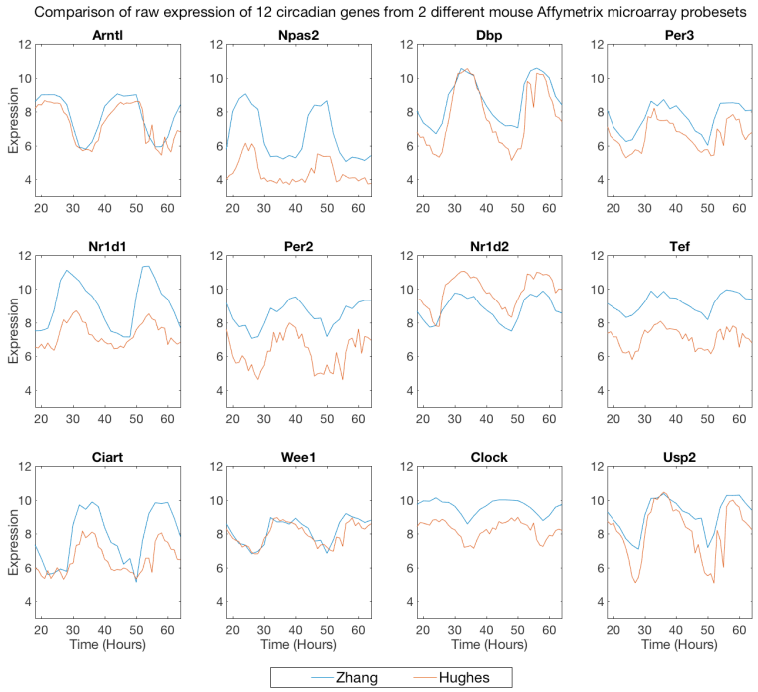


Figure 3.11: **Plot showing circadian timecourse data from two independent datasets.** Comparison of raw expression in the liver of mice [10], from two different experiments performed over two days, using two different affymetrix GeneChips.

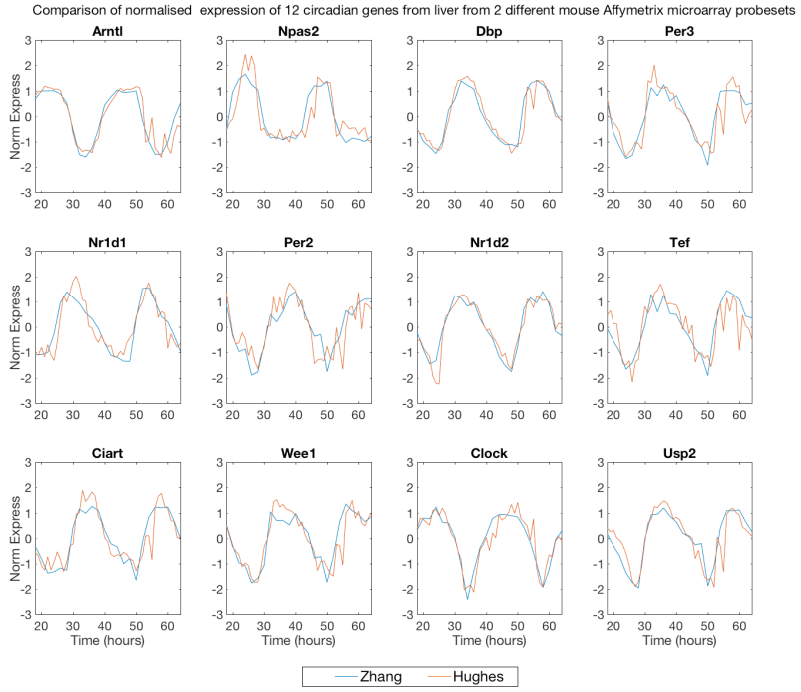


Figure 3.12: **Plot showing normalised circadian timecourse data from two independent datasets.** The overlays show an almost identical shape of the timecourses.

3.4 Human Timecourse Data

At the time of writing this thesis, very limited human tissue timecourse transcriptome data existed. Various studies have been published on the rhythms of human primary cells or stem cells [122, 6], and human blood timecourse was published in 2017 [14]. The *in vitro* data is not sufficient to compare with human *in vivo* circadian rhythms, as the cells are synthetically synchronised. The blood timecourse (which will be discussed further in chapter 4) uses custom microarrays, and cannot be compared to other datasets.

Suitable human timecourse was available, but is currently unpublished (as of April 2018). The human timecourse data used in this thesis was shared with us through a collaboration and MTA with Professor Georg Bjarnason of Sunnybrook Research Institute, Canada. This data set uses human tissue in the form of punch biopsies of ten individual's oral mucosa over 24 hours. There are 5 females and 5 males in the study, and the team in Sunnybrook Research Institute are investigating the differences in male and female circadian clocks[7].

The human Time-Teller model that will be presented in this thesis would not have been possible without this collaboration. The human timecourse oral mucosa data will be referred to in this thesis as the Bjarnason data.

3.4.1 Experimental design

Ten healthy human volunteers; five female subjects (# 05, 06, 13, 14, 18) and five male subjects (# 08, 09, 11, 12, 15), were recruited for the study. Mucosa tissue was collected at six timepoints 8 am, 12 noon, 4 pm, 8 pm, 12 midnight, and 4 am. Subjects were selected after screening by clinical history, physical examination, routine blood work (complete blood count, electrolytes, creatinine) and actigraphy to confirm regular sleep-wake patterns. Mucosa samples were collected by a dental surgeon, using a tissue punch biopsy. Subjects went to sleep in a dark room at their normal bedtime and were awoken for the midnight sample, and the 4 am sample[8].

After tissue mucosa samples were collected, they were immediately frozen in liquid nitrogen and stored at $-80°C$ until use. Total RNA was prepared by Trizol Reagent (Invitrogen) in accordance with manufacturer's specifications. RNA samples were quantified by optical density measurements at A260nm and A280nm. All samples were determined to be of high quality with A260:A280 ratios > 1.9. Total RNA (5μg) of each sample was used for microarray analysis on Affymetrix HG_U133_Plus2 chips. Cumulatively this chip represents 54,679 gene transcripts for analysis. Biotinylated cRNA were prepared according to the standard Affymetrix protocol (Expression Analysis Technical Manual,

[7]There is some evidence to suggest that there is a difference in male and female circadian clocks [123], but we will not discuss this in this thesis.

[8]The Research Ethics Board at Sunnybrook Health Science Centre approved the clinical protocol for this study.

2004, Affymetrix). Following fragmentation, 15 μg of cRNA were hybridized for 16 hrs at 45°C on GeneChip Human Genome U133 Plus 2.0 Array. GeneChips were washed and stained in the Affymetrix Fluidics Station 450. GeneChips were scanned using the Affymetrix GeneChip Scanner 3000.

Bjarnason *et al.* chose to use gcRMA normalisation of the microarray data, resulting in some low expression genes being excluded from the dataset. They also discarded any probes that were flagged as marginal at any time point for any individual. They used COSOPT analysis for rhythmicity detection. The result was 949 rhythmic transcripts for males, and 885 for females. All COSOPT results used in the following sections are a result of this analysis.

3.4.2 Analysis of raw human timecourse

The raw human data was shared in the form of 60 .CEL files, consisting of 10 individuals for 6 timepoints: 8 am, midday, 4 pm, 8 pm, midnight, and 4 am. The raw data was processed with the fRMA algorithm described at the beginning of this chapter. No filtering or quality control was performed. As we are only looking for highly significantly rhythmic and cross-observation-synchronised transcripts, the probability that any single bad reads will affect the results of these analyses, is extremely low. Also, we would not want to exclude a probe because its signal is low at the trough of expression, as the peak amplitude may be significant. Probes were annotated using probe information files from Affymetrix documentation. After fRMA processing, all expression values are in $\log_2$ format.

3.4.3 Synchronicity and rhythmicity detection of the Bjarnason data

A similar approach is taken in this section to the mouse analysis in the previous section. The differences here are that there are 10 individuals (not 8 organs), 6 timepoints over 1 day (not 24 over 2 days), and the Affymetrix HG_U133_Plus2 chips have 54,675 probes. JTK_CYCLE is not an appropriate algorithm to use for a single timeseries of 6 datapoints as any deviation from a monotonic behaviour between peak and trough in a 6 point timecourse results in a FDR value of 1. JTK_CYCLE can be used to look at the combined timecourse, but the results are not shown here. COSINOR is appropriate for rhythmicity analysis of sparse individual timecourses.

COSOPT analysis was carried out by Georg Bjarnason's group with specialist software, and so these values will also be used in this combined analysis. As gcRMA was used to normalise the raw data, QC cut-offs result in missing values for some probes. Also, only the top 1000 rhythmic genes were reported, so there are some missing values.

COSINOR regression

COSINOR was implemented in MATLAB for each individual and each probe over 6 time points. The algorithm was set to look for a period of 6, with a 95% confidence interval.

SVD analysis

SVD is performed after each timecourse was normalised to have 0 mean and standard deviation of 1, as described in detail in the previous section. The top 16 probes with the highest % variance explained by the first PC, are shown in timecourse expression in figure 3.13.

There are 10 genes represented by 16 probes in figure 3.13. All 10 of these genes are unquestionably core circadian clock genes, and the expression profiles are clearly synchronised in both shape and amplitude. Per3, Nr1d2, Arntl, Nr1d1, Per1 and Npas2 are represented by two different probes in this set. These duplicate probes have very similar timecourse profiles, but can have different expressions levels. This is especially noticeable for the Nr1d1 probes.

This finding, that human individuals have synchronised circadian clocks to such an extent, has always been expected. However, such evidence as this has never been published. This obvious and clean synchronicity in both phase and amplitude as seen in figure 3.13 is a significant and novel result. This was made possible due to the calibre of the Bjarnason data in conjunction with the SVD method of synchronised rhythmicity detection.

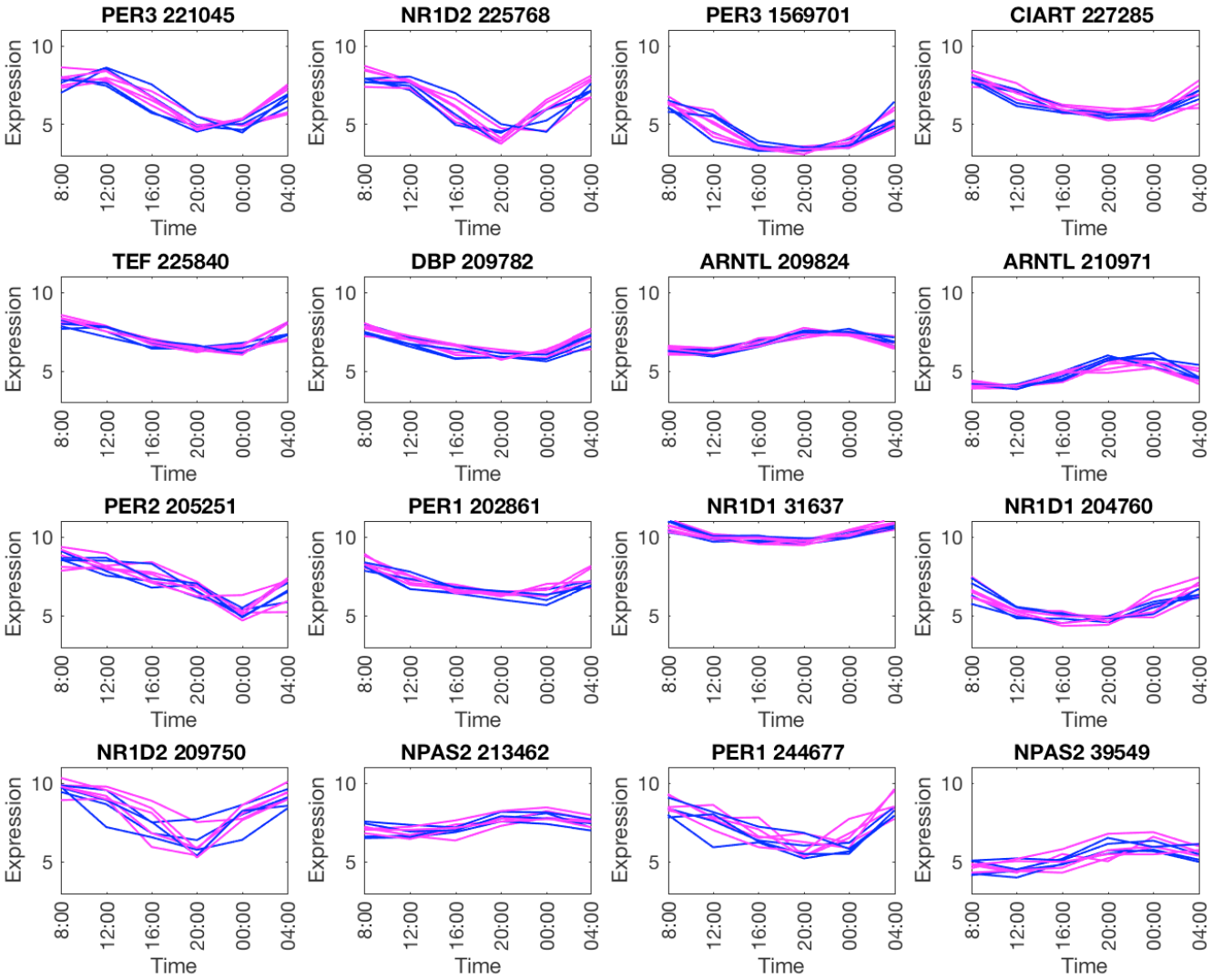


Figure 3.13: **Timecourse plot of the top 16 ranked "synchronised" genes for all 10 individuals.** All probes clearly have robust 24 hour, synchronised rhythms. Male (blue) and female (pink) lines do not show any differences.

3.5 Comparisons of rhythmicity detection methods

20 of the top 30 synchronised probes from SVD analysis are in the top 30 of the most rhythmic probes from the COSINOR analysis. These values summarised in table 3.6. Of the top 16 of these probes, 14 are present as rhythmic in Bjarnason *et al.*'s COSOPT analysis for both female and males, as shown in table 3.7. The exclusion of 202861_at (PER1) and 213462_at (NPAS2) in the COSOPT results is likely due to expression cut offs as a result of the gcRMA algorithm and filtering used prior to the COSOPT analysis,

Gene and ProbeID	% Explained	COSINOR geomean	Rank (SVD)	Rank (COSINOR)	Sum Rank
PER3_221045_s_at	0.87	8.00E-12	50	48	98
PER3_1569701_at	0.85	2.66E-13	46	50	96
TEF_225840_at	0.87	4.97E-11	49	47	96
NR1D2_225768_at	0.86	4.97E-11	48	46	94
PER2_205251_at	0.84	2.66E-13	44	49	93
CIART_227285_at	0.85	1.72E-10	45	45	90
PER1_202861_at	0.84	6.10E-10	43	43	86
DBP_209782_s_at	0.86	1.03E-08	47	39	86
NR1D2_209750_at	0.83	2.01E-10	41	44	85
ARNTL_210971_s_at	0.82	6.10E-10	39	42	81
ARNTL_209824_s_at	0.83	3.74E-09	40	40	80
NPAS2_213462_at	0.83	2.65E-07	42	36	78
NR1D1_204760_s_at	0.80	3.11E-09	36	41	77
NR1D1_31637_s_at	0.82	1.86E-08	38	38	76
PER1_244677_at	0.81	6.36E-08	37	37	74
NPAS2_39549_at	0.77	1.46E-06	35	33	68
NPAS2_1557689_at	0.73	1.10E-06	32	34	66
HLF_204753_s_at	0.70	6.49E-07	27	35	62
NPAS2_1557690_x_at	0.72	2.74E-06	30	32	62
GAREM1_219377_at	0.73	1.39E-04	33	24	57
LGALSL_226188_at	0.69	1.08E-05	24	31	55
HLF_204755_x_at	0.71	7.54E-05	28	26	54
SH3TC1_219256_s_at	0.74	1.83E-03	34	20	54
PER2_208518_s_at	0.70	2.37E-04	26	23	49
TMEM80_65630_at	0.69	1.31E-04	22	25	47
HLF_204754_at	0.67	1.75E-05	16	30	46
TSC22D3_208763_s_at	0.67	1.86E-05	15	29	44
MTERF2_225346_at	0.69	2.51E-03	23	18	41
BHLHE41_221530_s_at	0.66	3.66E-05	12	28	40
TPPP3_218876_at	0.70	7.08E-03	25	10	35

Table 3.6: **Summary of results of SVD and COSINOR analysis on the Bjarnason data.** Both the SVD and COSINOR are equally weighted in a ranking to find the most rhythmic and synchronised probes.

Probe	Genename	Male pMMC-β	Female pMMC-β
221045_*s_at*	PER3	0.00E+00	0.00E+00
1569701_*at*	PER3	0.00E+00	0.00E+00
225840_*at*	TEF	0.00E+00	3.58E-06
225768_*at*	NR1D2	0.00E+00	0.00E+00
205251_*at*	PER2	0.00E+00	0.00E+00
227285_*at*	CIART	0.00E+00	0.00E+00
202861_*at*	PER1	-	-
209782_*s_at*	DBP	1.68E-04	6.78E-03
209750_*at*	NR1D2	0.00E+00	0.00E+00
210971_*s_at*	ARNTL	3.58E-06	3.58E-06
209824_*s_at*	ARNTL	1.79E-05	3.58E-06
213462_*at*	NPAS2	-	-
204760_*s_at*	NR1D1	3.58E-06	0.00E+00
31637_*s_at*	NR1D1	0.00E+00	0.00E+00
244677_*at*	PER1	1.79E-05	1.43E-05
39549_*at*	NPAS2	5.01E-05	2.86E-05

Table 3.7: **Results of the COSOPT analysis performed by Georg Bjarnason** *et al.* pMMC-β values are measures of false discovery rates.

and not due to the rhythmicity detection.

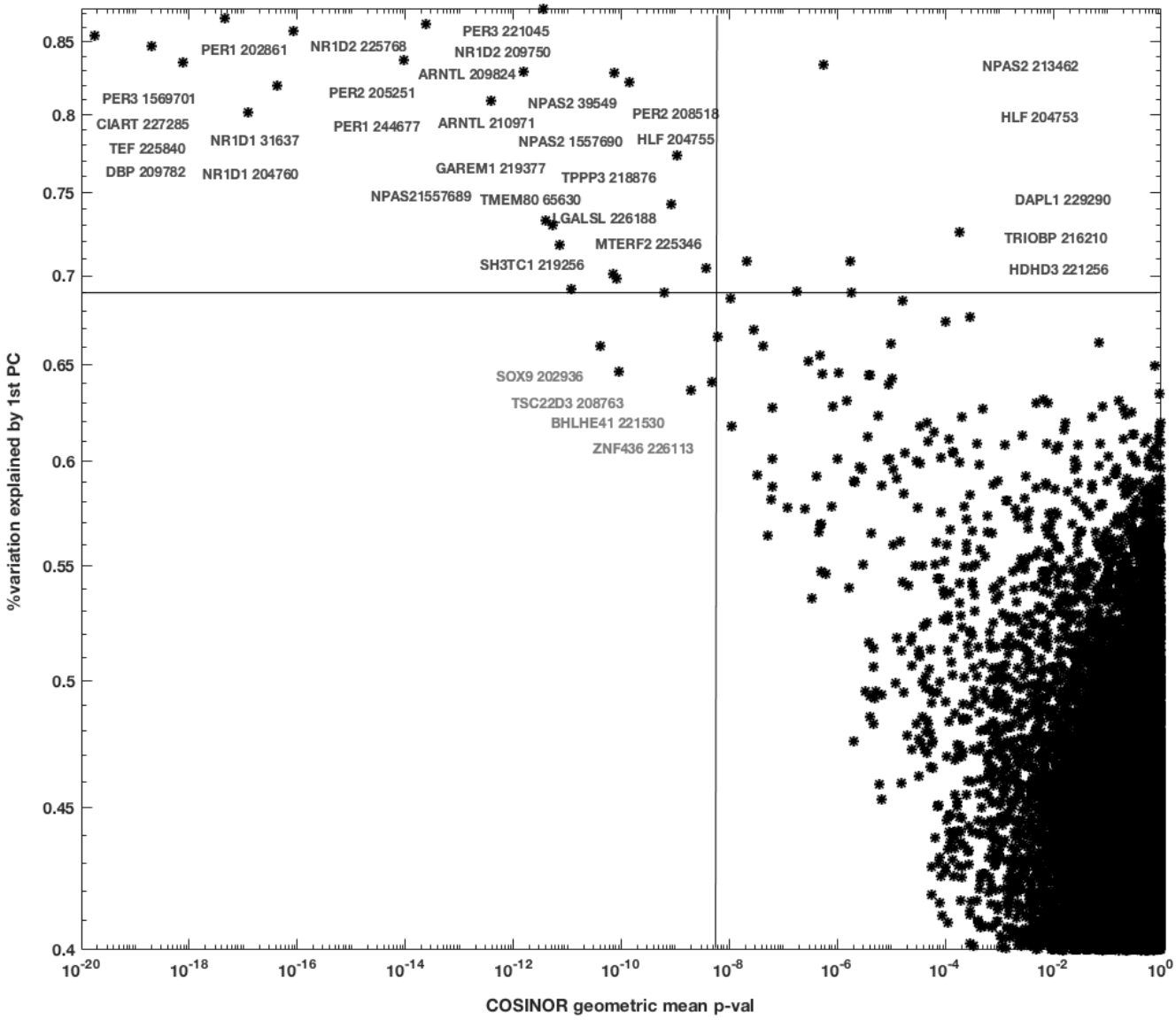


Figure 3.14: **Scatter plot showing % variation explained by 1st PC plotted against the geometric mean of COSINOR p-value for the Bjarnason data.** The top 30 genes for each metric are labelled, and 25 of these are shared, indicating in the human data that synchronised genes are rhythmic genes.

Figure 3.14 shows the results of the SVD analysis plotted against the geometric mean of COSINOR regression for each individual. The majority of the labelled top 30 genes for each metric are known core clock genes. 25 of the top 30 genes for each metric are shared, suggesting that for individual data, the majority of the synchronised genes are rhythmic genes. The mouse data showed many more genes in the top right quadrant of these plots, which suggested that there were genes whose expression was synchronised across organs,

but not 24 hour periodic. There are very few probes in this top right quadrant of figure 3.14, and those that are, are borderline.

No Clock, Wee1, or Ror transcripts were found to be rhythmic or synchronised in the human data, with a % variation explained metric <0.4 and with COSINOR p-values of 1. This suggests that these genes are mouse specific circadian clock genes. This could be valuable information to the circadian scientific community who often assume the circadian roles of Clock, Wee1 and Ror in human circadian studies.

Cry1 and Nfil3 are also significantly rhythmic and synchronised in this human data, but fall off the end of table 3.6.

The final 10 genes from 16 probes to be used in the human Time-Teller model are ARNTL (209824_s_at,210971_s_at), NPAS2 (213462_at,39549_at), PER1 (202861_at, 244677_at), PER2 (205251_at), PER3 (1569701_at,221045_s_at), NR1D1 (31637_s_at,204760_s_at), NR1D2 (209750_at, 225768_at), CIART (227285_at), TEF (225840_at), and DBP (209782_s_at).

3.6 Simulated Timecourse Data

The stochastic Relogio model shown in chapter 2 will be the source of the "dummy" *in silico* data that will be used in this thesis. Some hypotheses to test with the Time-Teller model require specific sets of test data that are extremely difficult to find or non-existent in the published online libraries. The stochastic Relogio model allows control over the noise in the generated timecourses, and as will be shown in chapter 5, it also allows "*in silico* knock-downs".

The noise from microarray data is a sum of biological stochasticity and error in experimental set up and measurement. It would be a huge study in itself to understand the noise of every contributing factor to this. We can however, crudely estimate it using the systemsize parameter, Ω, in the stochastic model, and the noise we see in the data we already have.

Some analysis of the stochastic model was performed in the previous chapter, but here we do more of a direct comparison of the stochastic model with the mouse timecourse data (the same could be done for the human Bjarnason data but is not shown here).

The Relogio ODE model was not designed with noise and stochastic variation in mind [7]. The building of a stochastic model to match the variance seen in real data would be an extremely interesting mathematical challenge, but is far out of scope for a side story to this thesis. Instead of changing what already exists, we compare the stochastic model with the real data in order to choose the most appropriate approximation for the system-size Ω.

3.6.1 Generation of *in silico* timecourse data

We used the stochastic Relogio model to simulate timecourses that could be compared with the Zhang mouse organ timecourses. In the previous chapter, it was shown that stochastic simulations lose synchronicity over time. It has been observed in multiple studies [125, 126] that this would also happen eventually to organisms living in an unforced environment. The mice in the Zhang study were synchronised with LD cycles for a week before being released into constant darkness, so there can be some form of alignment in the initial conditions of the data. The initial conditions for the stochastic Relogio model are chosen using a point on the limit cycle of the ODE Relogio model, γ_0 with added scaled Gaussian noise. This is in the form of:

$$Y_0 = \gamma_0\Omega(1 + N(0,1)/\sqrt{\Omega}) = \gamma_0\Omega + \gamma_0 N(0,1)\sqrt{\Omega} \tag{3.3}$$

so that Y_0 represents the gene expression at ZT0. To reflect the real data, samples are "taken" every 2 hours, by sampling 24 equally spaced values across 2 periods (approxi-

mately 52 hours).

The nature of the way time is sampled in the Gillespie algorithm means that there will not be 24 exactly equally spaced timepoints, but the closest rounded times. This does result in each output variable having a slightly different time to its label, but this noise is very small and just adds more realistic noise to the model, so is of little consequence.

The RNA terms for Arntl, Per, Cry, Rev-Erb, and Ror in the 19 variable Relogio model were saved as 8 (organ) by 24 (timepoint) matrices. The same matrices for Per3, Cry1 and Rev-Erbα, and Rorc in the Zhang data were used in parallel. Real and *in silico* timings were aligned visually in order to produce the normalised figure 3.15, where the coloured timecourse is real data (coloured by organ type), plotted with 20 trajectories of scaled simulated data (black), with initial estimate for Ω of 100.

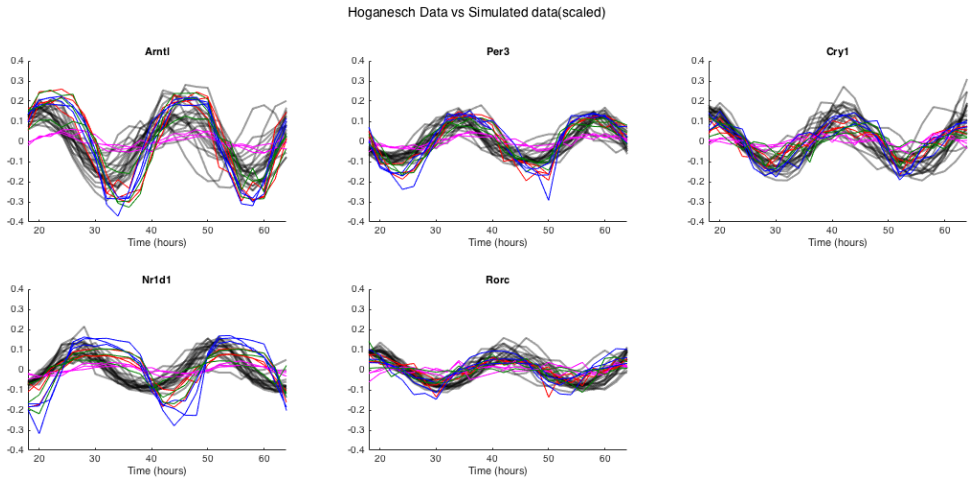


Figure 3.15: **Plot showing normalised Zhang and simulated timecourse data.** Zhang data is coloured, and plotted with 20 normalised simulations in black (with $\Omega = 100$. The pink timecourse represents the low amplitude brain data.

There is generally good agreement amongst phasing, with perhaps a slightly too early phase for Nr1d1. The Relogio model was designed before the Zhang data was available. This correlation is a very positive result for both the viability of the Relogio model and the hypothesis of this thesis that the circadian clock is well-behaved and robust across independent datasets.

Variation and stochasticity exist in the Zhang data, and varies by gene (as was measured in section 3.3.2), by organ, and probably also by time. SVD was used to measure the variance of each gene of the real data, and the same thing can be done here for the *in silico* data.

Estimating Ω: method 1

Figure 3.16 shows the results of the same SVD analysis as was done in section 3.3.2, but the %variance explained is now plotted (in black) for all 9 principal components[9]. The same analysis was performed for 9 sets of dummy in silco data with $\Omega = 100$ in blue, and $\Omega = 1000$ plotted in red.

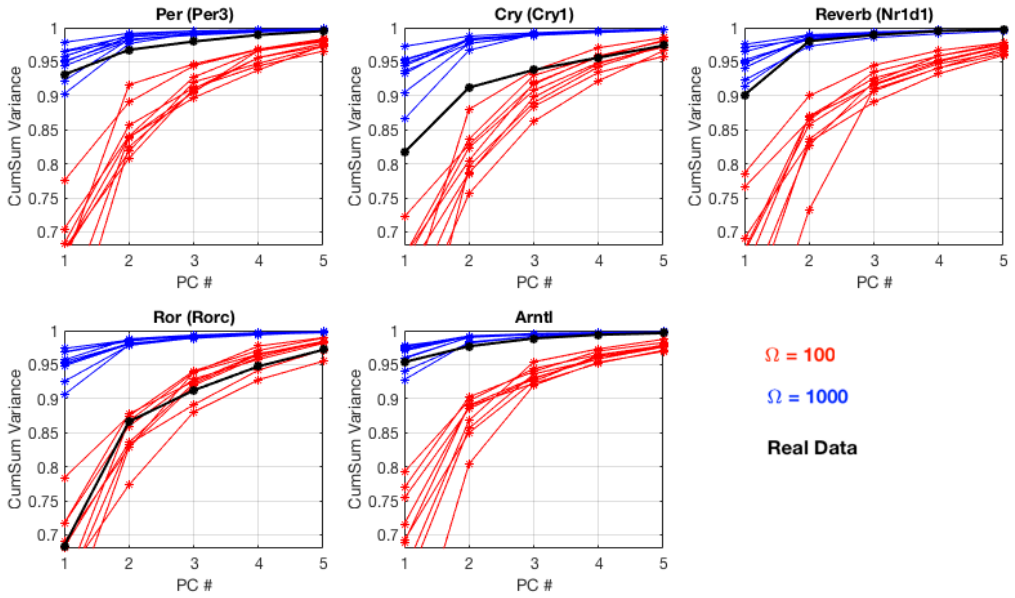


Figure 3.16: **Plot showing the cumulative % variance explained by singular values, for five genes, for real and simulated data.** Testing 10 sets of 9 simulations of the Relogio model simulation with $\Omega = 100$ (blue) and $\Omega = 1000$ (red) compared to the same analysis on real mouse data (black).

Cry and Ror have similar variance to a model with $\Omega = 100$. However, neither Cry1 nor Rorc are chosen as training genes for the mouse model. The real Rev-Erb data appears to have similar variance to a model with $\Omega = 1000$. Both Arntl (Bmal1) and Per are somewhere in between.

It is interesting to notice that the variance for the 5 simulated genes is extremely similar, but the real data shows significant differences. This shows that the stochasticity for each gene is different in reality, and this is not something that the Relogio model has accounted for. It would be interesting if a stochastic model were designed to be able to replicate this type of differing variance behaviours.

Estimating Ω: method 2

Another way to estimate Ω is to again perform SVD, but on matrices for each organ, where each matrix is 5 genes by 24 time points. The SVD is not measuring synchronicity

[9]In this analysis White Fat was also included.

here, but how many dimensions the elliptic orbit of the 5 clock genes are occupying. Figure 3.17 shows a plot of each set of decreasing singular values. The plots are coloured according to organ type, and simulations of two different Ω system sizes have been plotted in black (1000) and yellow (100). The reproducibility of the simulations for $\Omega = 1000$ is apparent, as is the variation of the $\Omega = 100$ simulations.

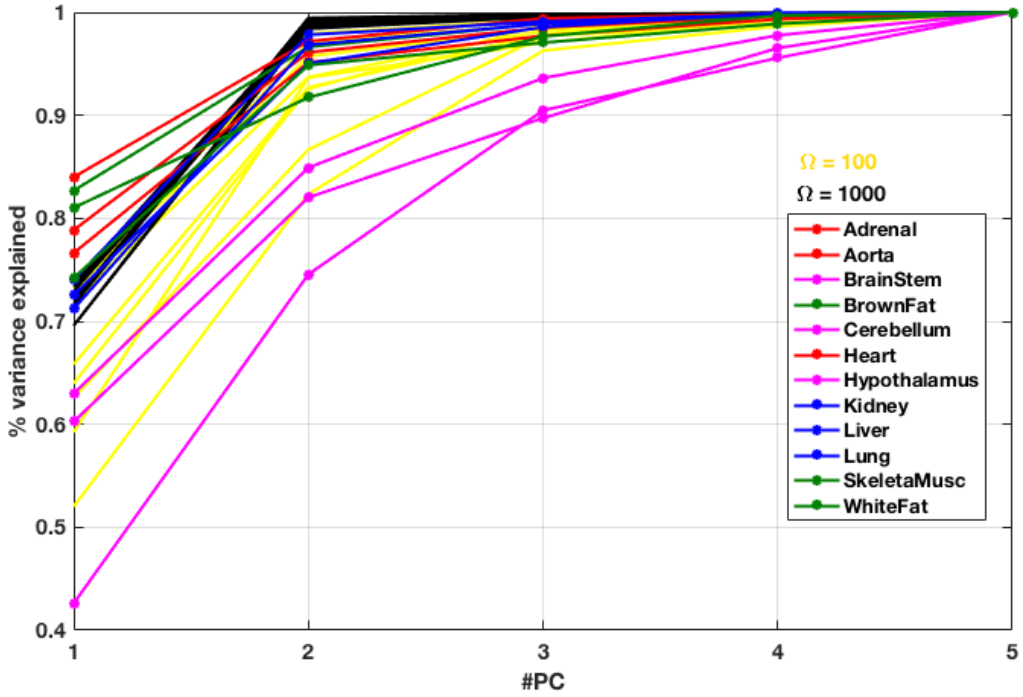


Figure 3.17: **Plot showing the cumulative % variance explained by singular values, for each organ, for real and simulated data.** The coloured plots are for the real data and coloured by organ type. The 10 black lines represent the singular values of simulated data for $\Omega = 1000$, and 10 yellow lines represent the same for $\Omega = 100$.

The three brain regions (in pink) show much higher noise than the other organs, as expected. Figure 3.17 suggest that the noise in the brain data would be represented by a stochastic system with a system size less than 100. The 5 gene dimensions are represented by the first two PCs by more than 90% for all other organs.

For $\Omega = 1000$ simulations, the first PC under represents the data in comparison to the real data but the second PC over represents it. Additionally, half of the $\Omega = 100$ simulations have more noise than the real data.

Although this is a rough analysis, we can conclude that an appropriate range for Ω is approximately between 100-1000.

Using more variables

The analysis above shows that the 5 mRNA terms in the Relogio simulate the real data well, in terms of phase and noise (when correct Ω is used). Five terms however, are not

adequate for a training set for a timetelling model, so it would be useful to be able to use more than 5 of the variables. As each variable trajectory is a periodic timecourse, it is reasonable to include a subset of the 19 variables in this type of dummy data analysis.

When all 19 variables are included (with $\Omega = 100$) the first 5 PCs explain about 99% of the variance in the data. The first and second principal components of the 5 and 19 variable systems are almost identical, as shown in figure 3.18. This is a consequence of the robustness of the model dynamics, and indicates that the non-mRNA terms can be used to generate data for a dummy data set.

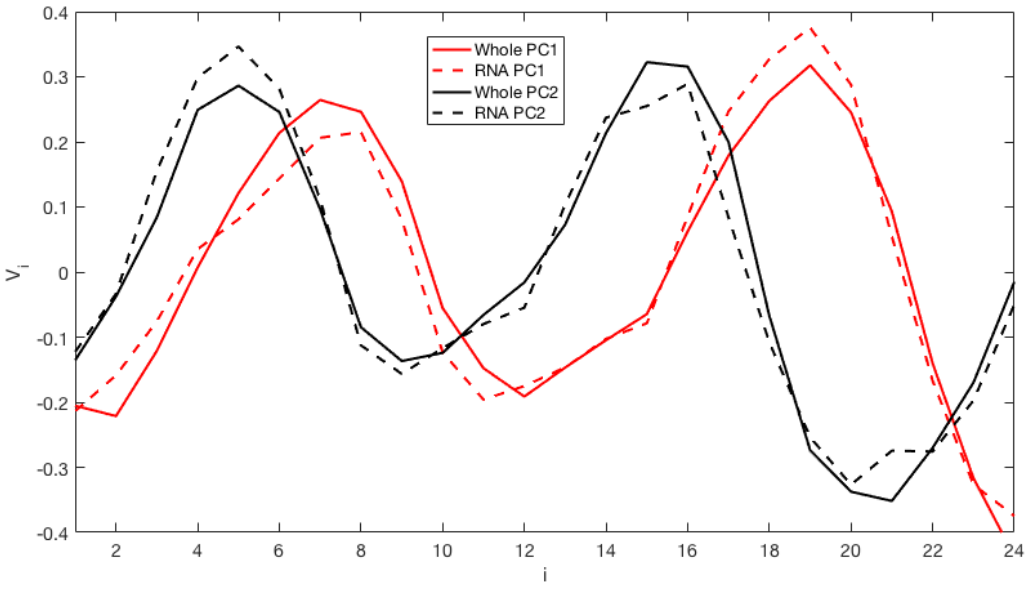


Figure 3.18: **Plot showing the first and second PCs of the dummy data for 5 mRNA terms only, and for the full 19 terms.** The PCs are very similar in profile as the 5 mRNA terms are enough to derive all the dynamics in the data.

Although this shows a high level of redundancy in the extra terms of the model, it also shows that the extra terms can be used without the risk that the "non gene" terms of the model show different behaviour to the "gene" terms. Non gene variables will be used in the dummy data Time-Teller.

This analysis can conclude that a system size in the range $\Omega \in [100, 1000]$ can be reasonably used to simulate data comparable to experimental data. Although more thorough and accurate methods could be used to find an optimum Ω, nothing would be gained from this for this thesis, and the validated range will be adequate.

3.7 Summary of Chapter

This chapter has presented three sets of timecourse data: mouse, human and *in silico*. The identified rhythmic and synchronised genes of the mouse and human data will be used as training genes for the mouse and human Time-Teller algorithms. The stochastic Relogio model will be used to generate dummy data of appropriate Ω, as both the training and test data for the *in silico* Time-Teller.

The use of the SVD was presented as a simple and computationally efficient way of finding rhythms by exploring synchronicity[10]. The use of multiple rhythmicity algorithms in parallel with the SVD allows interesting outliers to be identified, and allows higher confidence in the results.

Using SVD, COSINOR, and JTK_CYCLE, 11 mouse probes were chosen to use in the mouse Time-Teller. Using SVD and COSINOR, 16 probes (representing 10 genes) were chosen to use in the human Time-Teller.

In silico time courses were generated and the Ω range of 100-1000 was validated for use for dummy data in which to train and validate the *in silico* Time-Teller.

The correlation in the rhythmic and synchronised measures in the Bjarnason data suggests that no genes but the circadian clock genes are synchronised in expression pattern for human individuals. It also suggest that the more 24 hour rhythmic a gene is, the more synchronised that gene is to the general population. Tightly coupled autonomous oscillators were shown to be robust in chapter 2, and this result complements this. A tightly coupled oscillator system results in robust and rhythmic gene expression which is relatively unchanged in response to external noise, and hence results in synchronised gene expression profiles across the population.

This means that these sets of tightly regulated, robust oscillators, are genes that can be used to define the expected behaviour of a healthy circadian clock.

[10] As far as we know, no such analysis has been published like this before, and the basic approach may be a useful functional genomics tool to assess synchronised rhythmic gene expression for different tissues and observations.

Chapter 4

Time-Teller

The previous chapters have provided evidence for the robustness of the circadian clock. The main results have been that:

- both a deterministic and stochastic autonomous circadian oscillator show very robust dynamics.
- a set of mouse circadian clock genes have synchronised rhythmic expression across tissues and organs.
- a set of human circadian clock genes have synchronised rhythmic expression for 10 individuals
- the GeneChip or technology used shows the same behaviours for individual genes in timecourse, but actual expression levels differ slightly by GeneChip or technology.

Evidence so far now leads to the idea of an expected behaviour of the clock: that the transcriptome has an embedded time signature. In this chapter, by finding a transcriptomic signature related to time of day we develop a model for telling time from just one sample.

The timecourse data (Zhang mouse data and Bjarnason human data) will be used as the healthy standard to make predictions for other samples; i.e. these data sets are the model training sets. Making predictions of the unknown based on a labelled training set is a classic problem in machine learning.

4.0.1 Introduction to machine learning

Machine Learning is a broad term, referring to algorithms that are able to learn from data, without being explicitly programmed [127]. We give a short introduction to machine learning here, but a more in detailed introduction can be found in [128].

The two main types of machine learning are

- Supervised - A set of known inputs and matched outputs are presented so that a rule for a mapping can be learned.
- Unsupervised - Only inputs are provided and the machine tries to sort them so that some structure can be elucidated.

Supervised machine learning methods need labelled training data, and usually involve some sort of regression and/or classification. Support vector machines (SVM), k-nearest neighbours (k-NN), and linear discriminant analysis (LDA), all try to discretely classify data to known discrete labels. Regression algorithms, such as least squares, smoothing splines, partial least squares regression (PLSR) or gaussian process regression are supervised machine learning algorithms that try to find continuous relationships between inputs and outputs. Some deep learning methods such as convolutional neural networks are supervised methods, as the machine learns underlying patterns to the data starting with inputs and ending up at the known output.

Unsupervised machine learning methods are used to draw inferences from unlabelled data. It also includes many clustering methods such as k-means and Gaussian mixture models. Dimensionality reduction algorithms are technically unsupervised algorithms, but can also be used when the data labels are known (semi-supervised). Algorithms such as PCA or tSNE (t-Distributed Stochastic Neighbour Embedding) [129] are dimensionality reduction techniques where the data is (usually) reduced to 2 or 3 dimensions, and then the user can assess whether the known "classes" are distinct within the unsupervised decomposition. This can be done simply by looking at the data in different colours, or by using further clustering analyses for classification (in PCA only). Deep learning methods such as those that use autoencoders in neural networks can infer patterns from unlabelled data.

Machine learning and circadian rhythms

In this thesis, circadian transcriptome data is the input to all models, where each gene expression timecourse can be represented by a continuous, 24-hour periodic curve. Time is used as a continuous one dimensional output. Although we could treat time as discrete (all of the training data is in 2 or 4 hour intervals) this would hugely limit the model's accuracy and scope. For this reason, discrete classifier methods are not appropriate for time-telling models. This discounts use of k-nearest neighbours methods and stochastic methods such as tSNE in time telling models.

PCA, PLSR, Gaussian process regression, and neural networks are all potentially useful in the "one time-point timetelling" question. However, standard approaches need

to be tailored to be used with a periodic input. The next section explores these methods, and those that have been used in existing literature to answer the timetelling question.

The standard setup for all algorithms is to start with a matrix of training data $X \in \mathbb{R}^{n \times p}$ where n is the number of observations (e.g. organ, individual), and p is the number of features (i.e. genes). Each row of X is associated with entry n of output vector $T \in \mathbb{R}^n$ (e.g. circadian time or body time) [1].

An important step in many machine learning algorithms is the normalisation of the data. If any batch methods are used to normalise the training data, then the model will not work for single test samples - unless the same normalisation was used for every test sample.

4.1 Literature Review of Time Telling Models

At the start of this PhD project there was only one published study with the explicit aim of estimating the time from the transcriptome. However, since then, at least five studies have been published that present methods that are directly relevant to this aim of this PhD project. Some methods have similarities to the novel model presented in this chapter, and some take different approaches, but they all involve some form of machine learning.

A comprehensive literature review of the main"single time point time-telling" studies is given now at the start of this chapter, and a summary is given in table 4.1. Associated papers with the main publications, either pre- or post- publication, are shown in non-bold font.

4.1.1 Molecular timetabling method (MTTM)

The aim of the study in Ueda *et al.* [11] was to detect individual body time (BT) via a single-time-point assay so that BT information could be exploited to optimise medication strategies. This is still very much the aim of every study presented in this chapter. Although this study uses BT and not clock time in their estimation, the results in the study show that in controlled mouse experiments, they are equivalent.

Experimental methods

Ueda *et al.* created a set of microarray expression profiles from pooled livers of four male mice every 4 hours over 2 days under 12 hr light/12 hr dark (LD, ZT) or 12 hr dark/12 hr dark (DD, CT) conditions. Profiles were also made from independently sampled livers from 8 individual male mice at ZT12 ($n = 4$), ZT6 ($n = 1$), ZT18 ($n = 1$), CT6 ($n = 1$), and CT18 ($n = 1$), where n represents the number of mice sampled. Additionally, they

[1] except for the PLSR model estimating melatonin, not time, where the output is also periodic

Method Name	First Author	Citation	Journal & Year	Description
Molecular Timetabling	**Ueda**	[11]	**PNAS 2004**	Best cosine fits
Zeitzeiger	**Hughey**	[12]	**Nucl. Acids Res. 2016**	**Trained SPC model with MLEs**
Supervised PCs	Bair	[130]	2006	Prediction by supervised PCs
	Hughey	[131]	*Gen. Med.2017*	*Application to human blood dataset*
BIO_CLOCK	**Agostinelli**	[132]	**Bioinformatics 2016**	**Deep Neural Networks**
CYCLOPS	**Anafi**	[13]	**2017 PeerJ**	**Deep Neural Network model**
SVD for genome data	Alter	[133]	2000	Defines an Eigengene
PLSR	**Laing**	[14]	**ELife 2017**	**PLSR model with Melatonin output**
Partial least squares	Boulesteix	[134]	2006	PLS algorithm
ΔCDD	**Shilts**	[15]	**2018 PeerJ**	**Metric for clock dysfunction using correlations**

Table 4.1: **Table listing published Literature on Time-telling models.** Bold entries show the publication of the main algorithm, italised entries show follow up publications using the algorithms, and other entries show publications of methods underlying the algorithms.

sampled livers from 7 individual *Clock/Clock* homozygous mutant mice at ZT12 ($n = 4$) and ZT8 ($n = 3$) and 3 individual male mice at ZT8 ($n = 3$) to verify the feasibilty of expression-based diagnosis of circadian rhythm disorders. All mice were synchronised under LD conditions for 2 weeks from 5 weeks postpartum and then sampled.

Molecular time table method

To detect *time-indicating genes*, i.e. genes whose expression exhibits circadian rhythmicity with high amplitude, the expression profile of each gene was analysed for rhythmicity and amplitude. Methods similar to COSOPT analysis were used to determine rhythmic genes. The top N rhythmic genes were then chosen for use in the timetabling method.

They normalised the expression profile X_i of each gene $i \in [1, ..., N]$ using its mean μ_i and standard deviation σ_i in the molecular timetable. The normalised profile is described by Y_i.

To estimate the BT of each expression profile, they fit a cosine curve to each normalised profile:

$$Y_i = \sqrt{2}cos(2\pi(t - b_i)/24) \tag{4.1}$$

where each b_i denotes the phase of each gene. Cosine fits are restricted to have a resolution of 10 min intervals, hence there were 144 possible cosines each timecourse could be fit to.

The b_i phases represent the fixed timetable of phases for each gene.

To make the estimation for the time of a single sample, a single set of N gene expressions, of unknown time, is normalised with the same μ_i and σ_i as the training set (i represents each gene). The "best" BT estimate is defined by the b_i that gives best correlation value of these normalised values to each set of 144 cosines.

To evaluate the statistical significance of BT estimation, the authors generated a random expression profile Y_r following the same distribution of the normalised data, Y_i. They calculated the best correlation value c^r and the phase b_c^r. This was done 10,000 times to create a distribution of c^r values and phases b_c^r. The probability (P_r) that a random expression profile has a best fitted cosine curve giving correlations equal or greater than those of the real expression profile was determined by the correlation value c_r from 10,000 expression profiles.

Results of MTTM

Using 168 time-indicating genes, the paper reports errors (difference of estimation and real time) of under 2 hours when testing 4 samples from an independent data set of mouse liver transcriptome. The method also attempts to diagnose circadian rhythm disorders. Using mutant Clock KO mice, and WT mice, they were able to detect significant rhythms in WT mice, but not in KO mice. These results are shown in figure 4.1, where the top row with fitted cosines represents WT data, and the bottom row with no significant fit found is showing the KO data.

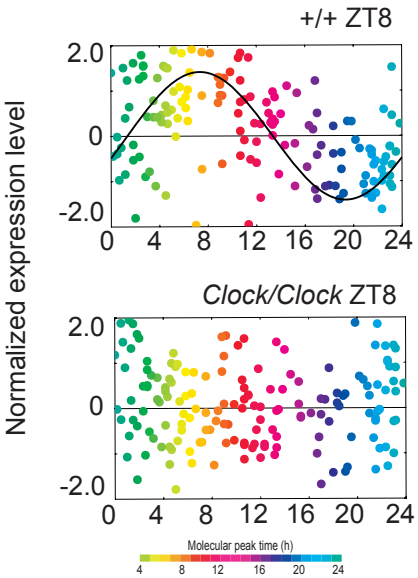


Figure 4.1: **Example plot for results of MTTM.** From [11]. At ZT8, figure shows significant rhythms in WT mice (top plots), and insignificant rhythms in KO mice (bottom plots). The x-axis represents each gene's calculated phase, where each dot is a gene.

Although this method has generally reported good results, its lack of power is clear.

The method needs fixed normalisation, expression shapes must be cosine, and every gene is treated as independent from others. Additionally, 168 genes is a much larger number of genes than any of the other models in this section will use, and it appears that computing time would be huge for this method. It was a very good initial algorithm for time-telling, but the method had much room for refining.

4.1.2 Zeitzeiger

Hughey *et al.* [12] published Zeitzeiger: a supervised learning method for high dimensional data from an oscillatory system. Zeitzeiger, meaning "Time-revealer" is presented as a method to predict a periodic variable (e.g. time of day) from a high dimensional observation. It is claimed that Zeitzeiger is faster and more accurate than the molecular timetable method.

Method

The Zeitzeiger method "learns" to use a particular set of genes to predict time (these may or may not be core clock genes). It uses training observations to learn a sparse representation of the variation associated with a periodic variable, then makes a prediction based on maximum likelihoods. Zeitzeiger is conceptually similar to supervised principal components (SPC) [130], as the genes used in the principal component analysis are chosen so that they show maximum variation over time. The full method is summarised into steps for clarity.

Step 1: Estimate the time dependent density of each feature j.

- Estimate the time-dependent mean, $f_j(t)$, by fitting a periodic smoothing spline to the behaviour of each feature j (gene) as a function of time.
- Estimate the variance of each feature as the sum of squared residuals from the spline fit s_j^2.

There is a homeoscedasticity assumption used here: that the variance of each feature (gene) about the mean is **constant** for all time. They claim that this is simpler and more robust than trying to estimate a time-dependent variance [2].

Step 2: Identify the major patterns that describe how features change over time. Construct a matrix $Z \in \mathbb{R}^{m \times p}$ in which the spline discretised into a number of time-points and scaled by that feature's standard deviation about the mean curve.

[2]The novel model in this thesis does not have to make this assumption.

If τ_i is the corresponding time-point for the ith row in Z, then

$$z_{ij} = \frac{f_j(\tau_i) - \bar{f}_j}{s_j} \tag{4.2}$$

$$\text{where } \bar{f}_j = \frac{1}{m}\sum_{i=1}^{m} f_i(\tau_i) \tag{4.3}$$

where
$\bar{f}_j$ is the mean of feature j over all timepoints.
i indexes time, where j indexes gene.
τ represents time scaled between 0-1
m is the number of timepoints (where m is also the maximum number of SPCs that can be used for prediction).

Step 3: Subject Z to a penalized matrix decomposition. PMD is performed on Z. PMD is similar to SVD except that it is adapted to be more suited to sparse datasets. Sparse data sets are large, noisy data sets with low rank [135]. This is suitable when all genes are being used, and not just a cherry picked subset. As a result, matrix $V \in \mathbb{R}^{p\times m}$ is generated, which is a matrix of m PCs, each of length p.

Step 4: Project the training data from high-dimensional feature-space to low-dimensional PC-space. Produce a new matrix $\tilde{X} \in \mathbb{R}^{n\times m}$ calculated as $\tilde{X} = XV$.

Step 5: Estimate the time-dependent density of each SPC. Denote the time-dependent mean of the kth SPC as $f_k(t)$ and the variance as $\tilde{s}_k^2$.

Step 6: Project the test observation from feature-space to SPC-space. A test observation $w \in \mathbb{R}^p$ is projected from feature space to SPC-space $\tilde{w} = wV$.

Step 7: Use maximum likelihood estimate to predict the time of the test observation. Assuming that each PC is Gaussian at any given time, the likelihood of time t given $\tilde{w}_k$ is,

$$\mathcal{L}_k(t|\tilde{w}_k) = \frac{1}{\tilde{s}_k\sqrt{2\pi}} e^{\frac{-(\tilde{w}_k - \tilde{f}_k(t))^2}{2\tilde{s}_k^2}} \tag{4.4}$$

Where $nSPC \leq m$ is the number of SPCs, the universal likelihood is calculated by

the sum of each PC's log-likelihood.

$$\mathbb{L}(t|\tilde{w}) = \sum_{k=1}^{nSPC} \log\mathcal{L}_k(t|\tilde{w}_k) \tag{4.5}$$

They make the comment in the study that this is not a mathematically valid thing to do, but that it works well empirically. It could be argued that each of these combined single Gaussian likelihoods should at least be weighted by the singular value associated with each SPC.

The predicted time $\hat{t}$ for test observation w is $\hat{t} = \arg\max_{t\in[0,1)}\mathbb{L}(t|\tilde{w})$, i.e the value of t for which the likelihood function attains its maximum value.

Zeitzeiger applied to Zhang data

Using the Zhang mouse organ data [9], Zeitzeiger was applied in a leave-one-organ-out methodical approach to estimate the time of single samples. Figure 4.2 shows the shapes of the first 2 PCs, and the 14 genes that contribute to them. These training genes include 10 of the 12 genes found in the previous chapter to be rhythmic and synchronised (excluding the genes Ciart or Clock). The data was first adjusted for organ specific differences, by the batch processing program ComBat [136]. This appears to be a redundant step, as all the data is just normalised in timecourse anyway.

Zeitzeiger versus MTTM

The authors state that Zeitzeiger is on average 1.2 hrs more accurate than the molecular timetable method, even when using 13 genes where the MTTM uses 110. Decreasing the number of genes used in the MTTM decreases accuracy. The authors state that Zeitzeiger is more than twice as fast at the MTTM. To run leave-one-out cross validation Zeitzeiger took around half of the CPU time than the MTTM did.

Zeitzeiger using independent samples

Multiple mouse datasets from WT mice were tested using the Zhang data as the models training data (details can be found in the paper). All datasets are normalised as time-courses so comparing samples from different technologies is not an issue with this method. It does mean, however, that Zeitzeiger can only be used to predict the time of samples that exit in timecourse.

The prediction results of Zeitzeiger are very accurate (< 1.5 hrs absolute mean error for most datasets). The Hughes liver timecourse data [10] is used as a validation dataset

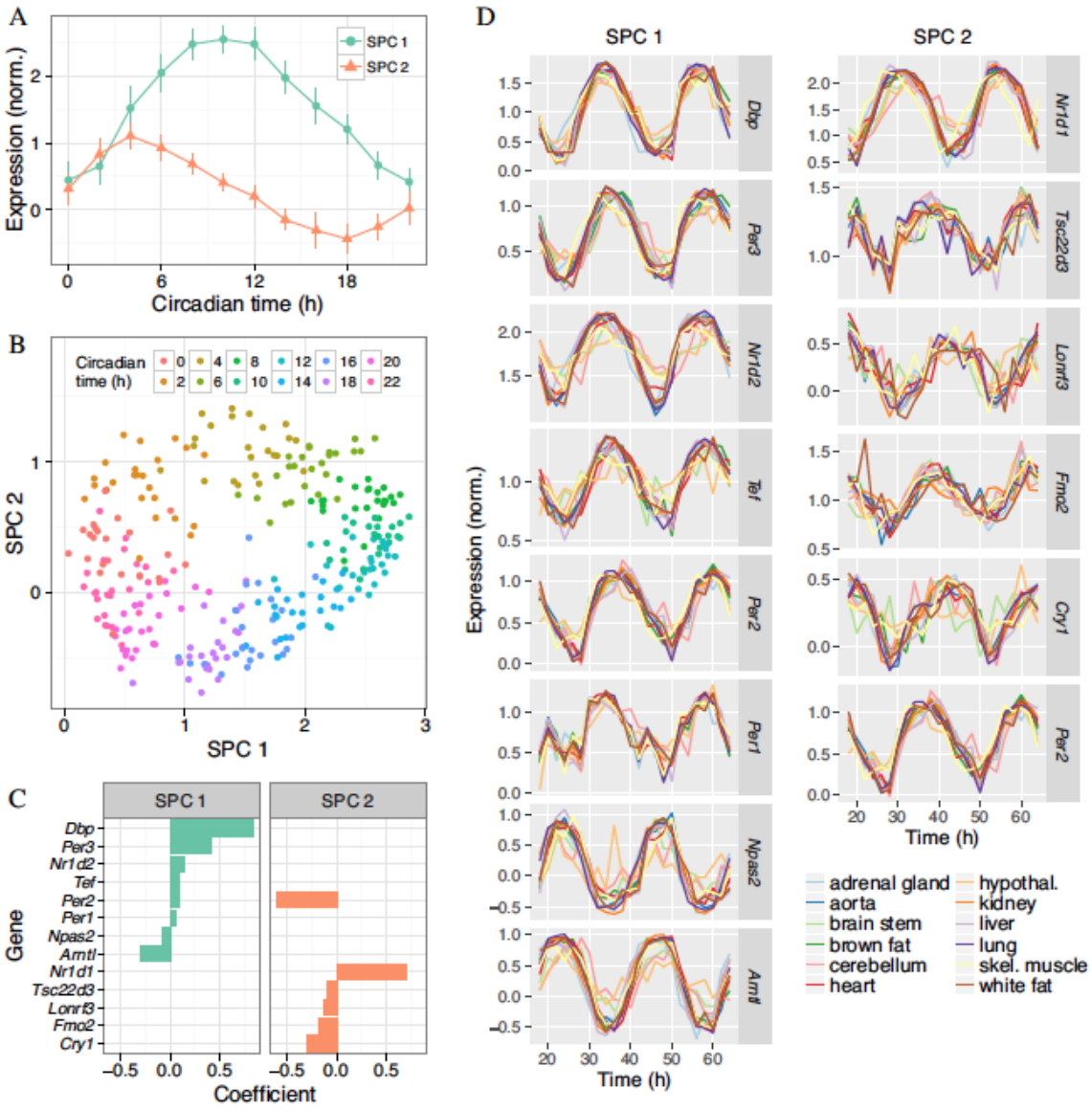


Figure 4.2: **Figure showing Zeitzeiger SPCs for Zhang data.**From [12]. (A) Trajectory of the two SPCs as a function of circadian time. (B) Gene expression of the samples in SPC-space. Each point is a sample, with colour indicating the (true) circadian time. (C) The loadings for each SPC. (D) Normalised expression versus time for the selected genes. Time is shown as the full 48 h of the experiment.

for this, and absolute prediction error is reported to be very low at < 0.5 hours. As was shown in the previous section in figure 3.11 the probe values do not align for these training and test datasets. Zeitzeiger's need for batch normalisation of timecourses is its biggest weakness, as it is unlikely that it will be useful to be able to predict the time of samples that already exist as a timecourse.

The study does mention a potential use of Zeitzeiger is to investigate if samples that exist as a timecourse have a dysfunctional circadian clock, stating "*We hypothesised that a dysfunctional clock might cause not only aberrant timing, but also poorer fit of the observed*

gene expression to what would be expected at a particular time." They observed that the log-likelihoods of their MLE predictions were significantly lower in mutant samples than in WT. They also noticed that the location of wild-type samples projections were in a similar space to that of the training samples, but the mutant samples were deviated away from this WT periodic trajectory. Zeitzeiger simply made these observations, but Time-Teller, the novel method presented in this thesis will use these observations to create a metric of clock dysfunction.

Zeitzeiger could be extended to use the methods developed for Time-Teller to measure the dysfunction of the circadian clock, but Zeitzeiger cannot help to achieve the aims of this thesis. This is because Zeitzeiger requires that samples be acquired throughout the 24 hour cycle, in order to make predictions. Hence Zeitzeiger is not applicable to testing single samples of unknown time.

Applications to human blood dataset

In a following study, Hughey applied Zeitzeiger to three publicly available datasets of human blood timecourse transcriptome [131]. Each dataset had a different experimental design; each consisting of data from individuals from control conditions, and conditions in which sleep and LD cycles were perturbed. Only 2 of the 15 genes used for training are known core circadian clock genes (Per1 and Nr1d2), the others having no previous connection to the circadian clock. This appears to be a feature of the data, and not the algorithm. Good results were reported despite this with median absolute error being $< 2hrs$. Errors for estimating time were generally higher for individuals with perturbed clocks.

In the section "processing gene expression data", Hughey states that the microarray datasets were processed using MetaPredict, which performs both intra-study and cross-study normalisation, so is not relevant for single time point estimation.

4.1.3 BIO_CLOCK

Agnostinelli *et al.* published a study called "*What time is it? Deep learning approaches for Circadian Rhythms*" [132]. The paper has two themes; one of period detection and the other of estimating time from single samples. Despite the name of this paper, there is little focus on the methods for time-telling. The methods for BIO_CLOCK are simply described as "a supervised deep learning algorithm using neural networks", with few method specifics discussed. In one example, they state that they train BIO_CLOCK using mouse data from Zhang *et al.* [73], and the genes Arntl, Per1, Per2, Per3, Cry1, Cry2, Nr1d1, Nr1d2, Bhlhe40, Bhlhe41, Dbp, Npas2, Tef, Fmo2, Lonrf3 and Tsc22d3. The reasoning behind this choice is not given, but it appears to be a mixture of a literature search and a rhythmicity analysis. They state that "training and test samples are normalized to have

mean of 0 and standard deviation of 1. They do not clarify if this is for each single sample or for each timecourse.

Different microarray datasets and RNA-seq data are used in this study, with good results reported. There is no mention of how datasets have been normalised in order to be able to compare microarrays and RNA-seq, or comparisons between different probes and annotations.

Due to the lack of information given in this paper towards the methods, it is difficult to review.

4.1.4 CYCLOPS

Anafi *et al.* published CYCLOPS [13]: a neural network that finds circadian patterns in datasets. This study does not present an algorithm that can (yet) tell the time of single samples (and does not claim to be able to), but it has clear value in this field.

The study uses a (quasi) unsupervised machine learning algorithm (a neural network) to construct a cyclic periodic timecourse, using unordered large datasets of expression measurements according to the clock genes, by mapping them to ellipses. These known rhythmic genes are given as prior information, hence this review's "quasi"-unsupervised renaming of the approach.

Methods

CYCLOPS builds on the ideas from a previous algorithm pulished by Leng *et al.*, called Oscope, which was designed to investigate the cell-cycle in single cell RNA-seq data [137]. Leng *et al.* saw that when they plotted pairs of genes against each other some gene expression pairs formed ellipses. The resulting algorithm to explore this is extremely computationally expensive and is not an ideal way to identify rhythmicity. Anafi *et al.* build on this idea of fitting data to ellipses, but by fitting "eigengenes" to ellipses, not pairs of gene expressions. **Eigengenes** is a term coined by Alter, Brown and Botstein in 2000 [133], when they looked at the cell cycle in yeast. An *eigengene* is defined as a characteristic expression pattern after SVD on a matrix of N genes and M (disordered time) samples. When plotted in expression space, the ordered eigengenes v_1 and v_2 were observed to be almost $\pi/2$ shifted, resulting in an elliptical shape when plotted together. The M time samples can be in any order in eigengene analysis, and would achieve identical results as all that matters is the comparative relationship between features (genes).

The number of eigengenes retained was set so that $> 85\%$ of variance was captured. CYCLOPS optimally weights and combines the eigengenes, in order to create the closest thing to an ellipse. These points are subsequently ordered along that ellipse using a neural network. We do not discuss neural networks here, but full details can be found

in the supplementary information in [13]. To briefly summarise, optimal weighting and combination was performed through use of a circular node autoencoder. Autoencoders are feedforward neural networks trained so that the output can reproduce the input. The outputs of the two coupled circular bottleneck nodes represent a single angular phase. The output with the minimum sum of least squares is chosen. A representation of this is shown in figure 4.3.

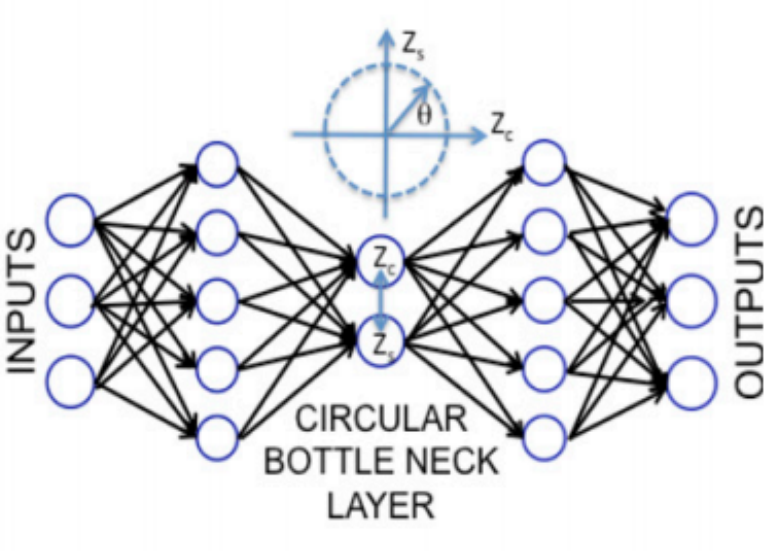


Figure 4.3: **A representation of the neural network used in CYCLOPS.** From [13].

Application to data

The RMA normalisation algorithm (discussed in section 3.1.1) was used on all raw data tested with CYCLOPS. Then they used the top 10,000 highest expressed probes, as sorted by the mean expression for all time samples. The expression $X_{i,j}$ of each probe i, sample j, and total samples N was scaled as;

$$S_{i,j} = \frac{X_{i,j} - M_i}{M_i} \tag{4.6}$$

$$\text{where} \quad M_i = \frac{1}{N} \sum_j X_{i,j} \tag{4.7}$$

Note that these definitions are different to those reported in [13], but the published definitions do not appear to be correct and the above definitions are likely to be what the authors meant to define. The $S_{i,j}$ data were expressed in eigengene coordinates $E_{i,j}$ using the methods in [133] - which is essentially just a singular value decomposition of every timecourse geneset. The number of "eigengenes" was chosen so that least 85% of variance in the data is represented.

Results

CYCLOPS is validated using mouse timecourse data (the Zhang data [9]) and human brain data [138]. They randomized the ordering of the Zhang data by time (each organ data separately), and tried to reorder it again with CYCLOPS. It worked well for high amplitude mouse tissues like the liver, kidney, and adrenal. It failed with lower amplitude tissues such as the white fat and brain regions. When allowing each organ timecourse reconstruction to use prior knowledge of what transcripts cycled in that tissue or in at least 75% of the other tissues, accuracy increased. The results for some of the mouse estimations are shown in figure 4.4A.

Anafi *et al.* used human cortex data that was taken at autopsies, and attempted to order these samples according to time of death [138]. As no appropriate human timecourse data was available to them, in order to choose the genes to use, they conducted research into evolutionary conservation and chose human homologs of the circadian mouse genes in the Zhang study for the set of genes to use. The results are shown in figure 4.4. Despite the fact that this is transcriptome of tissue that may have been dead for a few hours before being sampled, and the genes that are being used are informed from mouse data (e.g. the Clock gene is known not to be highly rhythmic in humans) the results are relatively impressive. There is good correlation of the real vs estimated scatter plot in figure 4.4B, and there is some obvious periodicity in the Chrono, Nr1d1 and Per3 probes in 4.4C.

When applying the methods to data from both healthy and cancerous tissue, they found that the method did not work for the cancerous tissue, hypothesising this is due to weak clock function in the tumours. The data they used is from samples of hepatocellular carcinoma (HCC), that also has matched control samples of healthy liver. The results are shown in figure 4.5, where the black data is from the "healthy" liver and the red data from the HCC samples. This method has no prior information, and only tries to fit ellipses in the data, but ellipses could not be found in the tumour data as they could in the healthy data.

Summary of CYCLOPS

The study highlights the advantages of this approach as;

- CYCLOPS does not need prior training (like Zeitzeiger). Supervised training approaches require a training library of samples with known circadian time - and the only human datasets available to them were the blood datasets [14] and the autopsy cortex data [138].

- It does not matter what organ, organism, or technology the data originates from.

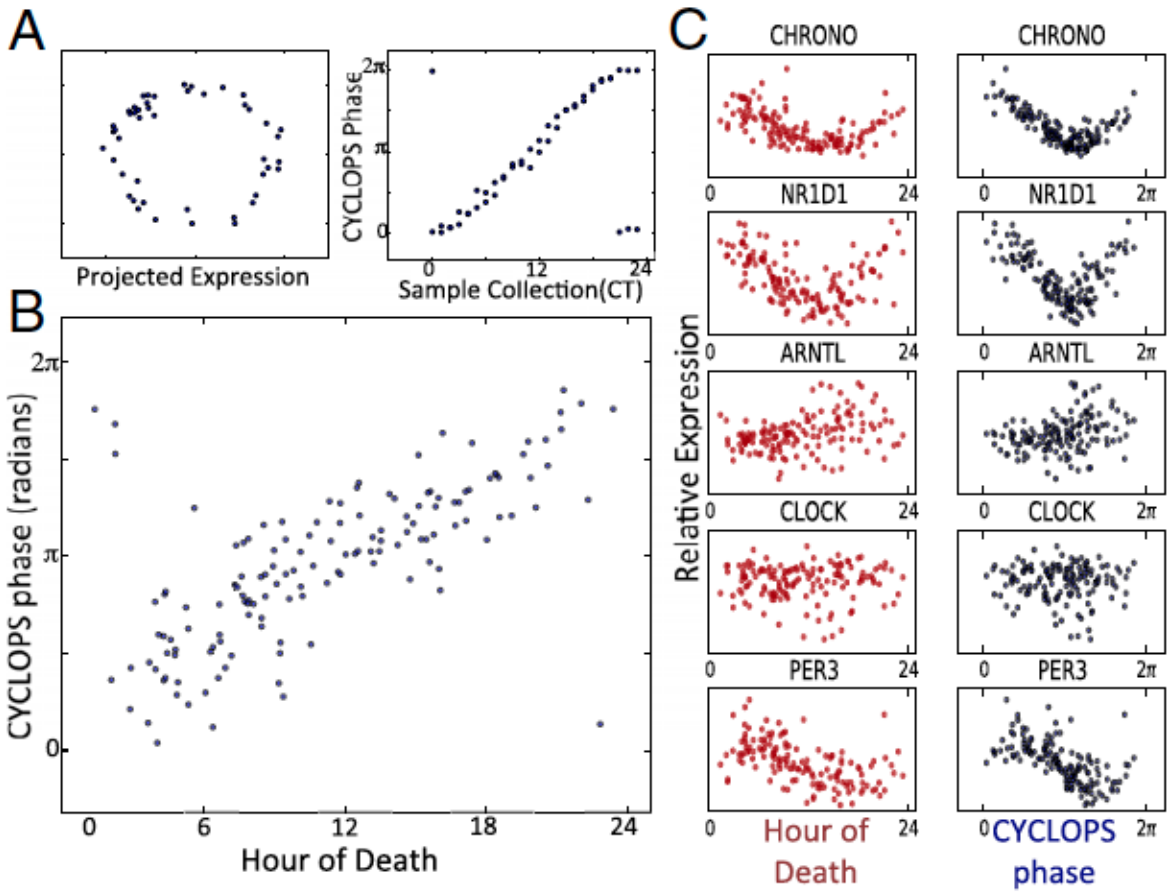


Figure 4.4: **Plots of some results from the validation of CYCLOPS.**From [13]. (A) shows the eigengenes in an elliptic shape in 2 dimensions, and the results for CYCLOPS phase and sample collection time, for Zhang data. (B) shows the results for the human dataset validation, where there is positive correlation for Hour of death vs estimated phase of sample.

The study does not highlight the disadvantages of CYCLOPS. They could be identified as;

- The technique requires large datasets and cannot handle single or small datasets in any way. It could be possible to combine comparable datasets from the beginning, but there is nothing reported on this.

- If an actual time marker does not exist, time cannot be estimated. Only the phase of each sample relative to each other sample is calculated.

- The concept of evolutionary conservation, applied by matching mouse-human orthologs, might not be appropriate. For example, Clock is rhythmic in mouse data (Zhang *et al.*), but its human ortholog has not been reported or identified as rhythmic in humans.

- Eigengenes are not always ellipsoidal, even though the dimensions are chosen to

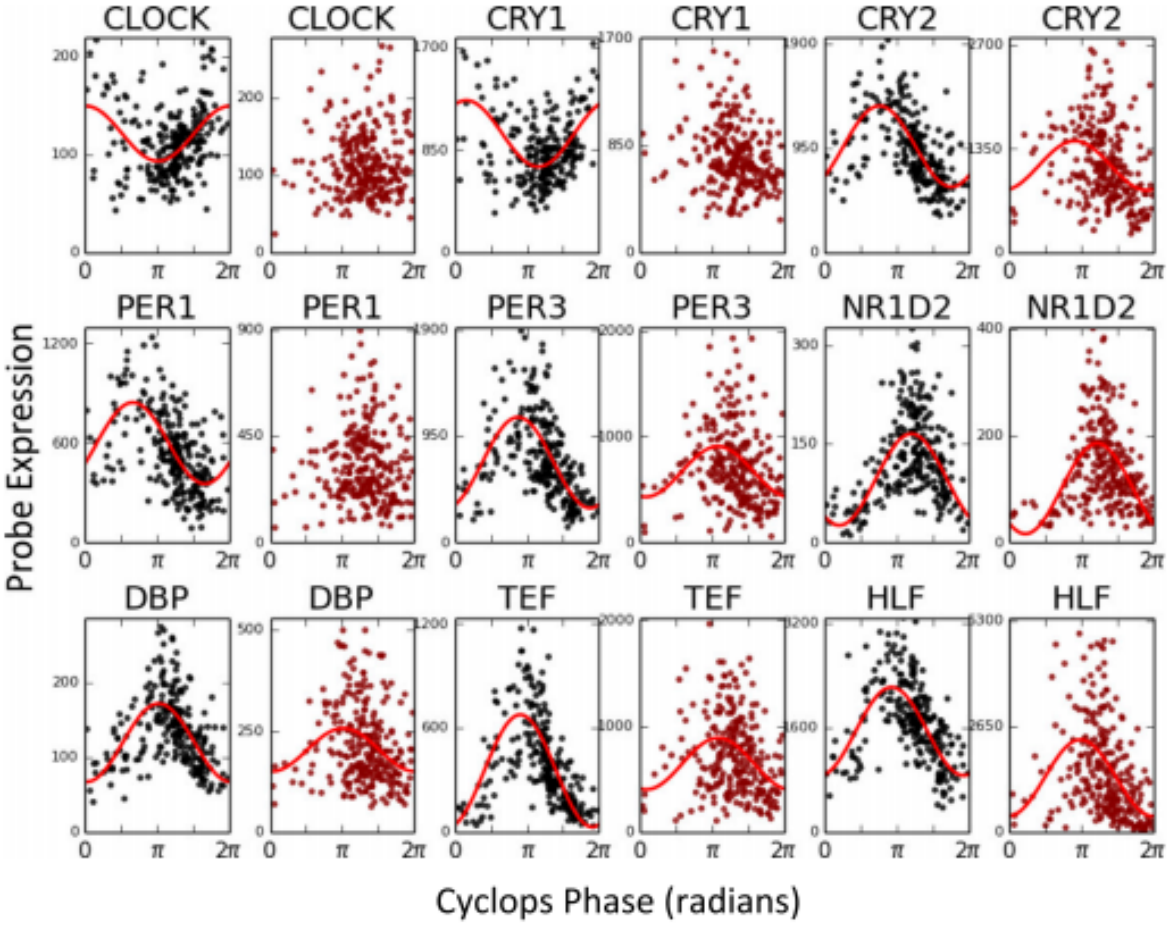


Figure 4.5: **Plots shown the results of CYCLOPS applied to human liver samples.** From [13]. Healthy human liver samples are plotted in black and HCC (cancer) samples are plotted in red. Cosines were fit to all 9 genes in the healthy data, but only 6 in the tumour data, suggesting stronger circadian rhythms in the healthy data.

maximise variation. Perfect symmetry is not expected nor observed in periodic gene expression.

- The accuracy and viabilty of the human data set used to validate the model is questionable. The data is from autopsies where time-of-death is taken as the true time for the model, but in reality the samples could have been taken hours after death. The dynamics of gene expression after death are not well understood, but this is not mentioned in the paper.

CYCLOPS is probably not a method that can be used in the one time point time telling manner that we show in this thesis. CYCLOPS, however, may be a powerful as tool when producing evidence for the dysfunction of the circadian clock in human cancers.

4.1.5 PLSR

Laing *et al.* [14] published a study that uses a large, novel human transcriptome dataset, which focuses on melatonin and gene expression in sleep deprived states. The study is not strictly speaking presenting a "timetelling" model. The authors instead attempt to predict melatonin phase from the transcriptome of blood samples from human volunteers. There is no new method developed in this study, instead a new application of an existing method, Partial Least Squares Regression [134]. This is a thorough study, where the authors also compared their method with MTTM and Zeitzeiger, but only a short overview will be given here.

Summary of partial least squares regression [134]

PLSR has similarities to principal component analysis, except that as well as decomposing the input matrix X, an output matrix Y is also decomposed, and a linear mapping between the 2 spaces is found. In this problem the dimensionality of the predictor set (transcriptome profile X) is reduced by projecting both the predictor set and the response variable (melatonin phase Y) into orthogonal latent spaces, called latent factors. These latent factors are relevant for predicting the response variable, without directly prioritising any underlying time dependency within the dataset. Factor loadings, the correlation between each feature and each factor, can then be used to select features to produce the circadian phase prediction method.

Note: PLSR works here as melatonin is also a periodic output. It would not be straightforward to adapt a PLSR model to predict time, which is linear.

Methods

mRNA abundance and melatonin data from 53 participants were collected in four sleep conditions
(i) Sleep in phase with melatonin,
(ii) Sleep out of phase with melatonin,
(iii) Total sleep deprivation, no prior sleep debt, and
(iv) Total sleep deprivation, with prior sleep debt.
This was partitioned into two groups: a training set of 329 mRNA samples from 26 participants and and a validation set of 349 mRNA samples from 27 participants. This data was collected from the following studies:

- GSE82113 - A total of 49 samples comprising 10 human subjects, for which 1/2/3 samples across multiple time-points/sleep conditions were collected.
- GSE82114 - A total of 23 samples comprising 4 human subjects, for which 2/3 samples across multiple time-points/sleep condition were collected.

- GSE39445 - A total of 438 samples comprising 26 human subjects, for which 20 samples across multiple time-points/sleep condition were collected.
- GSE48113 - A total of 287 samples comprising 22 human subjects, for which 14 samples across multiple time-points/sleep condition were collected.

For each participant in each condition, hourly melatonin samples and transcriptome samples every 3 or 4 hours were taken. This is summarised in figure 4.6.

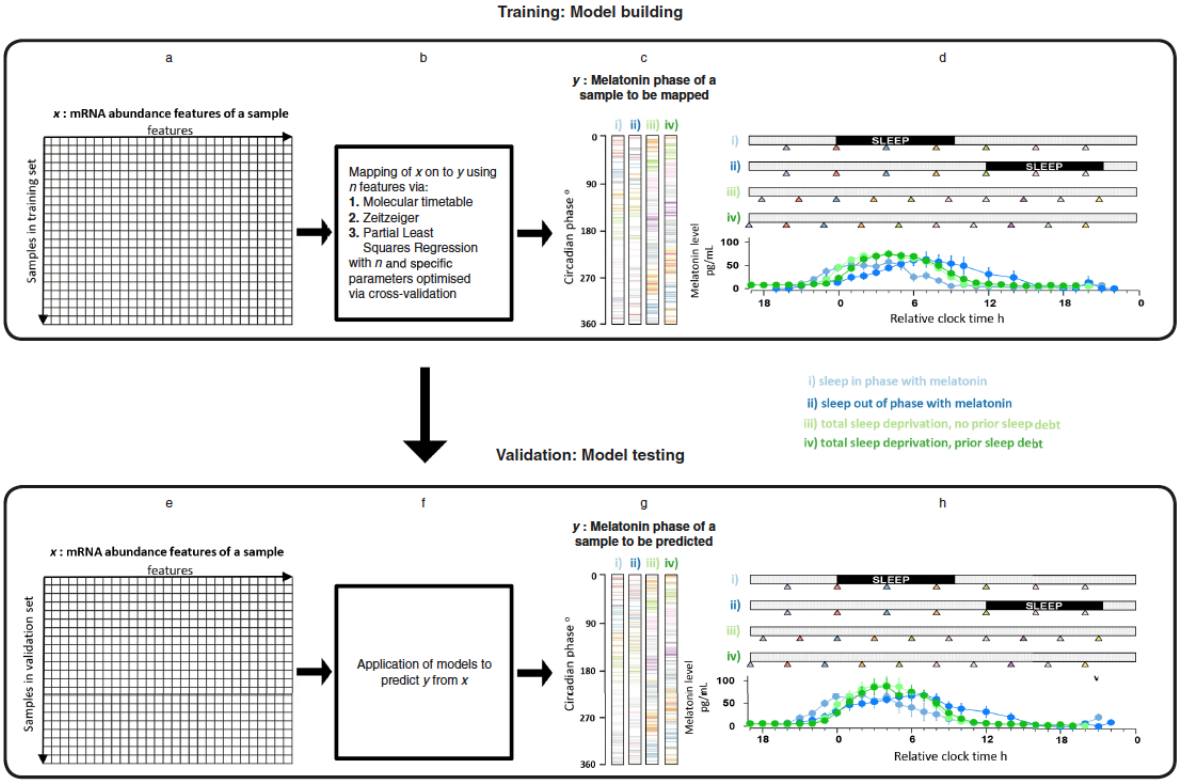


Figure 4.6: **Model design for predicting melatonin phase from transcriptome**. Three methods used were MTTM, Zeitzeiger and PLSR. Closed circles in plots represent the average melatonin profile of participants in a given experimental condition. Coloured triangles represent transcriptome sampling.

They state that they used the MTTM and Zeitzeiger, both of which performed poorly. This is not surprising as both methods were designed to tell time, and not melatonin phase. It is not explicitly stated how these methods were adapted to estimate melatonin phase, but it is likely that the time was estimated and the corresponding mean melatonin level was taken as the output. This could explain the extra error accumulation. The results are summarised in figure 4.7.

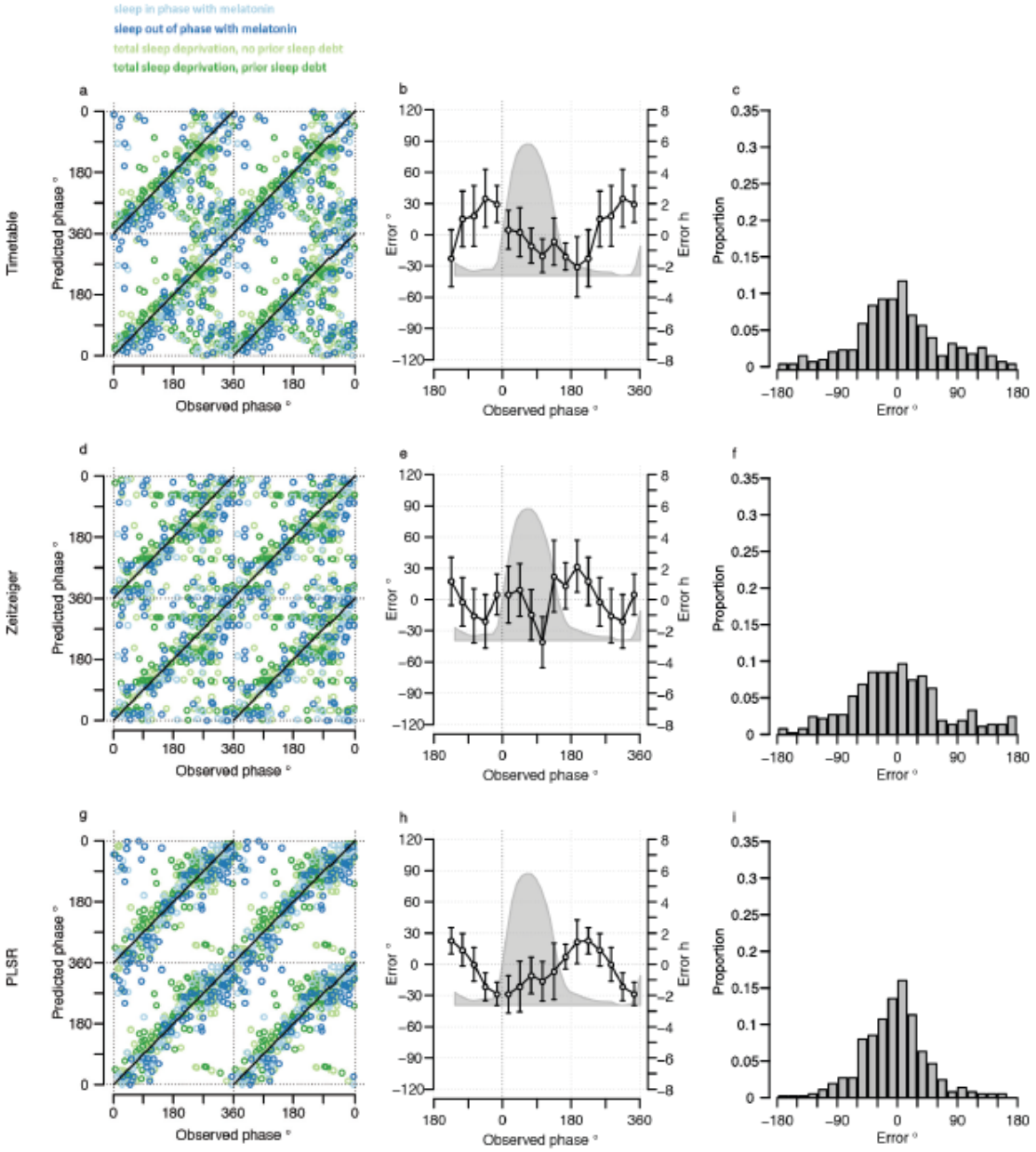


Figure 4.7: **Plots summarising the results of MTTM, Zeitzeiger, and PLSR for estimating melatonin phase.** From [14]. (a,d,g) show observed phase vs predicted phase of melatonin. (b,c,e,f,h,i) show error measurements, where the grey peak represents average melatonin phase, and 30° represents a 2 hr error.

4.1.6 ΔCCD

Shilts *et al.* [15] developed the ΔCCD clock coefficient of dysfunction. It is written by the same group as Zeitzeiger, but the ΔCCD methods are independent to Zeitzeiger. Acknowledging that Zeitzeiger cannot predict time of samples with unknown time, they develop a metric for clock dysfunction that does not rely on prior information, calling this

ΔCCD.

Using 12 clock genes (again from the Zhang data), they examine the (spearman) correlations between these genes and use these correlations as the "healthy" standard. Spearman's rank is used as a measure of correlation of two genes, denoted by the measure rho. Two in phase genes would have a correlation of 1. Two completely out phase genes would have a correlation of -1. As Spearman's test measures for monotonicity 2 genes that are $\pi/2$ shifted would have a correlation of 0, as their relationship would be circular.

The Zhang data provides a 12 by 12 table of rho coefficients which are taken as the healthy standard, as shown in figure 4.8. This agrees with all previous analyses; that Arntl, Npas2 and Clock have a similar phase, and that the phase of most of the other clock genes is 12 hours later.

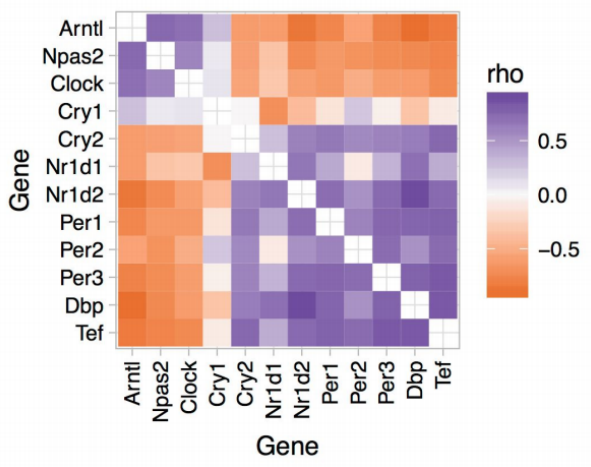


Figure 4.8: **Heatmap of reference Spearman's rank correlations for the ΔCCD metric.** The correlation for each pair of genes was calculated based on 8 mouse datasets. From [15].

These rho coefficient tables are produced for independent data sets. The ΔCCD metric is a simple euclidean distance metric that is a measure of distance between the 12×12 standard table, and the 12×12 test table.

Their results are generally that "healthy" data looks more like the standard, so has a higher ΔCCD. Data from tumours generally had a lower ΔCCD. They used the mouse standards from the Zhang data, as the healthy standard to compare to human data sets. The results for 4 cancer datasets that contain healthy controls are shown in figure 4.9. The correlations are far weaker in these healthy human datasets. One obvious reason for this is that these are human studies so all the samples will probably have been biopsied from patients in a short time frame, so not the whole timespace is being analysed here.

The paper admits that its methods are designed for simplicity, and that the correlation fitting does ignore a lot of the interesting behaviour of the clock components. The ΔCCD metric does however, provides an interesting insight into using a reference state and showing that there is *something* different with the tumour samples, and proves to be

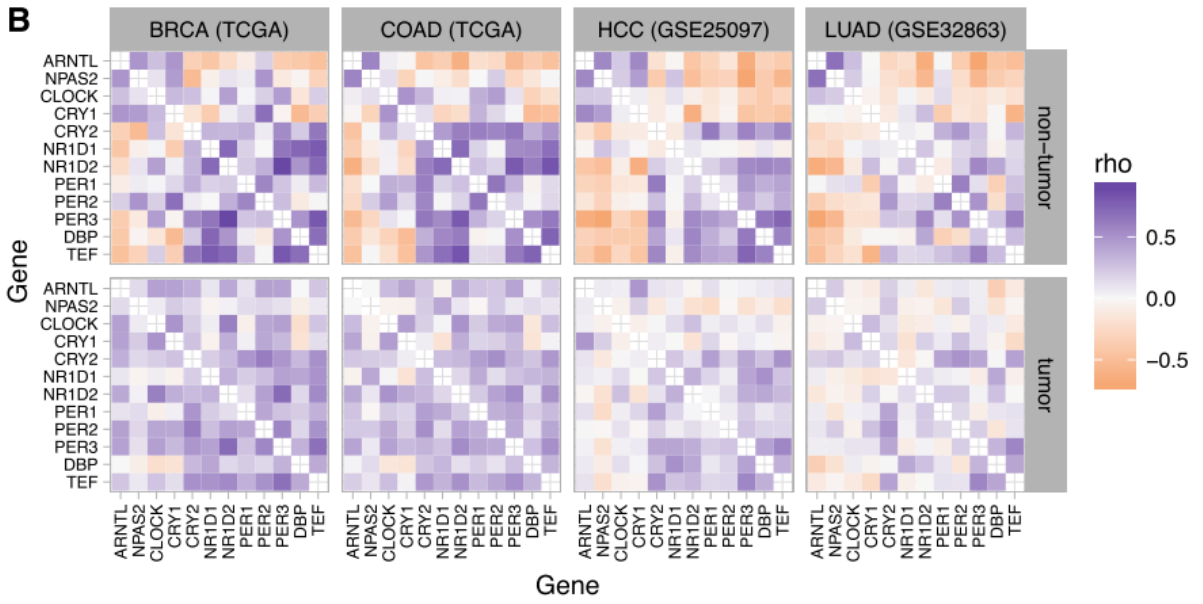


Figure 4.9: **Heatmaps of Spearman correlation between clock genes for non-tumour and tumour samples.** Two sets of data from the cancer genome atlas and two sets of data from NCBI GEO are used. From [15].

another tool that can help to provide evidence for the dysfunction of the circadian clock in cancer.

ΔCCD requires substantially sized datasets to form these correlations, and cannot test single samples - this is a limitation that we seek to overcome.

Other notable publications

Minami *et al.* used MTTM to the time using metabolites in mouse plasma dubbing this the "metabolite-timetable-method".

Kerwin *et al.* [139] expanded on the MTTM to study natural variation in the plant circadian clock using existing single time point microarray experiments from a recombinant inbred line population.

Matsuzaki *et al.* [140] built a statistical model of individual gene expression in rice integrating the effects of multiple environmental factors. Using Bayesian methods they estimated internal time based on the relationship between physical time of day and expression of multiple genes, claiming 22 minute accuracy.

4.1.7 Summary of existing methods

Although the methods presented above have value when producing evidence for the dysfunction of the circadian clock in cancer, none of the methods presented in this section are truly able to tell the time, or test the dysfunction, of a single sample. The next chapter presents a novel method that can do this.

The eigengene set-up (use of SVD) in CYCLOPS is used very similarly in the novel Time-Teller model in this thesis. Where in CYCLOPS the disordered data is projected to a an apparent ellipsoidal shape in PC space which is then used to train a neural network, Time-Teller uses a different approach. This approach also then has some similarities to Zeitzeiger, in the sense that we use Gaussian approximations, splines, and maximum likelihoods to make an estimate for time.

4.2 Time-Teller: a novel time-telling algorithm

The aim of the algorithm presented in this section is to be able to tell the time of a single sample by a subset of its trancriptome, without any prior knowledge or dependence on the sample existing as part of a larger set of data. In this chapter, different versions of Time-Teller are shown for synthetic data, mouse data, and human data.

A typical dataset that can be used to train Time-Teller would consist of multiple observations of timecourse transcriptome. The training data consists of G features (gene expression values), for N observations (simulation, organ, or individual), for T time points over a total length of time p (which is also the period or multiple of the period of the system).

In this approach we apply SVD to reduce the dimension of the data. This will define a projection of each normalised G-dimensional data vector into a smaller number d of dimensions (d will be 3 in our case). These 3 dimensions are equivalent to eigengenes [133]. For each of the observed times t_j we fit a d-dimensional multivariate Gaussian distribution ϕ_{t_j} to the set of vectors given by projecting the vectors $X_{t_j,n}$ from time t_j into d dimensions. We need to reduce the dimension in order to ensure that there are enough data points to enable a good fit for ϕ_{t_j}. As we only have data for discrete time points, splines are used to extend the ϕ_{t_j} to all times t to get distributions ϕ_t. Then given a new data vector Z we can project it into the d dimensions to study the probability $p(Z_t)$ that Z is of time t using ϕ_t. We can then estimate the time of Z by maximising the probabilities $p(Z_t)$ with respect to t. The method will be initially outlined in mathematical notation, with some sketches for aid of explanation. The method is further explained to the reader via examples using simulated data, before validation using real data is presented.

4.2.1 Model outline

Choice of training genes

The G genes are identified by the methods in chapter 3, where the most rhythmic and synchronised genes were ranked, and the top G genes chosen to be training genes. Choosing the training genes in this way provides very similar results to using methods in supervised principal components, or sparse principal components to identify the training genes [12]. The idea behind all three methods is that the genes used should show a large signal to noise ratio, and have a similar profile in all timecourses.

Normalisation

For each observation $n = 1, ..., N$ and each time time t_i where $i = 1, ..., T$, the data for each set of G expression levels is stored in vectors $X_{t_i,n} \in \mathbb{R}^G$. Each $X_{t_i,n}$ is then normalised to have a mean of 0 and standard deviation of 1, resulting in the vector $\bar{X}_{t_i,n}$.

As each sample is treated individually under both vector normalisation and the initial fRMA normalisation, there is no time-course batch bias in the Time-Teller method. This is not the case in many of the time-telling methods discussed in section 4.1. This means that every vector $\bar{X}_{t_i,n}$ is independent, and the only information in the vectors $\bar{X}_{t_i,n}$ is the relative ordering of expression of each gene in each sample. This is what allows Time-Teller to do real single time-point time-telling, and why the use of the same transcriptome quantification technology for the training and test samples is vital.

Principal component projections

The decomposition of the data into principal component space (forming the eigengenes [133]) makes up the initial steps of Time-Teller. Global PCA will refer to the standard PCA (see appendix A). Later in this section, the local PCA will be introduced and used in Time-Teller. However, we first start with the explanation of the method using global PCA, as the applications are almost identical, and easily adapted to local PCA.

The $G \times (TN)$ master matrix $\mathcal{Y}$ has as its columns all the column vectors $X_{t_i,n}$. This is decomposed into principal components using SVD:

$$\mathcal{Y} = \mathcal{U}\mathcal{D}\mathcal{V}^{\mathbf{T}} \tag{4.8}$$

where a set of G orthogonal principal components $U_1, ..., U_G$ form the columns of $\mathcal{U}$ and corresponding singular values $\sigma_1 \geq, ..., \geq \sigma_G > 0$ forming the diagonal entries of a diagonal matrix $\mathcal{D}$.

Now if $\mathcal{U}_d$ is the matrix consisting of the first d columns of $\mathcal{U}$ then;

$$\mathcal{Q}_{t_i,n} = \mathcal{U}_d^T \bar{X}_{t_i,n} \tag{4.9}$$

is the projection of $\bar{X}_{t_i,n}$ onto the first d principal components.

Fitting a Gaussian distribution to each set of time projections

Let each set of time labelled projections be denoted by $\mathcal{Q}_{t_i}$, i.e. $\mathcal{Q}_{t_i} = \{\mathcal{Q}_{t_i,n}\}_{n=1,..,N}$. We assume that $\mathcal{Q}_{t_i}$ are normally distributed and fit a multivariate Gaussian distribution Φ_{t_i} to the vectors in $\mathcal{Q}_{t_i}$. This is calculated by estimating the component mean μ_{t_i}, and covariance matrix Σ_{t_i}using an iterative expectation-maximization algorithm. We use the MATLAB function *fitgmdist* to do this (see Appendix C).

In this thesis, we take the number of dimensions, d, to be 3, hence each Φ_{t_i} is a 3D Gaussian capturing the first 3 PCs, $\mu_{t_i} \in \mathbb{R}^3$, and Σ_{t_i} is a 3×3 symmetric matrix. Using three dimensions captures as much variance as possible, whilst also allowing a statistically significant Gaussian fit with limited data (there are only 8 - 10 data points per Gaussian

in the training data for this thesis).

Each set of projected points $\mathcal{Q}_{t_i}$ belonging to a specific time t_i are now represented by a Gaussian distribution Φ_{t_i}.

Fitting a spline to Gaussian centroids

The result of this projection of periodic data, will result in the T Gaussians Φ_{t_i} being distributed along a periodic curve in space. This curve can be found using a cubic smoothing periodic spline through the means of each Gaussian. We make the assumption that the spline approximates the mean of the state of the system at the times in between the t_i.

A cubic smoothing spline fits a cubic polynomial to data, where each interval (between data points) is described by a cubic polynomial with local coefficients, where smoothness at each change of coefficients is ensured. Coefficients are determined by a maximisation of smoothness [141]. As we have so few time points, to fit the periodic spline to in 3D space, we fix that the spline will exactly pass through the mean of each Gaussian. The function *csape* (cubic spline interpolation with end conditions) in MATLAB, fits a spline allowing specification that data is periodic and that it passes through all data points[3]. The resulting function $\mu(t)$ where $t \in [0, p)$ is continuous and periodic and passes through all data points, i.e. $\mu(t)$=$\mu(t_i)$. Details of *csape* are given in appendix C.

Figure 4.10 shows a representation of the system in this form. The black spline is $\mu(t)$ and each ellipsoid is a 90% boundary of each Φ_{t_i} (where here $T = 6$ and $p = 24$).

Estimating intermediate covariance

Figure 4.10 shows that there are areas along the spline where the covariance of the data at that time is not known. The covariance changes significantly over time (the distribution is heteroscedastic), so estimating these intermediate regions is necessary instead of using a constant covariance (homoscadastic) approach that was used in Zeitzeiger [12]. TThe method adopted for this was to fit splines through each individual entry in the covariance matrices for a smooth transition. A shape preserving smoothing cubic periodic spline is separately fit thought each of the nine entries (six unique) in Σ_{t_i} for all i to result in the continuous function $\Sigma(t)$ defining the covariance for all times. A piecewise cubic hermite interpolating polynomial spline is used in this case. This type of spline is shape preserving, i.e. the second derivative continuity does not need to be a feature. This is suitable as, for example, if two co-variance matrix entries were identical for two consecutive time Gaussians, the hermite spline allows the value of the joining spline to stay the same in the space between, however a standard spline would enforce some change. This is also an interpolating spline so the spline passes through all points. This was done in MATLAB

[3]Csape is equivalent to using a standard periodic cubic spline fit with interpolation parameter at its maximum value (e.g.$p = 1$). The number of the knots is specified by the number of time points of the data, T, and the position is fixed at each μ_{t_i}.

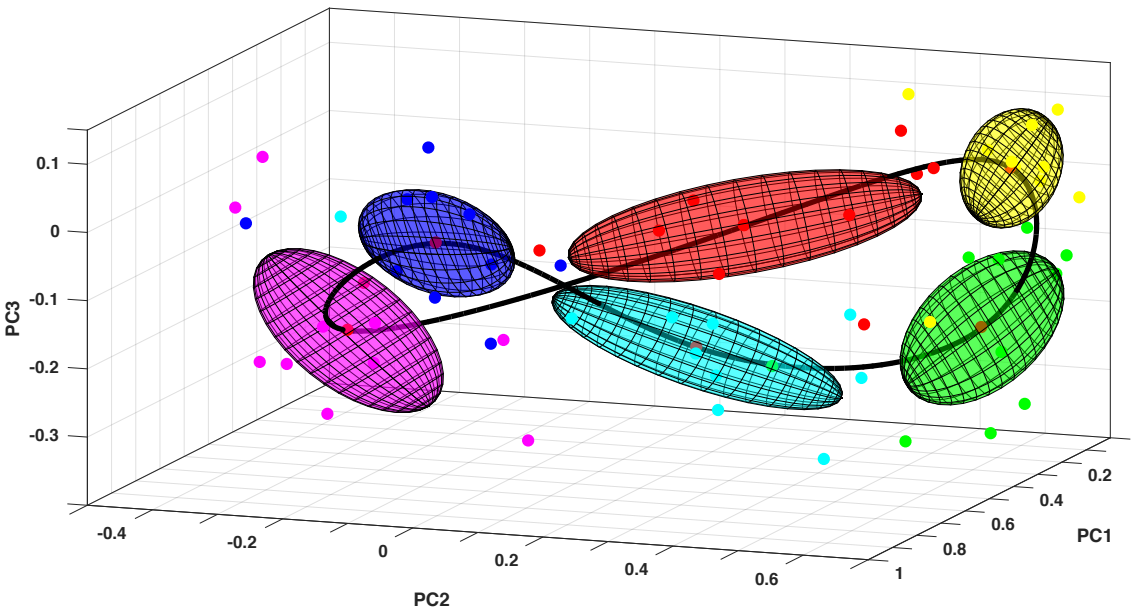


Figure 4.10: **A representation of 6 estimated ellipsoids along a spline that represents mean movement through time.** Each Gaussian is set to a 90% boundary. This plot has used real data (Bjarnason data), where the differences in covariance for each time is clear.

using the *pchip* function. See appendix C for details, and also for details of how positive definiteness was ensured for all $\Sigma(t)$.

Each time $t \in [0, p)$ is now represented by a 3D Gaussian:

$$\Phi_t = \mathcal{N}(\mu_t, \Sigma_t) \tag{4.10}$$

Now, the probability density function (PDF) of each 3D Gaussian allows the calculation of the probability that any sample x is of time t. For a fixed $t = \tau$, the PDF of Φ_τ for any vector x with projection Q_x is:

$$f(x;\tau) = \frac{1}{(2\pi)^{3/2}|\Sigma(\tau)|^{1/2}} exp\left(-\frac{1}{2}(\mathcal{Q}_x - \mu(\tau))^T \Sigma(\tau)^{-1}(\mathcal{Q}_x - \mu(\tau))\right) \tag{4.11}$$

Maximum likelihoods

For a given data set and probability model, a maximum likelihood estimate, or MLE, is the set of a model's parameters that give the highest probability for the observed data [142]. We calculate the MLE for the single parameter in our model, time t, for any sample Z using the following steps.

A single sample Z of length G and unknown time, is normalised to have mean 0 and standard deviation 1 and projected into latent variable space with the principal

components defined by the training data;

$$\mathcal{Q}_{\bar{Z}} = \mathcal{U}_d^T \bar{Z} \tag{4.12}$$

The likelihood function $\mathcal{L}$ is defined by each f in (4.11) for all times t. The likelihood that a fixed sample Z is of time t is the function:

$$\mathcal{L}(Z|t) = \frac{1}{(2\pi)^{3/2}|\Sigma(t)|^{1/2}} exp\left(-\frac{1}{2}(\mathcal{Q}_{\bar{Z}} - \mu(t))^T \Sigma(t)^{-1}(\mathcal{Q}_{\bar{Z}} - \mu(t))\right) \tag{4.13}$$

The maximum likelihood time estimation t_{pred} is then simply the maximum of the likelihood $\mathcal{L}(Z|t)$;

$$t_{pred} = \arg\max_{\tau \in [0,p)}(\mathcal{L}(Z|\tau)) \tag{4.14}$$

Note that in practice, t cannot be treated as continuous, and has to be discretised in order to generate all of the above quantities. The intervals that t is sampled in are chosen so there is a Gaussian distribution every 5-10 minutes. In Time-Teller this is a parameter that can be changed for the desired resolution, but there is little value added with making this interval smaller.

Figure 4.11 is a sketch showing what the likelihoods represent. This figure explains the likelihoods without the spline connected spaces, but the same explanations would hold for smooth likelihood curves. A projection outside the 2D doughnut distribution space which is a significant distance away from all distributions, would represent a weak clock. This likelihood curve has a peak, but with low height. A projection into the middle of the doughnut distribution shows no clock, as all times are equally likely, with a flat likelihood curve. A projection near to the centre of a single time Gaussian shows a high likelihood, with a single high peak. A more detailed analysis of this is discussed in the next chapter.

4.2.2 Local principal components

Projections using global SVD used a single projection to represent it in 3 dimensions. We could instead treat the data corresponding to each t_i separately using only the data for each t_i to determine the projection and thus finding different projections for each i. The potential advantage of this is that these local data sets are more likely to admit a good 3-dimensional linear representation than the global one since they are more localised and because the local representation might twist in the full space.

This local approach to SVD analysis of data has been proposed before (e.g. in [143]), but is an unorthodox approach. As far as we know, the local PCA approach has not been used to explore circadian data before. We will see that a local approach is an improve-

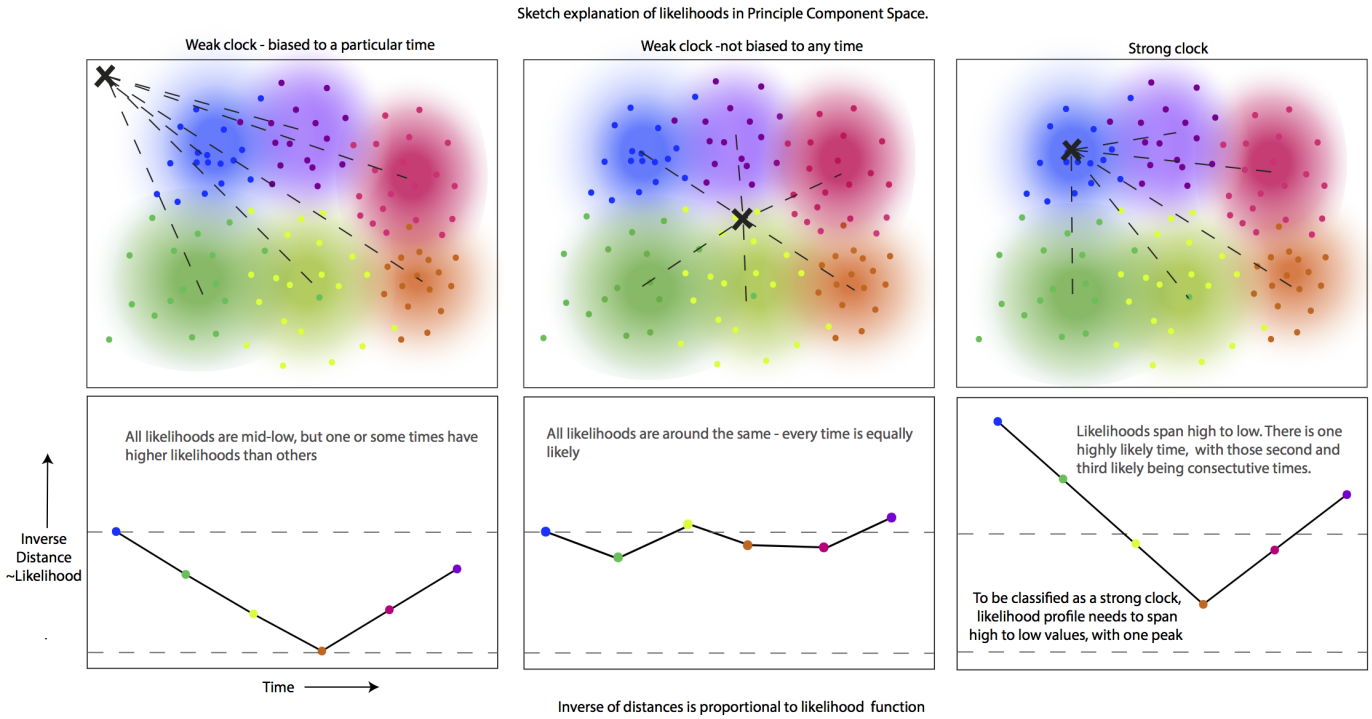


Figure 4.11: **A sketch likelihood shapes in a discrete projection space.**

ment to the global approach, as it helps to overcome problems that the symmetric elliptic distributions create.

Let $\mathcal{X}_{t_j}$ be the matrix whose columns are the N normalised vectors $\bar{X}_{j,n}$ for $j = 1, ..., T$. Each $\mathcal{X}_{t_j}$ matrix has dimensions $G \times N$, where there are T different $\mathcal{X}_{t_j}$. The SVD of each of these matrices provides the set of local principle components $\mathcal{U}_j$.

Projections

Every normalised vector $\bar{X}_{t_i,n}$ is projected using the matrix $\mathcal{U}_{j,d}$ consisting of the first d columns of $\mathcal{U}_j$. This results in $T \times T$ sets of projections, $\mathcal{Q}_{i,j}$ for $j = 1, ...T$ and $i = 1, ..., T$ that are produced using each set of time principle components with

$$\mathcal{Q}_{i,j} = \mathcal{U}_{j,d}^T \bar{X}_{t_i,n} \tag{4.15}$$

where j represents the time of the local projection vector and i represents the known time of the sample being projected.

Gaussian distribution fitting

As in the global case, but now for each i separately, we fit a multivariate Gaussian distribution $\Phi_{i,j}$ to the vectors $\mathcal{Q}_{i,j}$ for $j = 1, ..., N$. We again use the MATLAB function

fitgmdist to do this. The mean and covariance matrix of $\Phi_{i,j}$ are denoted by $\mu_{i,j}$ and $\Sigma_{i,j}$ respectively.

Fitting a spline to Gaussian centroids

Smoothing splines for the means and covariance matrices are found in the same way as for global PCA. Let the periodic smoothing cubic spline through $\mu_{i,j}$ for all values of i be represented by $\mu_j(t)$. Let the shape preserving smoothing cubic spline through each entry of $\Sigma_{i,j}$ for all values of i be represented by $\Sigma_j(t)$.

Projections of test samples

A single sample $\bar{Z}$ of length G and unknown time, which has been normalised to have mean 0 and standard deviation 1, is projected into latent variable space with each set of local principal components;

$$\mathcal{Q}_{j,\bar{Z}} = \mathcal{U}_{j,d}^T \bar{Z} \tag{4.16}$$

Now, the PDF of each 3D Gaussian allows us to generate the likelihood from the jth local PC using each time distribution $\mathcal{Q}_{j,\bar{Z}}$, that sample Z is of time t;

$$L_j(Z;t) = \frac{1}{(2\pi)^{3/2}|\Sigma_j(t)|^{1/2}} exp\left(-\frac{1}{2}(\mathcal{Q}_{j,\bar{Z}} - \mu_j(t)))^T \Sigma_j(t)^{-1}(\mathcal{Q}_{j,\bar{Z}} - \mu_j(t)))\right) \tag{4.17}$$

As we have T likelihoods, we are able to combine them to find the overall likelihood. To do this, we simply take their geometric mean over time;

$$\mathcal{L}(Z|t) = \left(\prod_{j=1}^{T} L_j(Z;t)\right)^{\frac{1}{T}} \tag{4.18}$$

so that the maximum likelihood estimate t_{pred} for sample Z using local PCs is

$$t_{pred} = \arg\max_{t\in[0,p)} \mathcal{L}(Z|t) \tag{4.19}$$

4.3 *in silico* Time-Teller

The Relogio ODE model was transformed into a stochastic model, and trajectories simulated using the Gillespie algorithm in chapter 3. A parameter representing system size (Ω) was incorporated by scaling parameters with units of concentration. We will use this model to create some dummy data, and use this data to show how the local approach is superior to the global approach. To generate some *in silico* dummy data to test the

model, we generated 90 trajectories of the stochastic Relogio model for both $\Omega = 1000$ and $\Omega = 100$. 5 sets of 9 simulations were generated for the training data, and 5 sets of 9 simulations were generated for test data, and the resulting data was saved in matrices of 12 variables by 12 time-points (representing 2 hour samples over 24 hours).

Time estimations for $\Omega = 1000$

The stochastic data for $\Omega = 1000$ was used to make a training model for both the global PCA and local PCA methods outlined above. Figure 4.12 shows the training data projected into the first 3 global PCs, coloured by time. The ellipsoidal distribution of the data is obvious. As there is low noise in this data, the 12 2-hour separated time points cluster closely with little overlap.

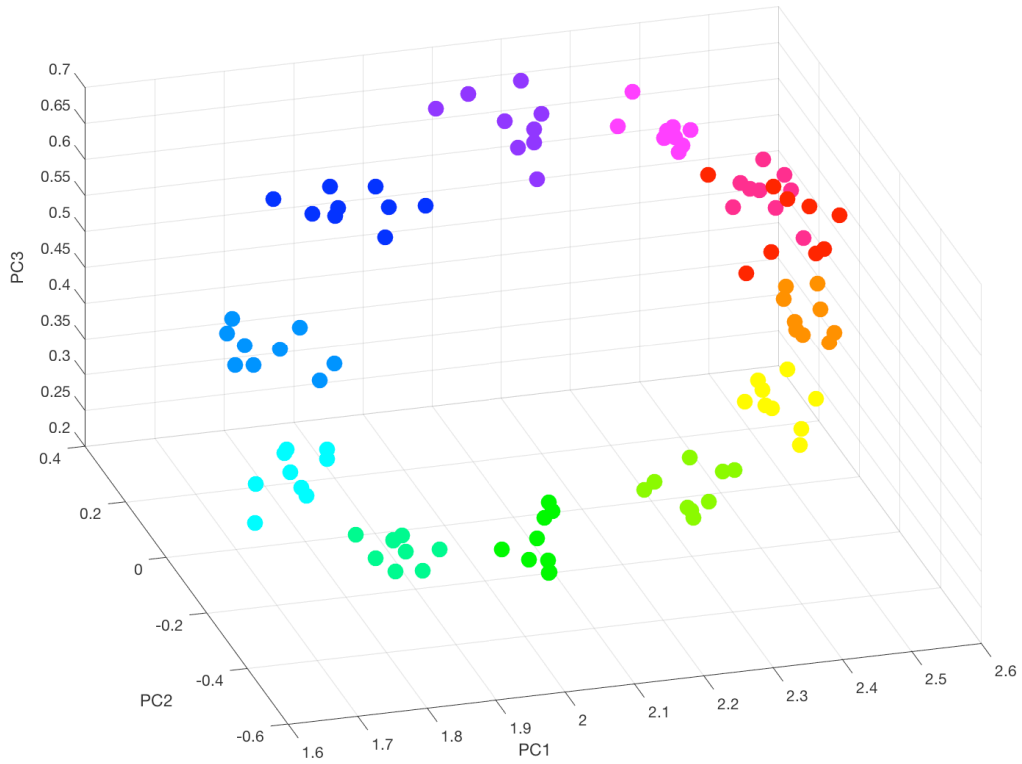


Figure 4.12: **Global PCA plot with each data point coloured by time, of simulated data.** Variance is very low in this low noise ($\Omega = 1000$) data. Colour represents time.

Figure 4.13 shows the training data projected into the first 3 PCs of every local time PC space, coloured by time. Each of the 12 local time principal component spaces show similarly ellipsoidal and time clustered data.

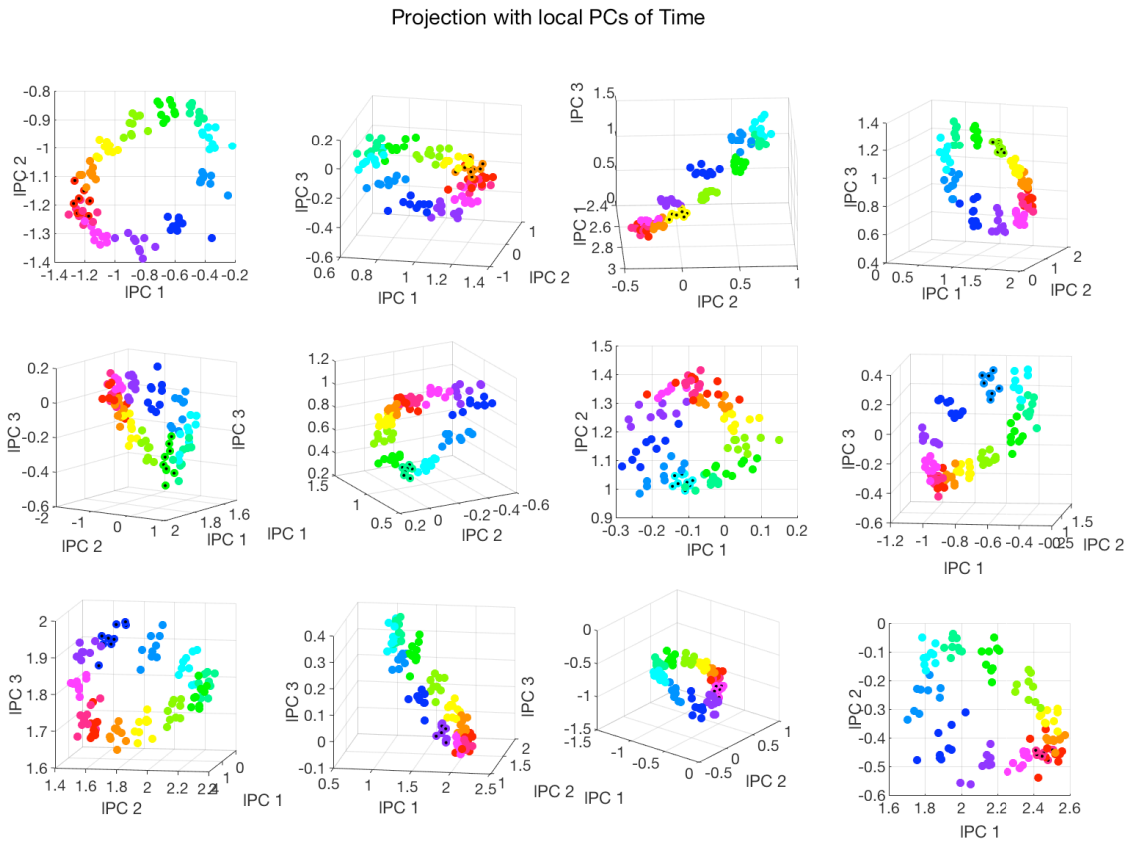


Figure 4.13: **Projections using local PCs from first 12 time points, using simulated data.** Each colour represents a time, and the points with additional black dots represent the time data for which the local PCs were calculated. Axis are manually turned to show the third dimension where the "hole" is not apparent in 2D.

The projection space using the one universal set of PCs appears to be less noisy and have more distance between opposite times, as we would expect: the projections are closer to the perfect ellipse of the underlying limit cycle. Although one might think that this is desirable, in this situation symmetry can be an adversary.

Take the following example using the dummy data. One simulated transcriptome of unknown time is projected into the test space of both the global PCs and the local PCs, and the PDF value is used to try to calculate the maximum likelihood time of that sample. The resulting likelihoods are shown in figure 4.14. The "real time" of the sample is CT4, the MLE of the local PC method is around CT5 and the MLE of the global PC method is either CT5 or CT17.

This particular data point will have been projected fairly centrally to the ellipsoid in figure 4.12, so that there is large and equal distance to CT5 and CT17 side of the elliptic distribution. The local PCs can overcome the issue of symmetry, as each of the time PCs

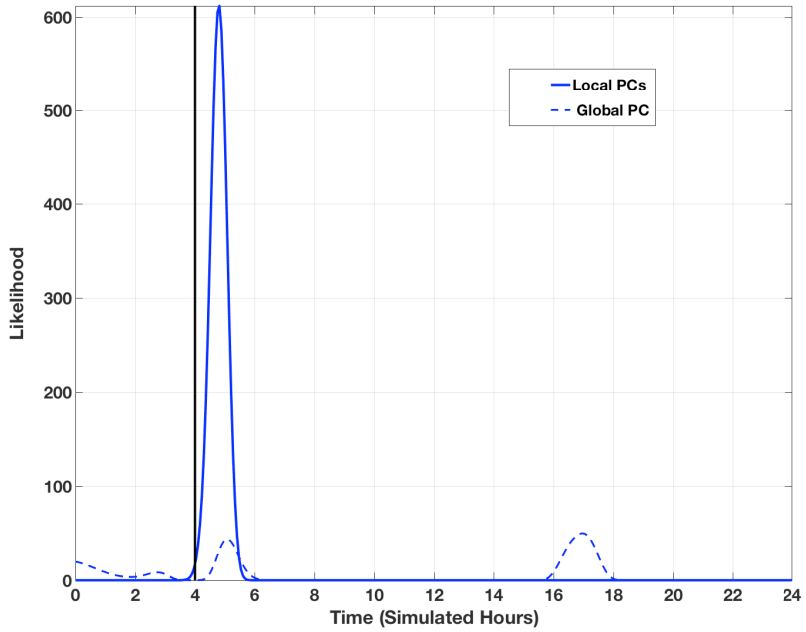


Figure 4.14: Example likelihood curves for the same data point for global and local PCA. Global PC likelihood is shown with the dashed line and local PC likelihood is shown in bold, for the same training and test data. Black line shows the true time of the sample.

generate a different likelihood, which can be (geometrically) averaged to find the overall likelihood, as shown in figure 4.15.

A simple way of measuring accuracy of an estimation is to use the least squares measure of a fit to $x = y$, where all of the data is transformed to have a maximum distance of 12 hours from the $x = y$ line for the R squared calculation. This is plotted in figure 4.16.

The R-squares for the global and local PC methods are summarised for 5 independent sets of training and testing data, in table 4.2. There is low noise in this data so the estimations are generally very accurate. Although in 3 sets of data the results are identical, estimations using global PCs are less accurate for dataset 1 and 4 (run 1 is what is shown in the figure 4.16.)

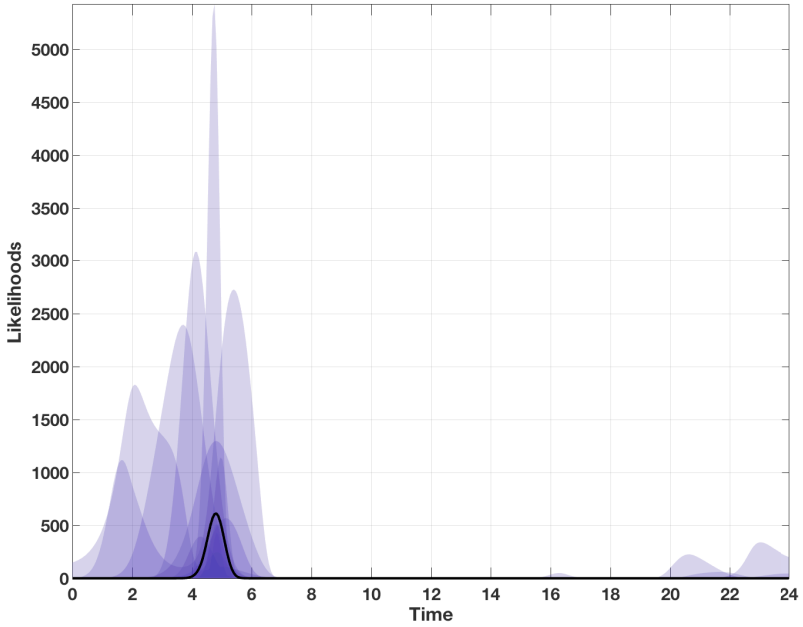


Figure 4.15: **Plot showing how individual local PC likelihoods are averaged to find the combined likelihood.**The likelihood curves for the local PCs are (geometrically) averaged, to get the overall likelihood. This helps to overcome problems with symmetry.

Time estimations for $\Omega = 100$

The advantage of using local PCs and not global PCs is even more apparent when using a noisy data set. The exact same analysis was performed as above, except for with 90 simulations of the Relogio model using $\Omega = 100$. The accuracy of the calculations fell, as expected, as shown in table 4.3, but the local PC results are significantly better than the global PC method. The estimated versus real time plot for run 3 are shown in figure 4.17.

Final comments on method development

It is clear that in a low noise system, one PC is the easier option. However, multiple projections allow for the combination of likelihood curves, minimising the risk of false estimations that arise from symmetrical distributions. The rest of this thesis will use the local PC method.

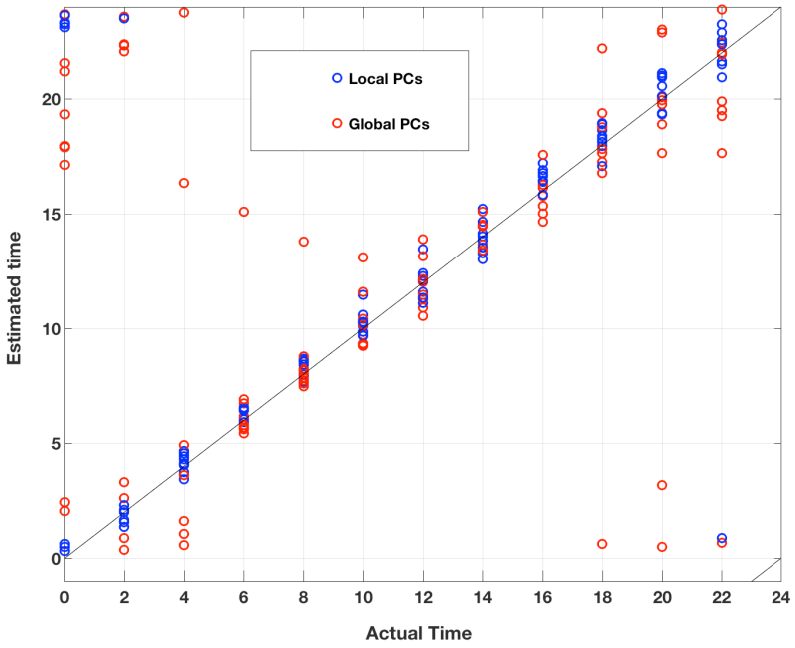


Figure 4.16: **Scatter plot showing estimated versus real time for low noise simulations, using local and global PCA.** Correlation is generally very high, with some obvious symmetry issues with the global PC estimations. The red point at $()8, 13)$ is the estimate resulting from the dotted likelihood in figure 4.14.

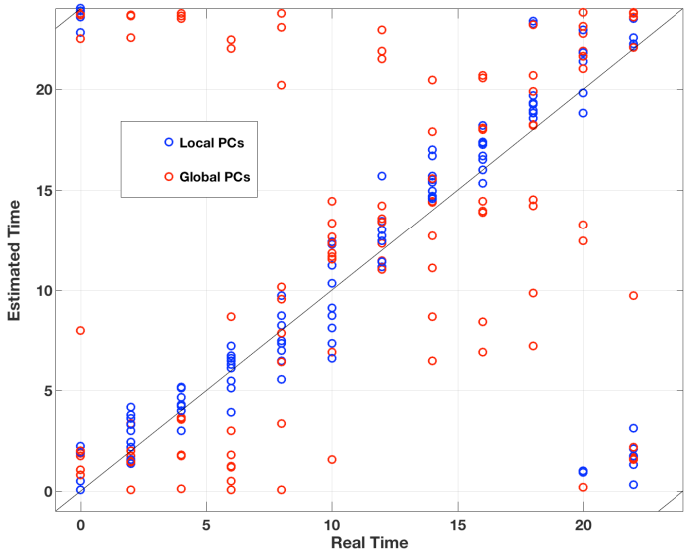


Figure 4.17: **Scatter plot showing estimated versus real time for high noise simulations.** The training and test data is identical. The method for multiple simulations is more accurate than the method using only one.

Run	Local PCs	& Global PCs
1	0.99	0.81
2	0.99	0.99
3	0.98	0.98
4	0.99	0.81
5	0.99	0.99

Table 4.2: **Table showing R-square values for linear fit to $x = y$ of estimated versus real time, for $\Omega = 1000$ simulated data.** Values are generally accurate due to low noise. R-Squares are similar except for runs 1 and 4 , where accuracy of local PC method is higher than for global PC method.

Run	Local PCs	Global PCs
1	0.69	0.45
2	0.74	0.62
3	0.80	0.60
4	0.92	0.64
5	0.79	0.43

Table 4.3: **Table showing R-square values for linear fit to $x = y$ of estimated versus real time, for $\Omega = 100$ simulated data.** R-square values for linear fit to $x = y$ of estimated versus real time. Values vary due to high noise in the simulated data. R-squares for local PCs are significantly higher than for global PC.

4.4 Mouse Time-Teller

The training data for the mouse Time-Teller is the Zhang data from the 8 tissues and 11 probes determined in chapter 3. A leave-one-organ-out approach is taken here to validate the model. As the mouse data spans over 48 hours, the decision can be made to pool data so that all data is between 0-24 hours, or there can be duplicate times from two periods. Here, the latter is chosen, so that there are 24 local principal components from 24 times between 0-48 hours.

The results of a leave-one-organ-out approach are shown in the correlation plot in figure 4.18. The real time is plotted along the x-axis, where Time-Teller's estimate is plotted along the y-axis. The black lines represent a perfect estimation, where the estimation can be ± 24 hours. The accuracy of the estimations is apparent, and no phase shift was found for any of the tissues. The use of the full 48 hour space allows us to observe, for example, that the transcriptomic time signature at CT18 can be estimated to be at CT32. This means that there is no significant change to the circadian clock genes after the mice have been in the dark for an extra 24 hours.

All of the likelihood curves that gave the MLEs are shown in figure 4.19. The shape of each likelihood is irregular due it being the combination of 24 likelihoods from each local PC. The red lines represent the real times of the samples, and it is apparent that the likelihood curves are indicating a time estimation around this real time.

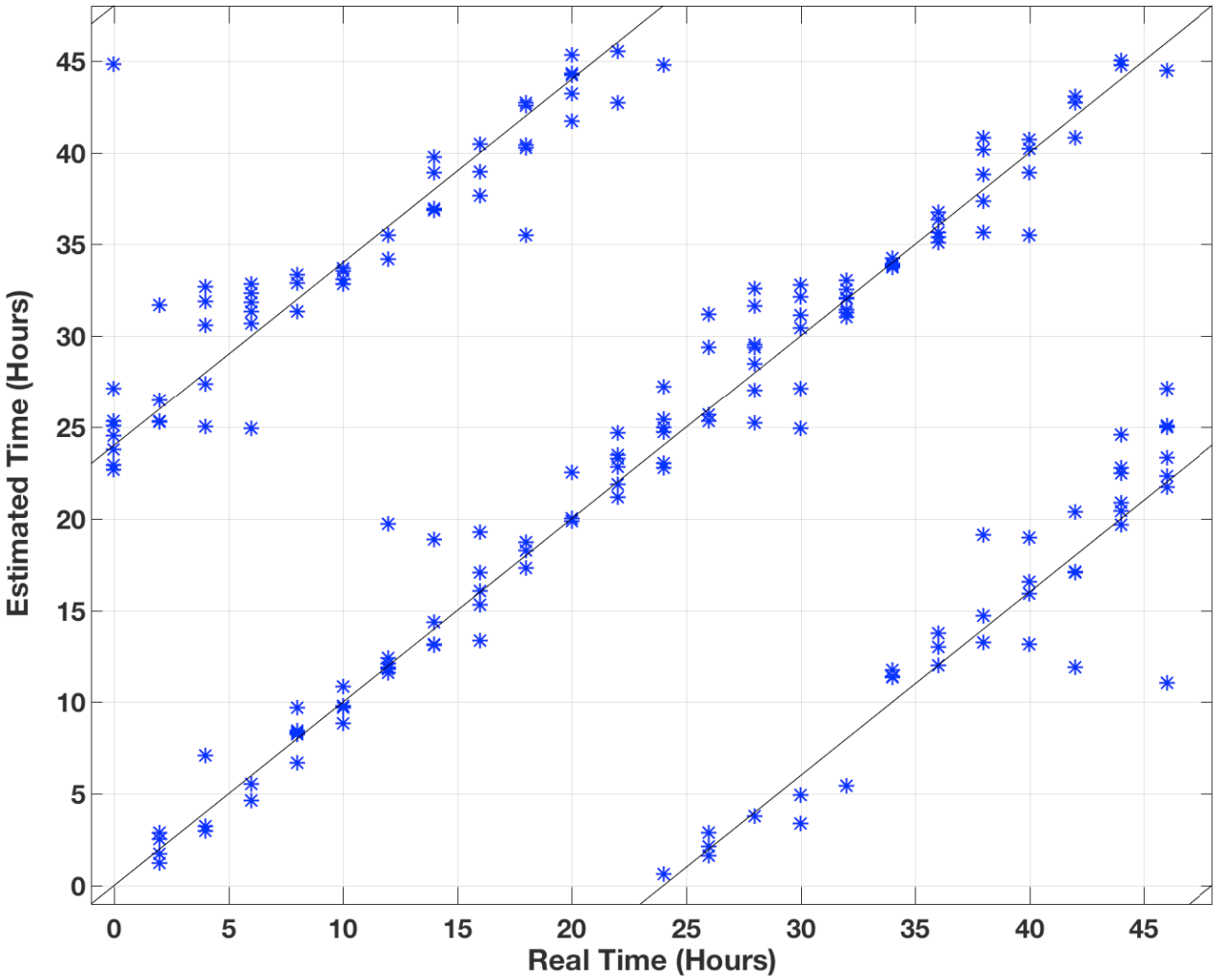


Figure 4.18: **Correlation plot for actual versus predicted time, for the results of the Time-Telling model on the Zhang data using 11 probes and 8 organs.** This is the result of a leave-one-organ-out approach.

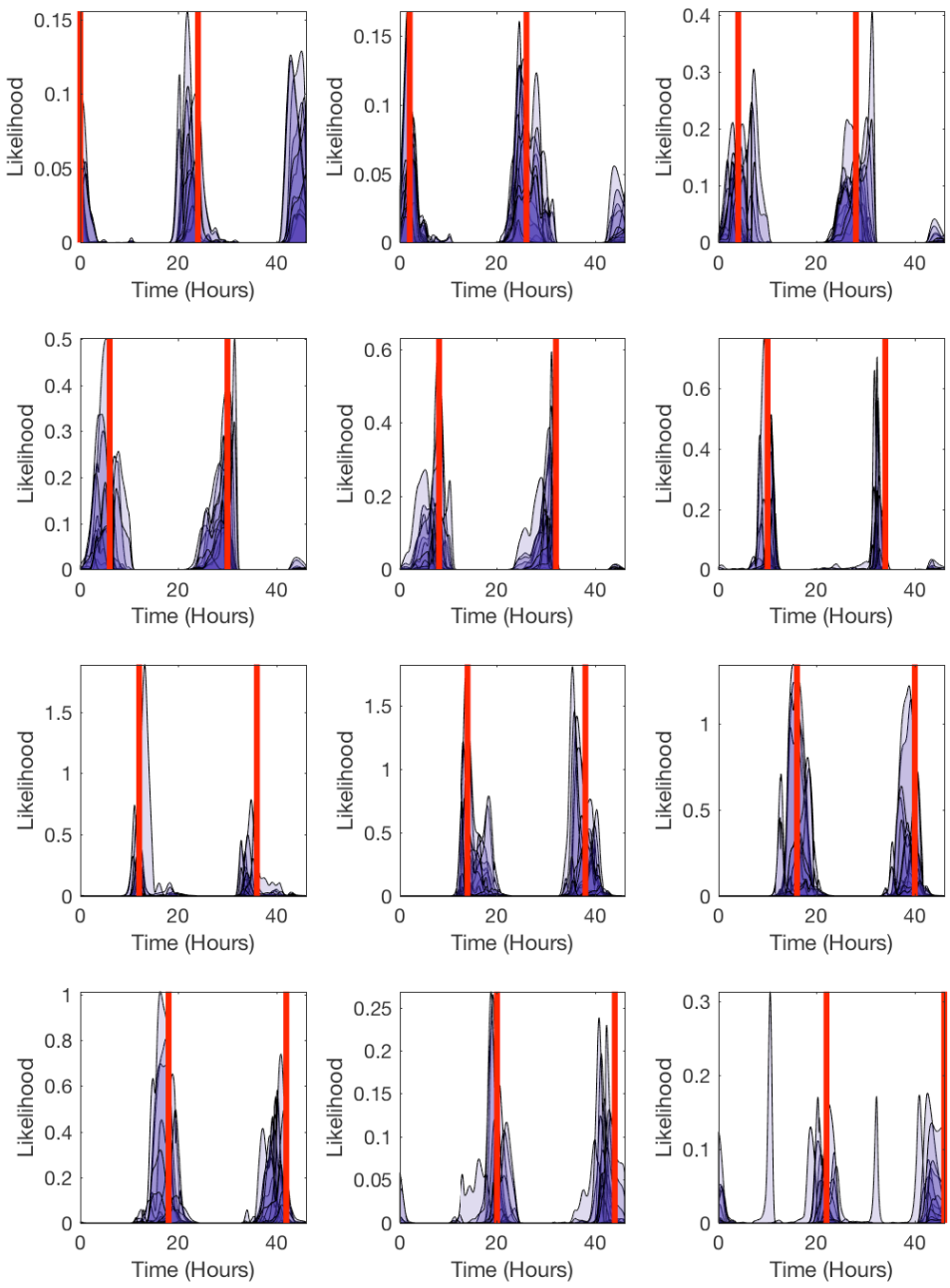


Figure 4.19: **Plot showing the results of all leave-out-organ-out, time estimation using the Zhang data.** The red markers represent the real time.

4.5 Human Time-Teller

So far in this thesis, Time-Teller has been shown to be able to tell times for *in silico* and mouse datasets. However, the data that make up these models has a high level of control and standardisation, with mice being kept in labs and being almost genetically identical. The generation of a human time-telling model is reliant on the finding in the previous chapter that the there is a set of circadian clock genes that were synchronised in expression profile and amplitude across individuals.

4.5.1 Training set and validation

The human data is at a 4 hour resolution over 24 hours, so only has 6 local principal component spaces. This human model uses 16 probes as the model features. The local PC spaces with fit splines are shown in figure 4.20.

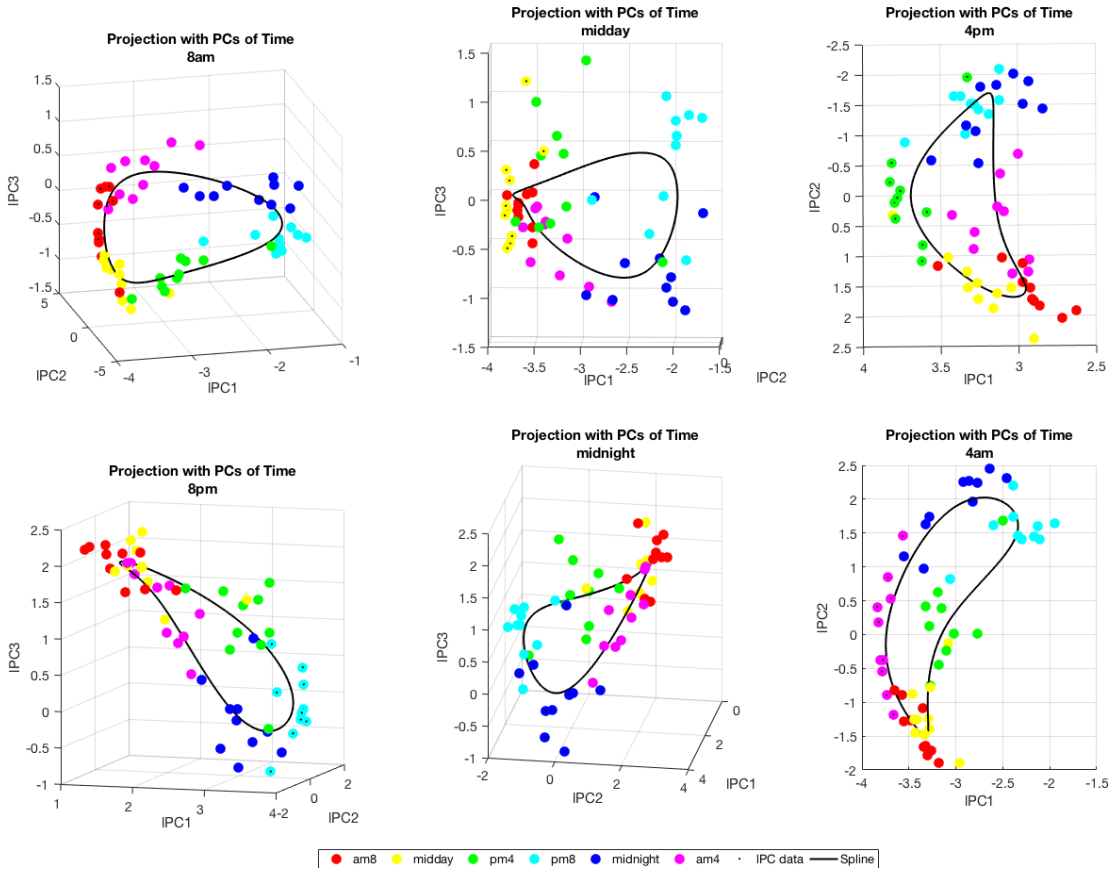


Figure 4.20: **All 6 local PC spaces for human data.** Data is projected into spaces, and coloured by time. Splines through the means of each set of 10 time data points show clear (distorted) elliptical shapes.

Despite the level of stochasticity expected when using transcriptomes of humans, the time organisation in local time principal component space is apparent. The differently

twisted shapes of these ellipses allows the local PC method to avoid some problems with symmetry.

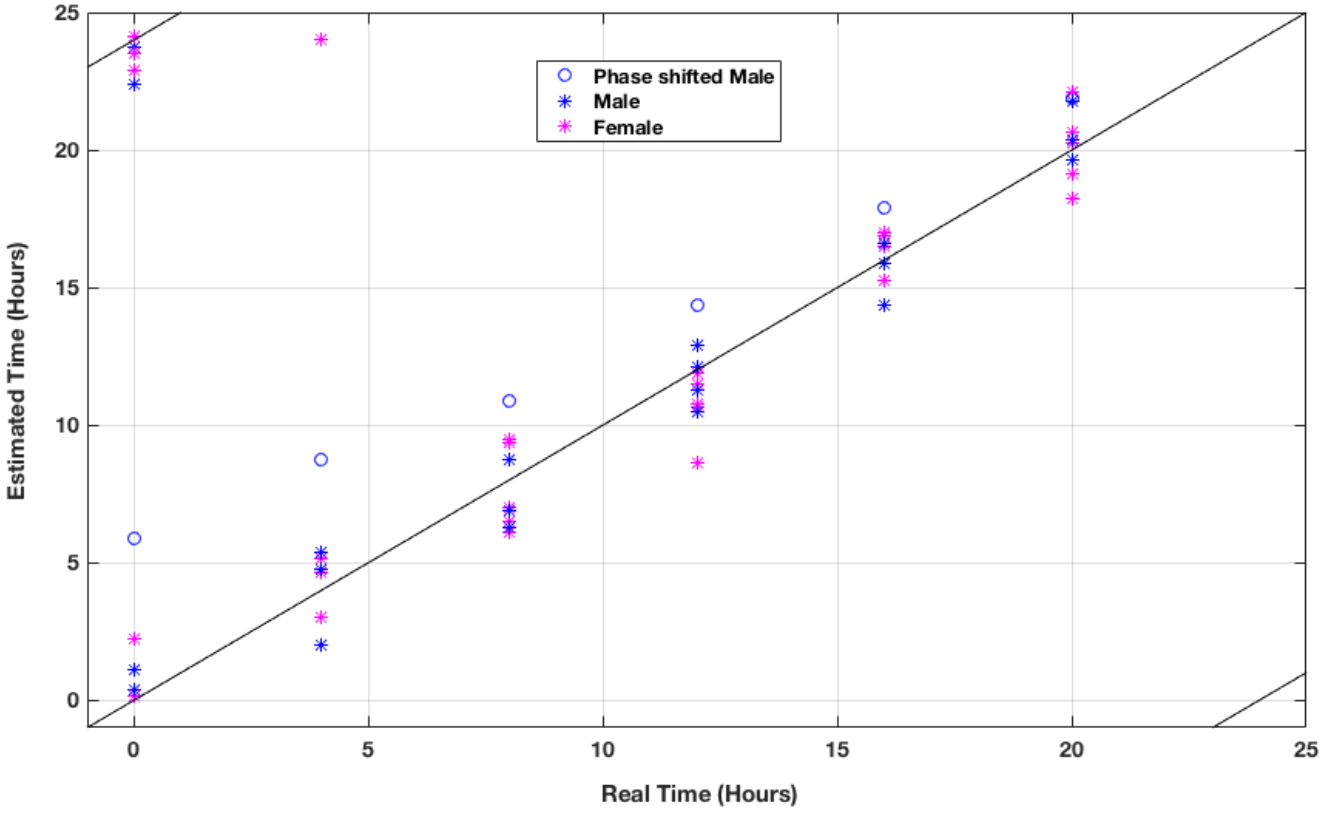


Figure 4.21: **Correlation plot for Time-Tellers predicted time versus actual time, for 10 human individuals, using a leave-out-out method.** One male has a circle marker to show that that individual is consistently phase shifted forwards 1-5 hours.

Figure 4.21 shows the results of a leave-one-individual-out approach to estimation. The method is very accurate, except for one male individual, whose time-estimations are consistently phase forwarded by 1-5 hours. Otherwise, the estimations are distributed about the real time. Table 4.4 shows this data and the distance from real time. The mean absolute error of estimation is about 1hr 20 minutes.

This mean absolute "error" is not true error as it includes the error introduced by the obvious natural phase shifts in Male15 and Female18. This "error" is due to both measurement error and natural inner body clock variation amongst individuals. This model says that when compared to the other 9 individuals, Male15 is more than 3 hours phase advanced. Similarly, Female18 is phase delayed by over 2 hours. We make the assumption in this thesis that a healthy population's body time is normally distributed about the population average, with small variance, and hence we estimate body time as real clock time.

We include Male15 and Female18 in the training data for the Human Time-Teller for

Individual	8am	Midday	4pm	8pm	Midnight	4am	Mean	Mean abs
M15	5.84	4.70	2.82	2.30	1.78	1.76	3.20	3.20
M12	-0.37	1.10	1.45	0.81	-0.83	-0.48	0.28	0.84
M11	1.12	1.35	0.70	-0.81	0.54	1.64	0.76	1.03
M8	-1.74	0.73	-1.78	0.06	-0.21	0.27	-0.45	0.80
M9	0.37	-2.01	-1.16	-1.55	-1.70	-0.48	-1.09	1.21
F18	-1.24	-1.02	-1.53	-3.42	-0.83	-1.84	-1.65	1.65
F13	0.00	-2.01	1.45	-1.43	0.91	-0.97	-0.34	1.13
F14	-0.62	19.88	1.33	-0.19	0.79	2.01	3.87	1.51
F5	2.24	1.10	-1.91	-1.31	0.41	0.52	0.18	1.25
F6	0.12	0.60	-1.04	-0.56	-0.83	0.15	-0.26	0.55
Mean	0.57	2.44	0.03	-0.61	0.00	0.26	0.45	
Mean abs	1.37	1.87	1.52	1.24	0.88	1.01		**1.32**

Table 4.4: **Table showing variation of Time-Teller's time estimates and real time.** All units are in hours. Male #15 (in green) is phase shifted forwards by around 3 hours, and female #18 (in red) is phase shifted backwards by 1-2 hours

future use, even though their body time is not the same as clock time. These phase shifts will be treated as a natural variation in body clock, and they help to define the shape of the Gaussian ellipsoids.

It is possible to adapt Time-Teller to estimate BT, and not clock time. BT could be determined by the expected phase of Bmal1, for example. Additionally, the 16 probes that have been used to create this model were chosen on the assumption that the most synchronised probes would inform the model the most. This is not necessarily true, and it could be possible to optimise this set to reduce the noise. This was explored within the Bjarnason data, and found to not add any value to the method. Too much optimisation within the training data resulted in over-fitting of the model, and resulted in worse results when applied to independent datasets.

4.5.2 Body time versus clock time

As internal body time is naturally variable about some population average, it would be clinically useful to be able to estimate the phase of a gene such as Bmal1 or Rev-Erbα (i.e. body time), instead of calculating the real clock time. It would be necessary in this case, to know the real time that a sample was taken, and use Time-Teller's output to calculate the shift from a defined population expectation as the input to a chronotherapy treatment regime.

COSINOR analysis provided estimates of the phases of each of the 16 probes, from the analysis in the previous chapter. Figure 4.22 shows the mean absolute "errors" in table 4.4 plotted against the estimated phase of each gene. There is a significant positive linear correlation for all probes, showing that Time-Teller is using the group average to

make predictions for clock time, which can be adjusted to the body time using the real time that the same was taken.

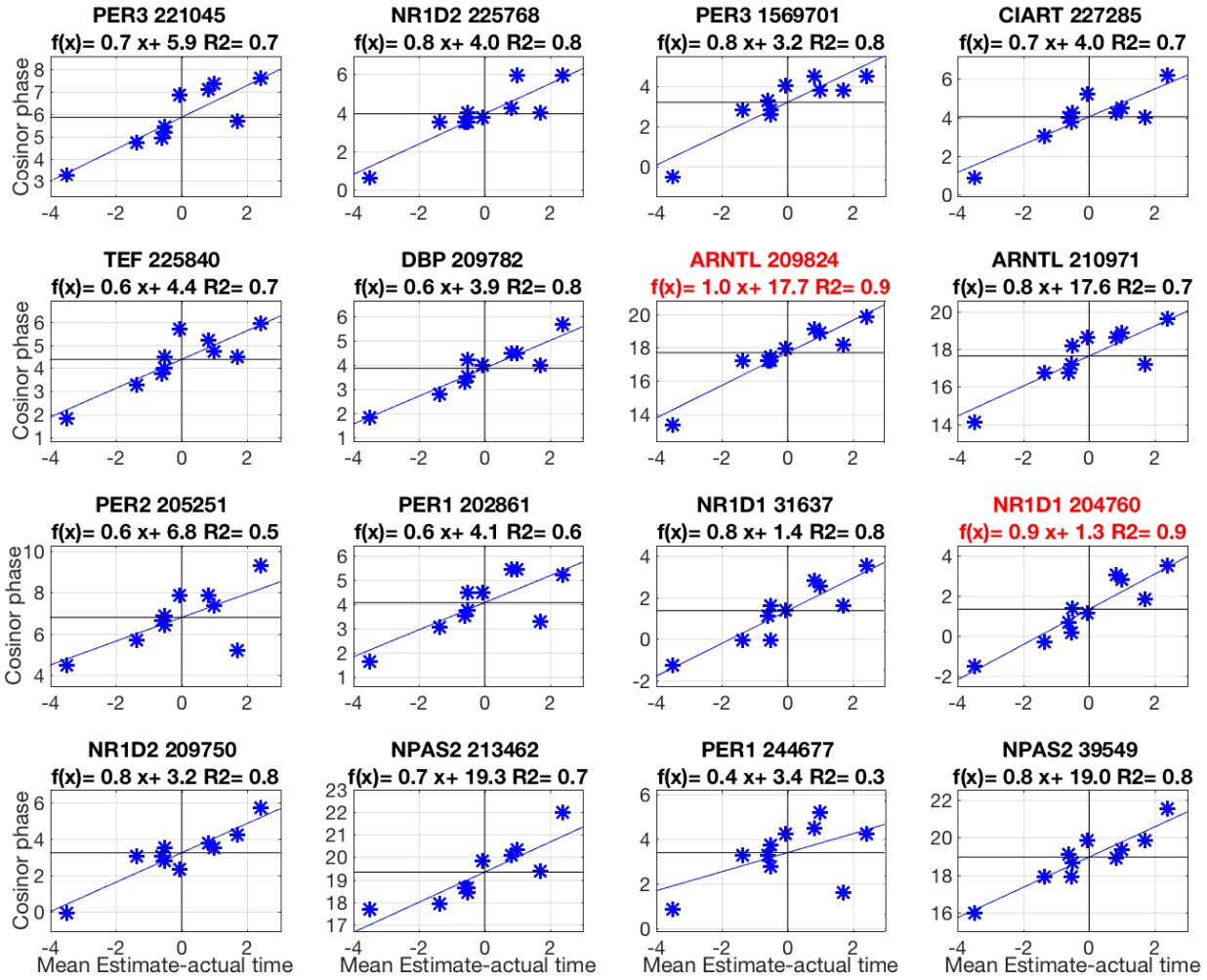


Figure 4.22: **Plot showing each clock genes phase plotted against the mean estimation "errors".** There is a strong positive correlation between estimated time difference from real clock time in all 16 probes. and Bmal1 and Rev-Erb$\alpha$are highlighted in red and show the highest correlations. This shows that Time-Teller can predict the phase of Bmal1 or Rev-Erbα with good accuracy if the real clock time of the sample is known.

This suggests Time-Teller can predict the phase of Bmal1 or Rev-Erbα with good accuracy and would be of use to the design chronopharmalogical optimisation therapies for drug administration.

4.6 Summary of chapter

This chapter summarised the main time telling algorithms that exist, and highlighted that none of them can truly tell the time from one single independent sample of unknown time. The Time-Teller algorithm was presented as a method that can predict the time of one independent transcriptome. All stages of data processing are free from batch biases, including the use of the fRMA preprocessing and only normalising with single sample mean and standard deviation (not timecourse mean and standard deviation as in Zeitzeiger). The advantage of using a local PC approach was shown using *in silico* data with low and high noise, showing that the local approach was always superior. A leave-one-organ-out approach to estimating the time of the Zhang data showed that Time-Teller is very accurate. It also showed that data samples between CT18-CT40 and CT42-CT64 are directly comparable.

Time-Teller can accurately calculate the time of human samples using a leave-one-individual-out approach. It was shown how the error in estimation of clock time is actually a good measurement for internal body time of individuals. This showed that Time-Teller can accurately calculate body time, if the real time that the sample was taken is known. We will discuss in the next chapter how we can measure if these individuals are simply phased forward or phase delayed, or if there is something else resulting in the measurements not showing the times we would expect.

The following chapter uses the full timecourse transcriptome datasets for training sets, and estimates the times of samples from independent studies. In order to do this, first we must be discussed how to measure the confidence of each time estimation.

Chapter 5

A Metric for Clock Dysfunction

This chapter concerns measuring confidence in the maximum likelihood estimations from the Time-Teller model. The assumption is made that if a sample's time cannot be estimated with a particular confidence threshold, then the circadian clock is dysfunctional in that sample. This assumption is reasonable due to the evidence that has been provided in this thesis for the robustness of the circadian clock, and for the consequential expectation of a specific behaviour. This assumption is only made when there is realistic expectation for an independent dataset to be comparable to the training data that determines this "specific behaviour" of the clock.

In the machine learning sense, we assume that if a test sample does not look significantly like the training set data, then something is wrong with that test sample, i.e. we assume that the training data represents the only "functioning clock" criterion, and anything that does not look like the training data we say has a "dysfunctional clock". Here we present a metric Θ that gives a value for the functioning of the circadian clock (which is equivalent to a confidence estimate in the MLE), where the threshold for which Θ that represents a functioning clock is provided by the training data.

The first section of this chapter will present the mathematical definition of the clock dysfunction metric, Θ. The *in silico* data is then used as an explicative example and as method validation. Independent datasets to the training timecourse mouse and human transcriptome are then tested with the Time-Teller model.

The metric Θ can be measured for just one independent sample, where the time that the sample was taken is unknown. This is not reported to be possible by any of the studies described in the previous chapter.

5.1 A metric for confidence in the MLE

The previous chapter showed that Time-Teller works well for predicting time of day from a sample in a leave-one-out sense. However, if Time-Teller's only output were the MLE, then an estimate will be made for *any* data, no matter how incomparable it is to the training data. Time-Teller is powerful as it can calculate the time of independent samples, and provide a measure of confidence for how good that estimate is.

The estimate for the time of a sample is given by the maximum point of the likelihood curve. When reporting a MLE, the shape of the likelihood curve is rarely used. We will now present a novel way of using the shape of the MLE to calculate the confidence of the estimate.

5.1.1 Shapes of likelihoods

The likelihood curves produced by Time-Teller are generated by probability densities of Gaussians that are distributed in a periodic shape in 3D space. These periodic 3D distributions were described as "distorted doughnuts" in previous chapters. Results that give easily interpretable likelihoods, as is sketched for the 2D case in figure 5.1, would arise from these ideal regular-shaped doughnut distributions. We will use this ideal scenario to begin to build our metric of clock dysfunction, and add levels of complexity as we progress.

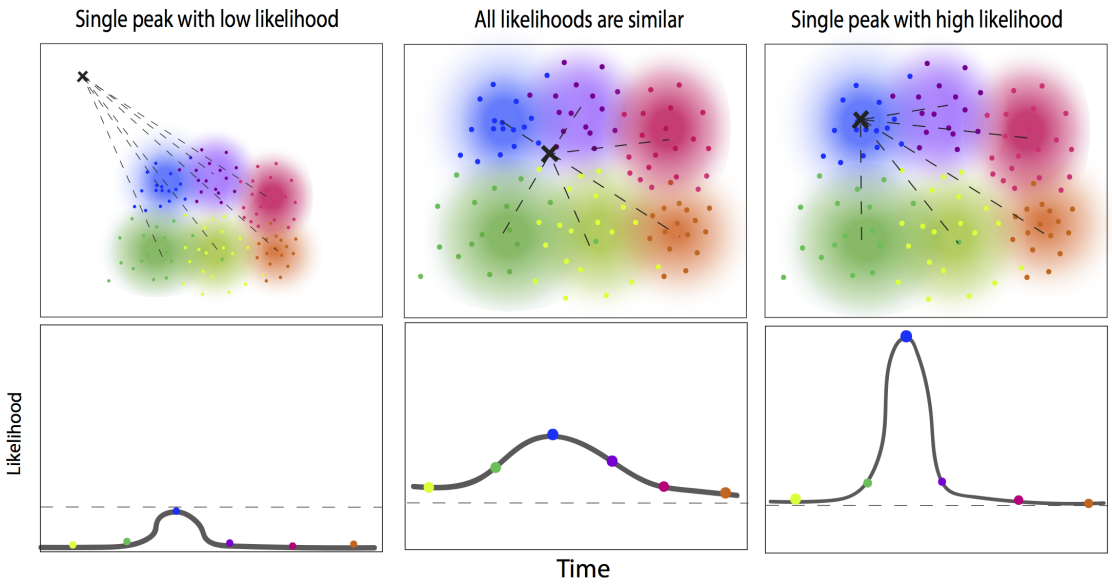


Figure 5.1: **Figure explaining likelihood shapes.** Sketches show likelihood shapes where a test projection is far from the training data space, in the middle of the training data space, and on the training space torus.

Figure 5.1 shows three scenarios that would lead to a MLE of the time represented in blue, from a 2D elliptical distribution. The left most sketch of figure 5.1 shows a test projection that is far away from the training data space, but the MLE is at the time that it is closest to, and all likelihood values are very small. Similarly, the middle sketch shows the projection could be near to the middle of the elliptical distribution, but slightly closer to the blue time Gaussian, and all likelihoods have similar magnitude. The right sketch shows that a MLE for a strong clock would have a high likelihood value, and other times would have a lower relative value.

Notice here that the value of the likelihood at the MLE would be a sufficient metric to measure the clock function in this case. However, as will be shown later in this section, the likelihoods we are actually working with are not as simple as this example, and the height of the likelihood is not a sufficient metric to represent the functioning of a working clock. Hence we continue with a different method.

Minimum threshold

To start building the clock function metric, first we recognise that the MLE in the first sketch, although being the maximum point of a single, clear, peak, should not be recognised as having a "good" clock because the maximum likelihood observed is low. If we set a minimum value for the likelihoods, m_t, then we can discount the false MLE that the first scenario would result in. This value is represented by the dashed line in the sketches. We set all values of any likelihood that are smaller than this threshold, to this threshold, i.e. any time at which test sample Z has likelihood $\mathcal{L}(Z|t) < m_t$, we fix the likelihood to the threshold for all the times at which this is true $\mathcal{L}(Z|t) = m_t$. The likelihood curve in the sketch on the left of figure 5.1, for which all values are less than this threshold, would now sit on the dashed line representing m_t, and the other two likelihoods would not change.

Now, we need a way to distinguish between the second and third scenarios. To do this, we use likelihood ratios to construct a metric. Likelihood ratios are generally used in hypothesis testing using the Neyman Pearson lemma.

Neyman-Pearson lemma

Suppose we have a probability model with PDF $f(x;\theta)$ and we wish to test $H_0 : \theta = \theta_0$ versus $H_1 : \theta = \theta_1$. If $L(\theta, x)$ is the likelihood function then the best test of size α has a critical region of the form

$$\Lambda(x) = \frac{L(x|\theta_1)}{L(x|\theta_0)} \geq A \tag{5.1}$$

for some non-negative constant A, where α is the probability of a false positive.

Use of likelihood ratios to define $\bar{\Theta}$

As Time-Teller is predicting the most likely time t_{pred}, one can view this as testing the hypothesis that the true time is $t \neq T$ i.e. by testing
$H_{t_{pred}}$: The true time of the sample is t_{pred}
H_t : The true time of the sample is t
According to the Neyman Pearson lemma, the best way to do this is by considering the likelihood ratio

$$\Lambda(x,t) = \frac{\mathcal{L}(Z|t)}{\mathcal{L}(Z|t_{pred})} \tag{5.2}$$

Let $\bar{\Theta}$ be the proportion of times t such that

$$\Lambda(x,t) \geq \eta \tag{5.3}$$

so that $\bar{\Theta}$ represents the proportion of the time space t for which $\Lambda(t,x)$ is greater than threshold η. The Λ likelihood ratio curves are shown in the sketch in figure 5.2 for the same likelihoods as in figure 5.1. A test projection to far outside of the elliptical space would result in $\bar{\Theta} \approx 1$, a projection near to the centre of the torus, would result in $\bar{\Theta} \approx 0.5$ (in that magnitude), and a projection in which we are highly confident in, would result in a small $\bar{\Theta} \approx 0.05$.

Periodically penalised ratio threshold

Now, the final likelihoods used in the Time-Teller model are from 3D distributions. It is possible that the distributions along the spline in each time PC space are slightly folded (see figure 4.20) and could result in likelihood curves with double peaks. Additionally, as we are using local principal components and combining likelihoods, it is possible that the final likelihood has multiple peaks.

The $\bar{\Theta}$ metric described does not deal adequately with multiple peaks. If the likelihood had two peaks with another high peak roughly 12 hours away from the MLE, we would want the $\bar{\Theta}$ metric to penalise this, but if the two peaks are close then we would not want this because that is compatible with good clock function. As it stands, $\bar{\Theta}$ would not distinguish between these two cases. Some examples of likelihoods and how we would want them to be classified are shown in figure 4.20. To summarise our desires for the $\bar{\Theta}$ metric:

- a single, narrow, high, peak would result in a high confidence estimate;

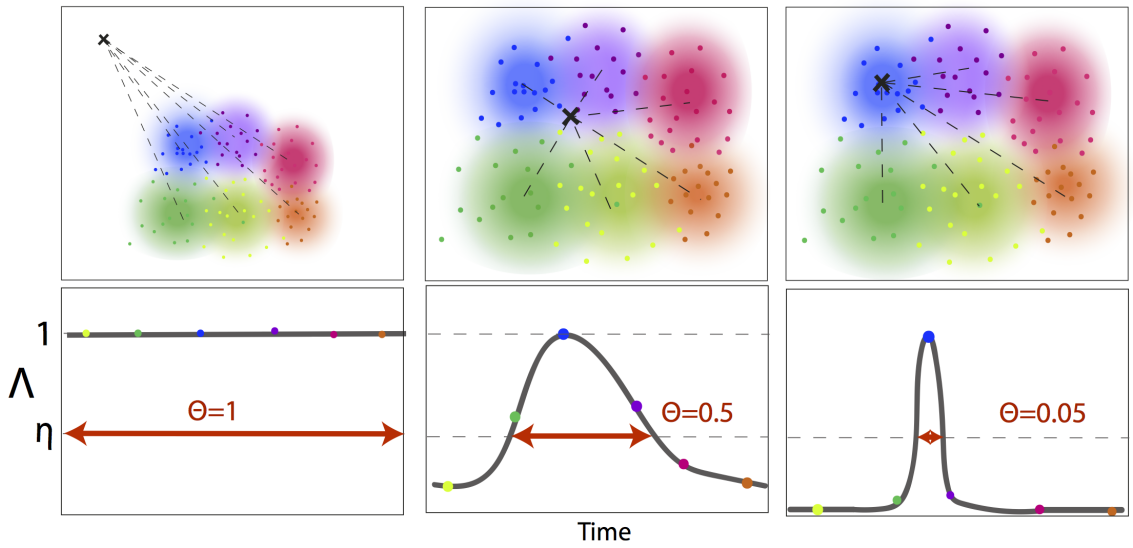


Figure 5.2: **Figure explaining likelihood ratio shapes, and the $\bar{\Theta}$ metric.**Sketches show likelihood ratio shapes where a test projection is far from the training data space, in the middle of the training data space, and on the training space torus.

- a wide peak would result in a comparatively lower confidence estimate;
- a high second peak near to the MLE would result in good confidence in the MLE;
- a low second peak around 12 hours would result in good confidence in the MLE;
- a high second peak around 12 hours would result in low confidence in the MLE; and
- a uniform likelihood function, or one with low peak(s), would provide a low confidence estimate.

To do this we implement a penalised likelihood ratio, where we use a cosine function to scale the impact that a secondary (or more) peak will have on the final metric based on distance to the MLE.

Let

$$C(t|T,\epsilon) = 1 + \epsilon + cos\left(\frac{t-T}{24}2\pi\right) \tag{5.4}$$

be a simple cosine curve transformed so that $\epsilon \leq C \leq 2+\epsilon$, where $C(T|T,\epsilon) = \epsilon$ and $C(T+12|T,\epsilon)) = 2+\epsilon$. We define $\epsilon > 0$ so that $C > 0$. The larger ϵ is, the less anti-phase peaks impact the final confidence metric.

We use this cosine transformation to scale the threshold η in (5.3) so that the final clock function metric Θ is defined by the proportion of time t which satisfies:

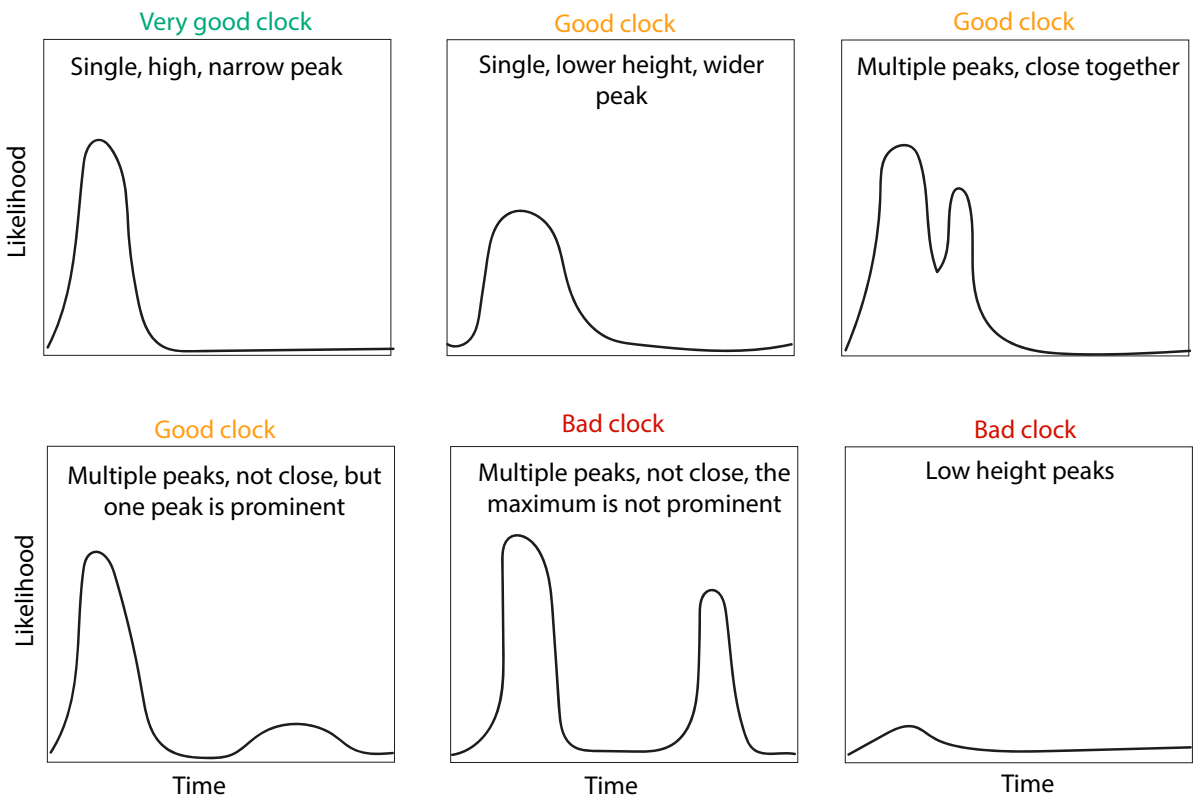


Figure 5.3: **Sketch showing some possible likelihood shapes, and how they should be classified.**

$$\Lambda(x,t) = \frac{\mathcal{L}(Z|t)}{\mathcal{L}(Z|t_{pred}))} \geq \eta C(t|T,\epsilon) \tag{5.5}$$

This results in a changing threshold for the likelihood ratio Λ, where, for example, the MLE must be $(2+\epsilon)/\epsilon$ times higher than a second peak 12 hours away, in order so that the other peak is not taken into account in the dysfunction metric.

A sketch of this is shown in figure 5.4. The left panel show single peak in a likelihood curve would not be affected by the penalised cut-off. The middle peak shows how if there is a secondary peak near to the MLE peak, then Θ will not be penalised. The right panel shows how if there is a secondary peak far from the MLE, then it will penalise Θ.

Choosing parameters

In order to generate Θ metrics using Time-Teller, the values of ϵ and η must first be appropriately chosen. First, η and ϵ must have some constraints so that we can ensure $\Theta(x) > 0 \ \forall \ x$ (if Θ were allowed to be 0, then we risk losing all information in the

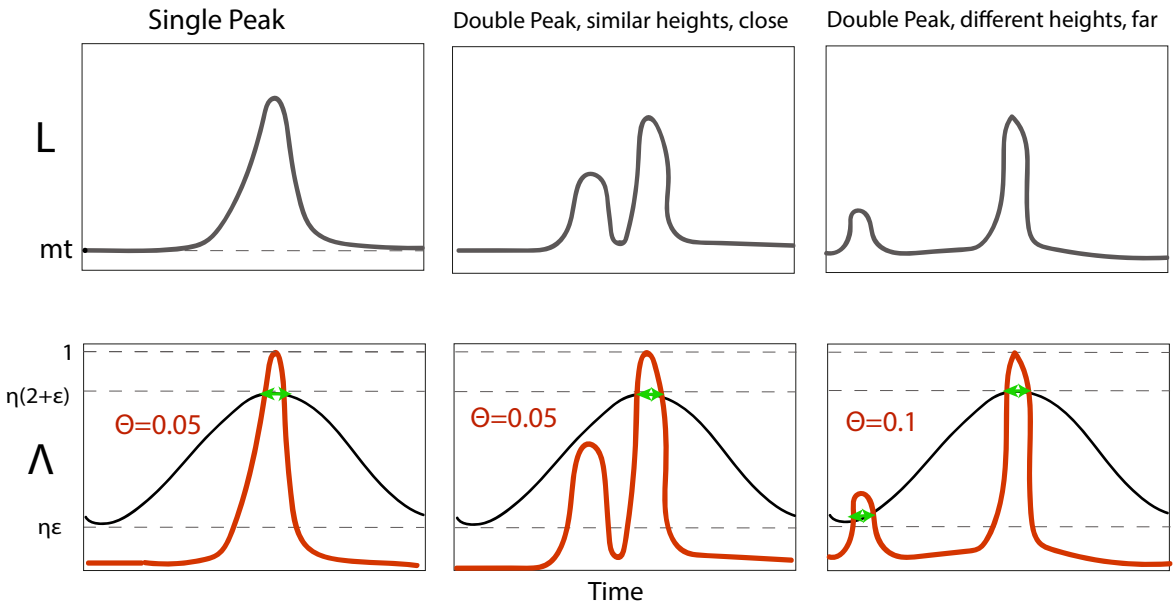


Figure 5.4: **Figure explaining the use of the penalised Θ.**Sketch of likelihoods (top), and equivalent likelihood ratios, Λ (bottom). Left panel shows how a single peak will not be affected by the changing threshold. Middle panel shows how a secondary peak, near to the MLE of comparable height, will not penalise Θ. Right panel shows how a small peak anti phase to the MLE will penalise Θ.

likelihood) hence the first condition is that

$$0 < \Lambda(x,t) \leq \Lambda(x,T) = 1 \; \forall \; x \tag{5.6}$$

From (5.5), it is clear to ensure the positivity of Θ we must enforce that

$$\eta C(t|T,\epsilon) < 1 \tag{5.7}$$

As $\max(C) = 2+\epsilon$, and $\min(C) = \epsilon$

$$0 < \eta\epsilon < \eta(2+\epsilon) < 1 \tag{5.8}$$

and the final constraints on ϵ and η are

$$0 < \eta < 1/(2+\epsilon) < 0.5 \tag{5.9}$$

$\eta\epsilon$ defines the minimum value of the threshold for which the likelihood ratio Λ can be intersected, and $\eta(2+\epsilon)$ defines the maximum. We set $\epsilon = 0.4$ and $\eta = 0.35$, which sets the maximum threshold (at the MLE) to $\eta(2+\epsilon) = 0.84$, and the minimum threshold (anti-phase to the MLE) to $\eta\epsilon = 0.14$. This means that if there is a peak in the likelihood

exactly anti-phase to the MLE that is more than 14% of the maximum, then it penalises Θ.

ϵ, η, and the minimum threshold m_t are hyperparameters of the Time-Teller model. $\epsilon = 0.4$, $\eta = 0.35$, and $mt = 0.0067$ are used in this thesis, but future users may want to change the parameters. Any parameters that satisfy the conditions in (5.9) will (in theory) provide a valid result, and an example of this is shown in the next section. As $\epsilon \to \infty$ the method is the same as a non penalised approach (so same as $\bar{\Theta}$), and as $\epsilon \to 0$ any anti-phase secondary peak above m_t will penalise Θ. An example of this is shown in the next section.

m_t is a measure of the signal-to-noise threshold. It means that the value of the likelihood curve at at the MLE must be must be far greater than this limit for it to be significant, i.e. $\mathcal{L}(Z|t_{pred}) >> m_t$. $m_t = e^{-5}$ was chosen somewhat manually, by observation of the log-likelihood curves. Optimisation of this parameter could be possible using the distances of the opposite sides of the "doughnut" distributions, but for now this remains fixed.

5.1.2 Exploring clock function using simulated data

As with the previous chapter, we use the stochastic Relogio model to generate dummy data to use for method validation and explanation. The strength of using data from this model is that we can create data with a "dysfunctional clock" by doing an *in silico* knock-down.

Bmal1 is widely reported to be crucial for the proper functioning of the circadian clock [144]. To knock down Bmal1 in the Relogio model, V_{max5}, the rate of Bmal1 transcription, is set to 10% of its original value (see appendix B). A trajectory of the knock-down (KD) over 2 periods is shown in figure 5.5, and is clearly disrupted when compared to the nicely rhythmic wild-type (WT) trajectories. These simulations have a system size $\Omega = 500$, used as a middle-ground noise level. The first few hours appear comparable between WT and KD. 10 equally spaced samples from the KD data set were saved (as vectors of 12 "gene expression" values), and the Time-Teller model was used to estimate the time of the samples as outlined in the previous chapter. The WT data was used as a training set in the same way as has been used previously.

Examples of projections into local principal component spaces are shown in figure 5.6. Where the training data forms the familiar elliptical space, the WT test data (black line) occupies a similar space, and the KD data (dashed line) is a distance away from the training data. As the KD simulation starts with the same initial conditions to the WT data, the trajectory starts off in the WT space.

The likelihoods resulting from these projections are shown in figure 5.7. The WT

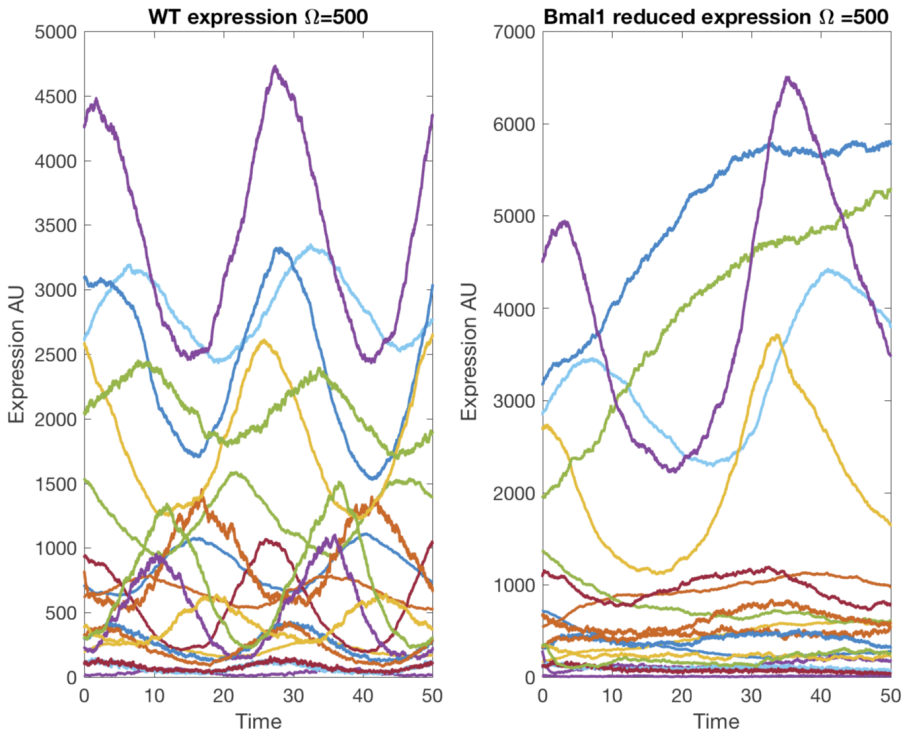


Figure 5.5: **Plots of WT and KD stochastic Relogio models.** For $\Omega = 500$ the trajectories of a simulation over 2 periods of the stochastic Relogio model with $V_{max5} = 1$ for WT(left), and $V_{max5} = 0.1$ for *in silico* KO (right). The colours on the plot represent the 19 variables of the Relogio model, and the amplitudes are arbitrary.

(black) data shows good shape likelihoods over the whole 2 days of data, and the KD (red) data shows a good likelihood for the first test time-point as expected, but there is no visible peak thereafter.

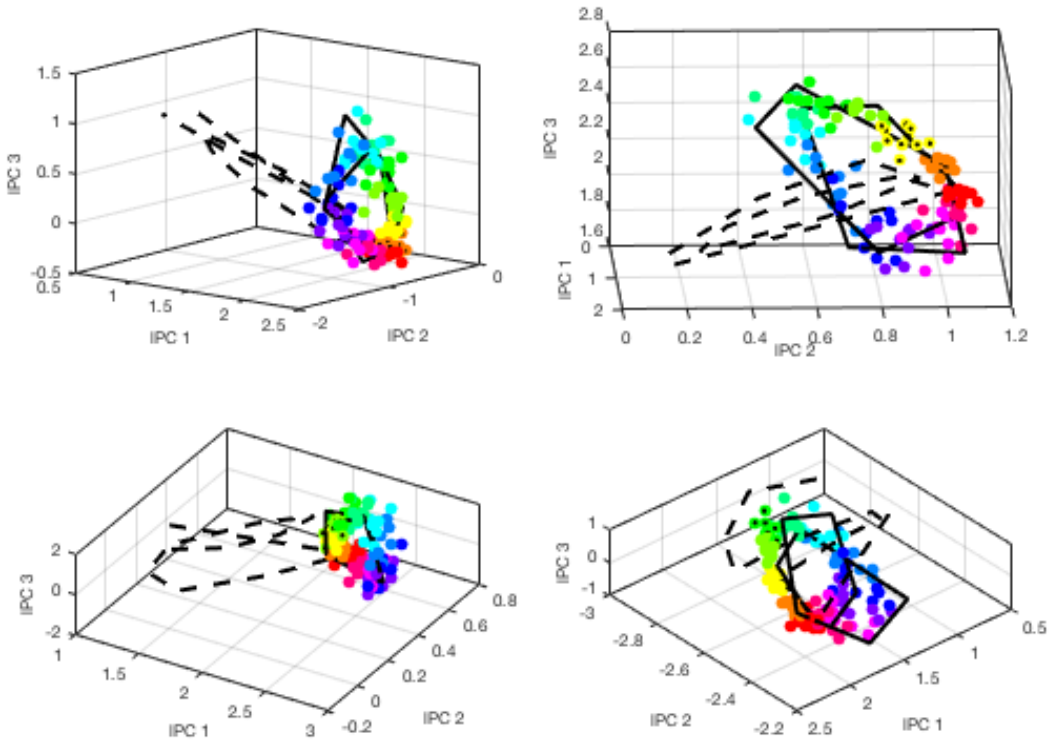


Figure 5.6: **Examples of 4 local principal component spaces, with KD projections existing outside of the doughnut distribution representing WT data.** The coloured training data shows the usual elliptical distribution shapes. Some WT test data is projected into the space, these points are shown joined by a solid, black line. The KD data (joined points with a black, dashed line) is a distance away from the training data in PC space.

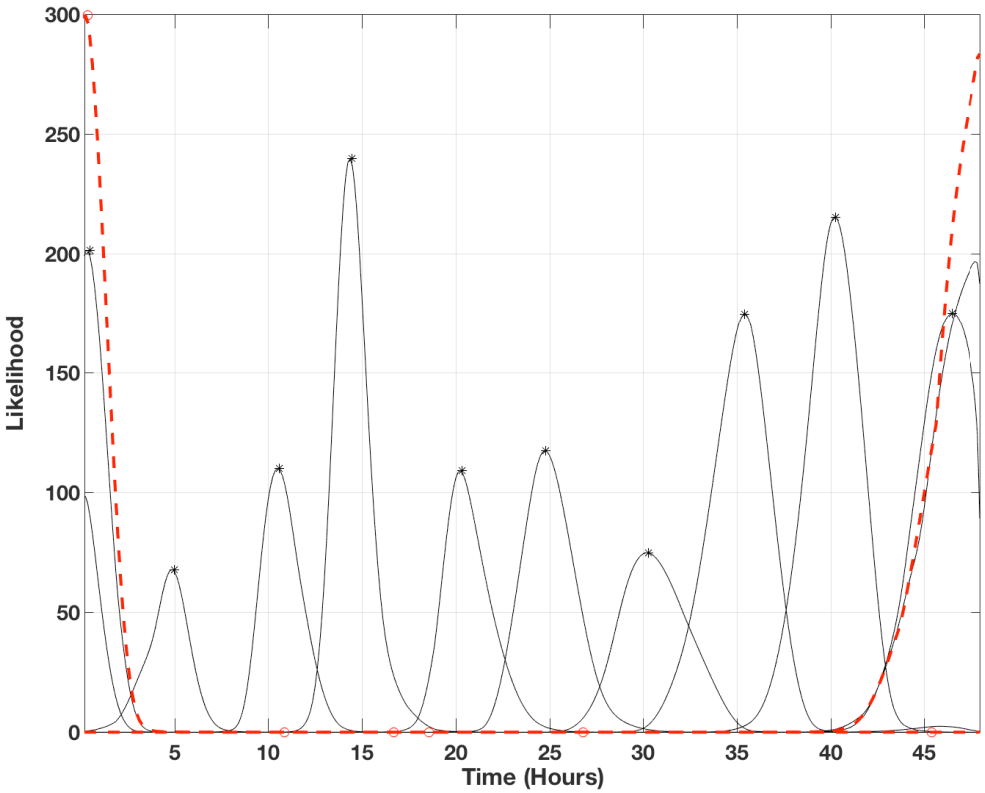


Figure 5.7: **Plot showing likelihoods for one timecourse of WT data and one of KD data**. The WT (black) data shows good shape likelihoods over the whole 2 days of timecourse, and the KD (red) data shows a good likelihood for the first test time-point as expected. There are no other peaks for the KD data as they are so flat, they are not visible in the figure. The black asterisks shows the MLE for each WT likelihood, and red circles show MLEs for the KO likelihoods.

5.1.3 Measuring Θ

Using the simulated data from the stochastic Relogio model, we attempt to show how Θ increases as the magnitude of the knock-down increases. Again using $\Omega = 500$ for 10 training simulations over 52 hours, we test how Θ changes as the magnitude of the level of knock out V_{max5} is decreased from the WT value of 1 to the KD value of 0.1, in steps of 0.1. Each simulation ran for 52 hours, where 10 equally spaced time-points were extracted as the test sets.

The first time-point of each test set is the same point from the limit cycle (with added noise) so are expected to have good Θ values even for the KD data[1]. The 10 data point timecourse of the Bmal1 variable as the V_{max5} parameter is shown in figure 5.8, where the change in circadian behaviour is clear. Time-Teller attempted to tell the time of each of the 10 data points for each trajectory. The Θ values for each of the 10 data-points of 10 different KD intensities are summarised in the box plots in figure 5.9.

Figure 5.10 shows example likelihoods for WT, and 3 levels of *in silico* knock-downs. The WT shows a single high peak in the likelihood. $V_{max5} = 0.6$ shows a clear peak, but with very small height compared to the 0.0066 threshold, and $V_{max5} = 0.4$ shows 2 peaks that are even smaller compared to the baseline. $V_{max5} = 0.3$ shows no likelihood above the baseline. Figure 5.11 show the equivalent figures for the likelihood ratios in red, with the cosine thresholds overlaid. Values for Θ are shown on the figure, where the WT likelihood is 0.025, and $V_{max5} = 0.6, 0.4, 0.3$ are $\Theta = 0.182, 0.727, 1$ respectively.

[1] As two periods are being used in the training data, the likelihood curves for each period are now averaged (geometric mean)

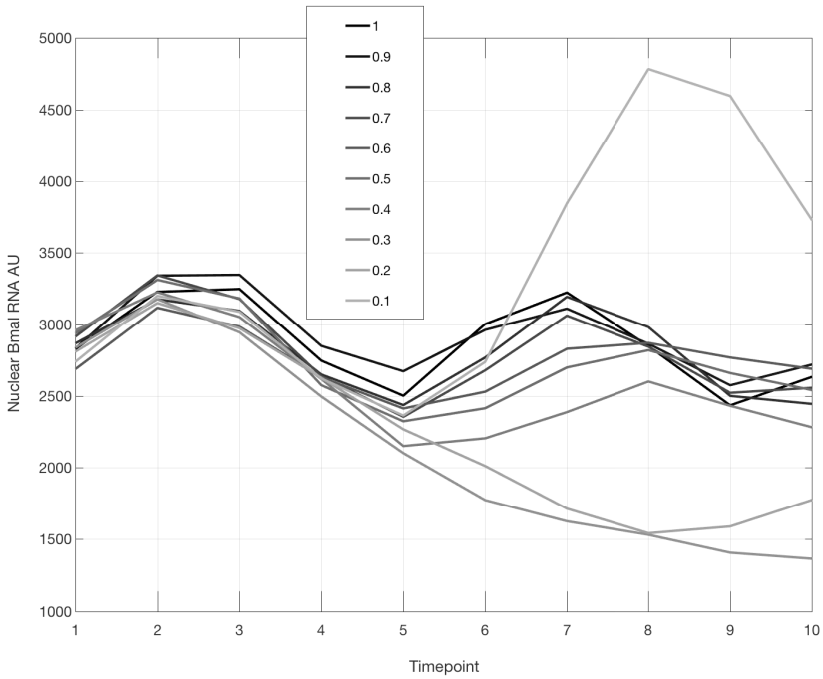


Figure 5.8: **Plot showing the change in trajectory of the Bmal variable of one simulation each of the Relogio model as the intensity of knock-down is increased.** V_{max5} decreases from 1 (black) to 0.1 (lightest grey), where rhythms are lost.

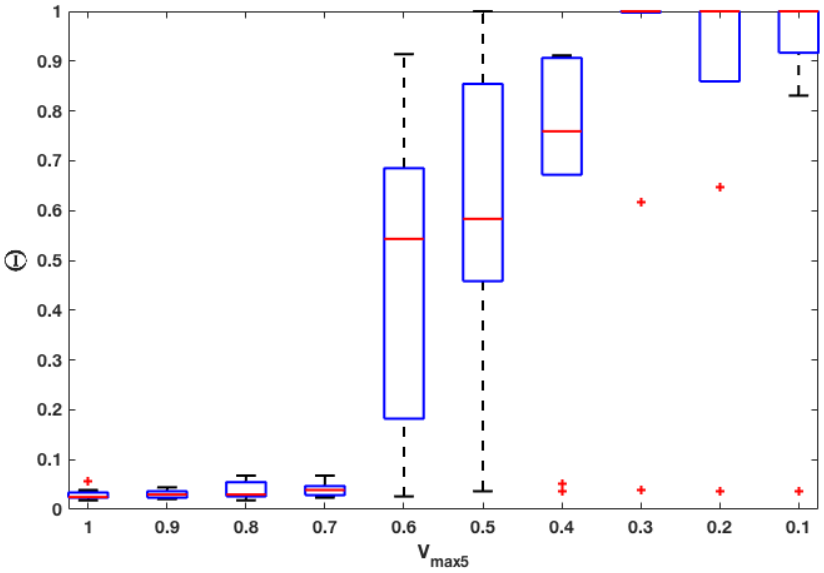


Figure 5.9: **Box plots showing the change in dysfunction measure Θ as V_{max5} is decreased by 0.1 from 1 to 0.1.** V_{max5} clearly has a large effect on the dynamics of the system as it falls below 0.7. Outliers in red around $\Theta = 0.05$ are due to each simulation starting with the same initial conditions, and acts as an *in silico* control.

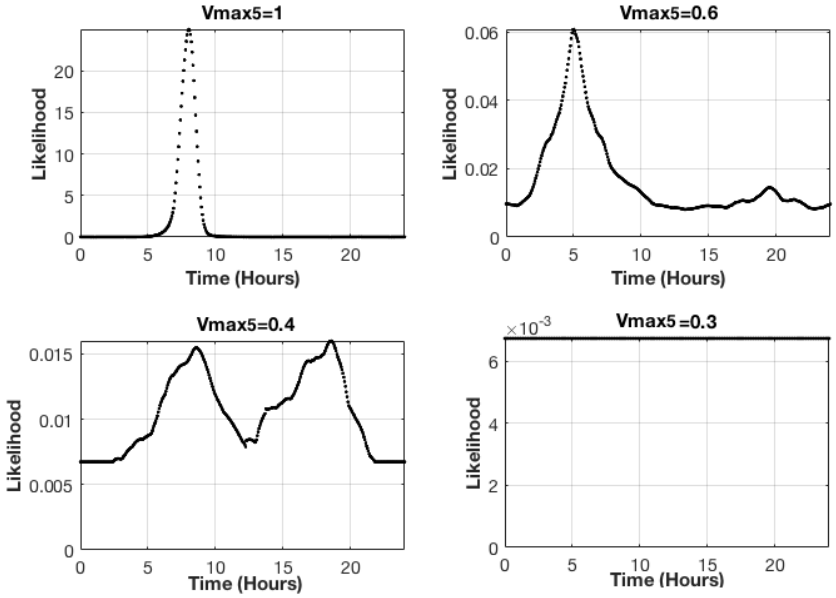


Figure 5.10: **Plots showing the likelihoods for 4 magnitudes of knock-down in the stochastic Relogio simulations.**

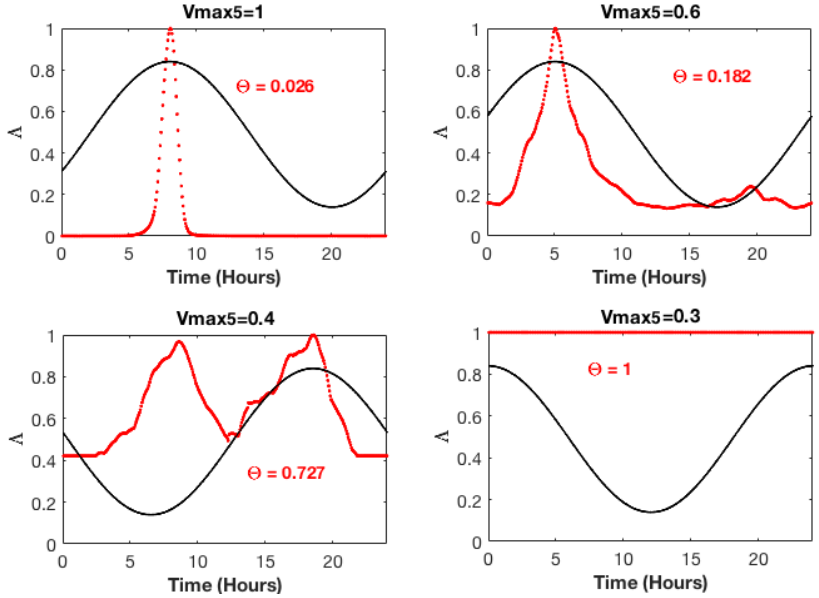


Figure 5.11: **Plots showing the likelihood ratios for 4 magnitudes of knock-down in the stochastic Relogio simulations.** Values of Θ are shown, which represent the proportion of time that the likelihood ratio is above the cosine (in black).

Notes on the values of η and ϵ

The box plots in figure 5.9 showed that as V_{max5} falls below 0.7, the functioning of the clock is severely affected. There is an (almost) monotonic relationship between Θ and the change in the V_{max5} parameter. This plot was generated with parameter values $\epsilon = 0.4$, $\eta = 0.35$, and $mt = 0.0067$. To show the effects of changing ϵ and η, 6 different parameters (still satisfying conditions (5.9)) were used to generate equivalent figures. These box plots are shown in figure 5.12, and all figures show that the Θ metric indicates a functioning clock for V_{max5} until it reaches 60-70% of its original value.

All boxplots show similar behaviour, where the intensity of the KD is correlated with Θ. However, there are some differences. Setting $\eta = \epsilon = 0.01$, the method is very sensitive. When $\eta = 0.2$ and $\epsilon = 0.1$ the results show a more "all or nothing" result. $\eta = 0.33, \epsilon = 0.9$ and $\eta = 0.49, \epsilon = 0.01$ show many Θ values that are not 1 for the extreme knock downs, so are not sensitive enough. $\eta = 0.05, \epsilon = 10$ represents a parameter set with weak anti-phase peak penalising, but will penalise heavily for wide peaks. $\eta = 0.3, \epsilon = 0.2$ shows results comparable to 5.9, as the parameters are of similar ranges. This shows that the method is not reliant on the specific choices of η and ϵ.

This *in silico* analysis should provide an intuitive explanation as to what the clock dysfunction metric Θ represents. These parameter choices for η and ϵ produce appropriate results for the purposes of this thesis. It should be noted that we do not treat Θ with the normal 0.05 significance cut-offs that one might be inclined to do. Θ is a metric, and not a test statistic. "Functional clock" or "dysfunctioning clock" ranges for Θ are defined by each training set. This is due to different levels of expected variation. Θ will be referred to as the clock function/dysfunction metric.

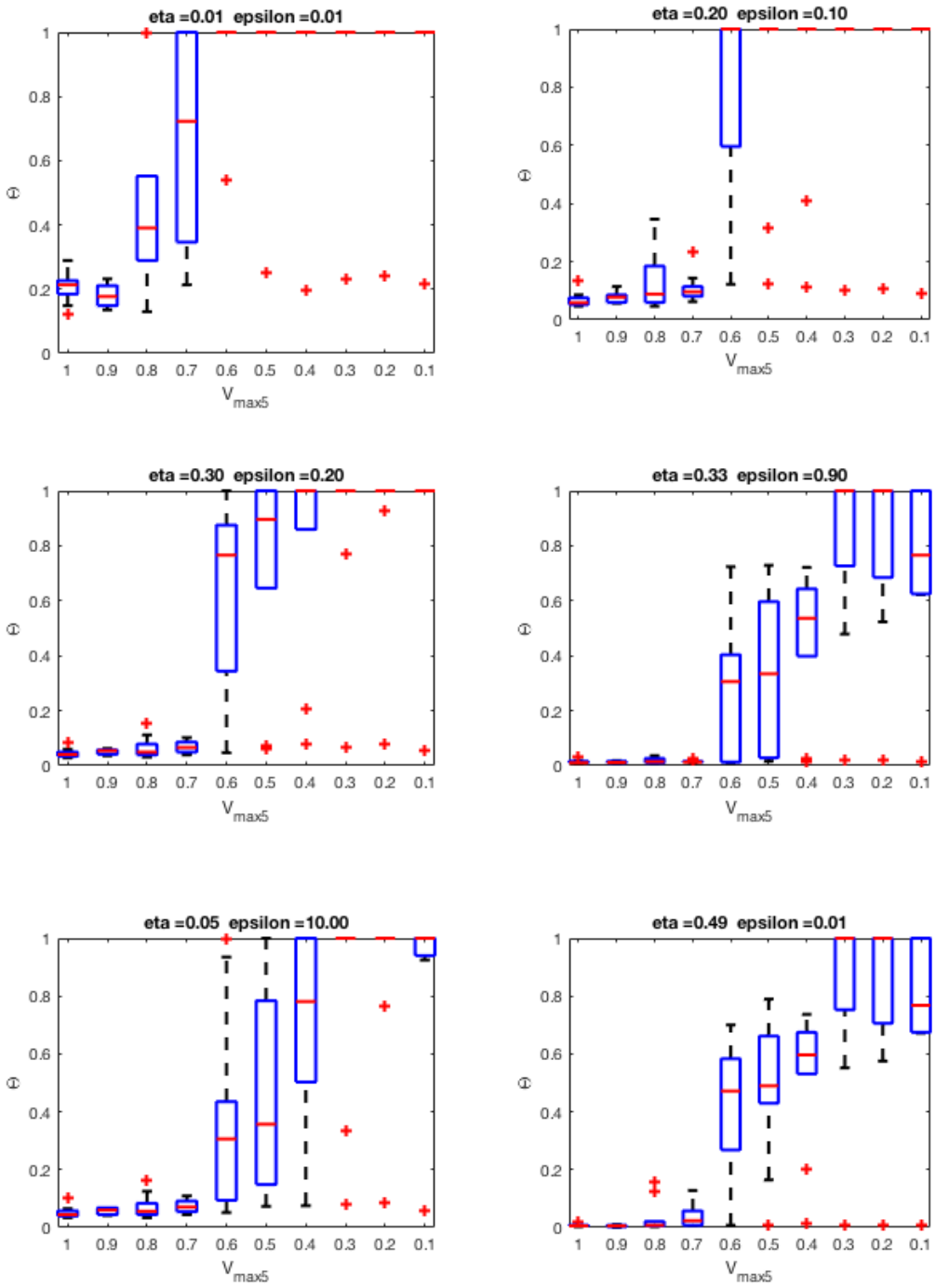


Figure 5.12: **Examples of Θ distributions for different levels of knock downs, for different ϵ and η parameter sets.**

5.2 Mouse Time-Teller applied to independent datasets

In this section, Time-Teller is applied to independent mouse data sets. All data sets were found by a literature search, for data labelled with time, using the same male mouse C57BL/6, and with microarrays performed using the Affymetrix MoGene 1.0 ST GeneChip.

5.2.1 Distribution of Θ for Zhang timecourse data

To investigate the Θ clock function metric for the training data, we do not use a leave-one-out approach. The test samples are contained in the training set, as to make it possible to compare to other data sets. This is a control and validation step, and does not impair the power of the test. The 10 genes Arntl, Npas2, Dbp, Per3, Nr1d1, Per2, Nr1d2, Tef, Ciart, Wee1, and Clock (with Ciart having 2 probes) are used. The range of Θ values resulting from applying the Time-Teller model to the training samples are shown in the histogram in figure 5.13. This range helps us to define what Θ values represent a functioning circadian clock. It is quite clear here that majority of the data has $\Theta < 0.1$, but there is a tail on the distribution until $\Theta < 0.2$. We use this distribution to simply define that mouse samples with $\Theta < 0.2$ have functioning clocks, and we are confident in these sample's estimates.

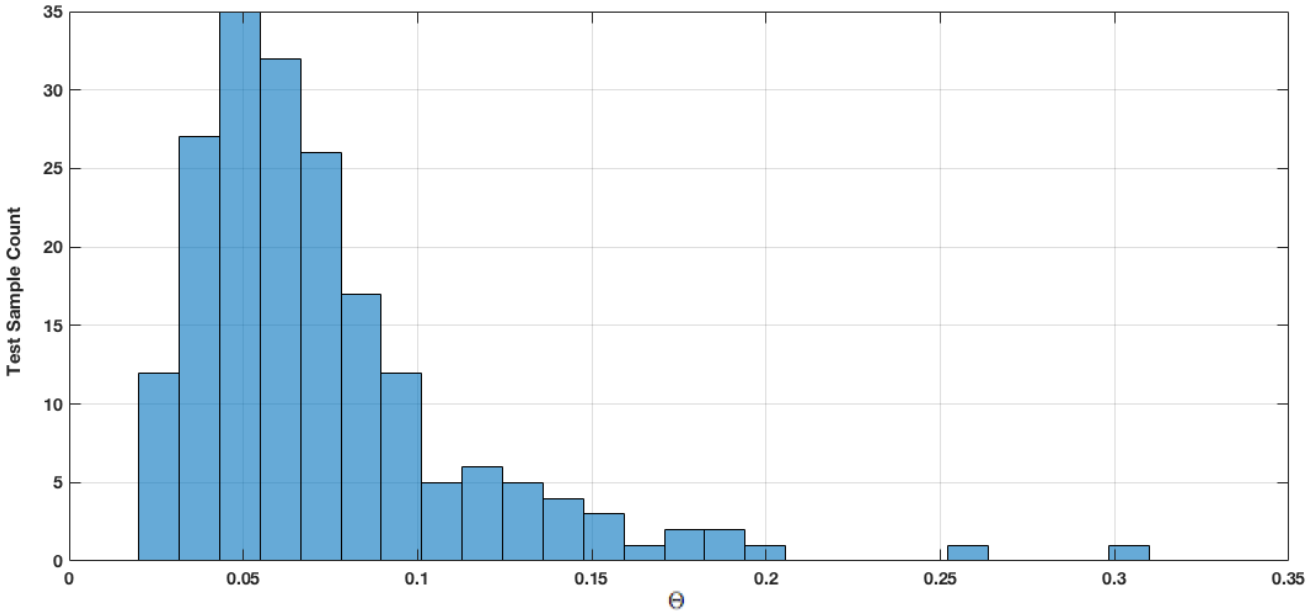


Figure 5.13: **Histogram showing the distribution of Θ values for the Zhang data.** For all 182 samples of the Zhang data, the distribution of Θ values has a median around 0.05, with a tail to around 0.2.

5.2.2 LeMartletot data: simple timecourse

The simple timecourse dataset created by LeMartelot *et al.* [107] is an ideal dataset to test the power of the mouse Time-Teller on an independent dataset. It is made up of microarrays of pooled RNA extracted from the whole liver of 5 mice. Samples were collected at times ZT2, ZT6, ZT10, ZT14, ZT18, ZT22, ZT2(+24), after the mice had been entrained to LD cycles for 2 weeks.

Although there are only 7 data points in this dataset, it provides crucial validation for the Time-Teller model. Time-Teller's predicted results for the LeMartelot timecourse are summarised in figure 5.14, showing Time-Teller's predictions versus the real times of the samples. The mean absolute error for time estimation is less than one hour. The Θ values range between 0.02-0.11 with one $\Theta = 0.17$, so all values fall within the "functional clock" criteria ($\Theta < 0.2$) defined in the previous section. The data for each time point was treated independently, and any of these estimations could be made in one analysis with completely reproducible results. These clearly accurate estimations, with Θ values that indicate a functioning clock, show that the Time-Teller model works with no need for data manipulation or batch bias. It is also notable to recognise that the results appear distributed around the $x = y$ expected result line, indicating no overall phase shift bias.

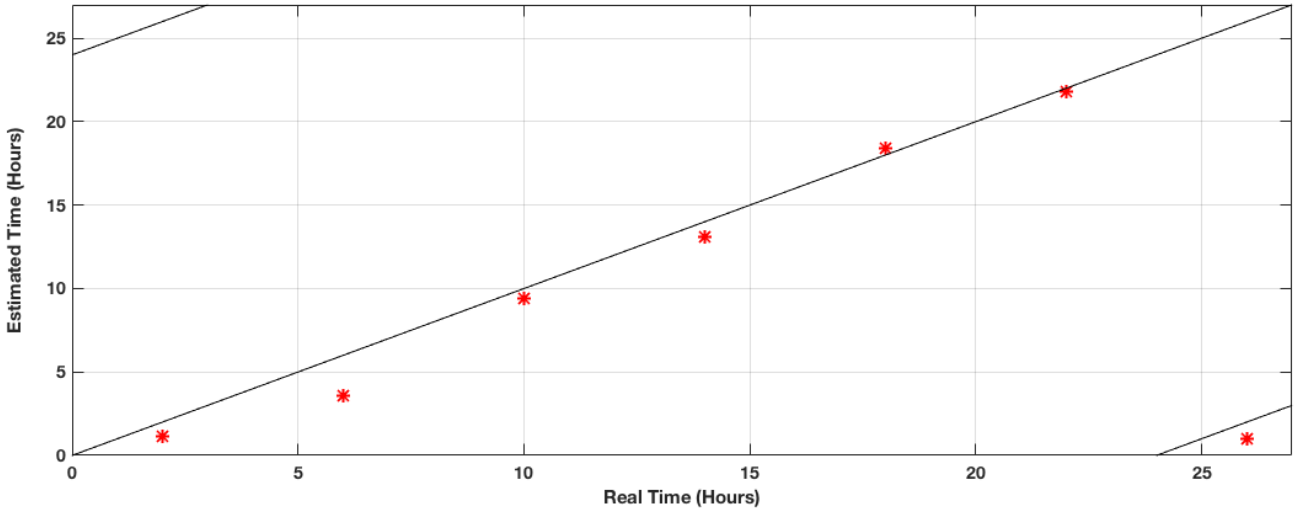


Figure 5.14: **Scatter Plot showing the real time versus estimated time for the LeMartelot liver timecourse data.** The estimations are highly accurate.

5.2.3 Fang data: WT and Rev-Erbα KO

A dataset comparable to the *in silico* WT versus KO data from the previous section was published by Fang *et al.* [105]. This dataset allows the validation of Time-Teller by testing its estimations on samples where we can assume that there is no functioning clock, versus samples from the same experiment where we do expect a functioning clock.

In Fang *et al.*'s [105] experiment, 5 Wild Type and 5 Rev-Erbα KO mice were entrained to LD cycles and euthanised at ZT10, where liver samples were taken. As the training data is labelled between CT18 and C64, this corresponds to CT34 in the Time-Teller model. The publication reports that the knock-out was successful.

The time of these single samples was estimated using the Time-Teller model. The likelihoods are plotted in figure 5.15. The WT (blue) likelihoods are wide and irregularly shaped, but produce relatively accurate estimations of ZT 36.6, 36.8, 35.7, 33.0, and 36.4, with corresponding Θs between 0.03-0.13. There is a mean absolute error of around 2 hours for time estimations of the WT data. This estimation error, but good Θ values, could be explained by discrepancy arising from the use of mice in constant darkness to train Time-Teller to estimate the time of mice that have been in regular LD cycles.

Time-Teller reports that the Rev-Erbα KO mice have no functioning clock, and reports with extreme significance the difference between Θ values between the WT and KO groups ($p = 0$). The KO (red) likelihoods are almost entirely flat in figure 5.15. The KO data produces Θ values between 0.75 and 1 so time estimations are not accurate, and indicating that there is no functioning clock. The box plots in figure 5.16 show the difference in Θ values for the WT and KO data.

Time-Teller accurately and confidently predicted the time of the WT samples and gave low confidence metrics for the KO data, indicating that there is no measurable clock.

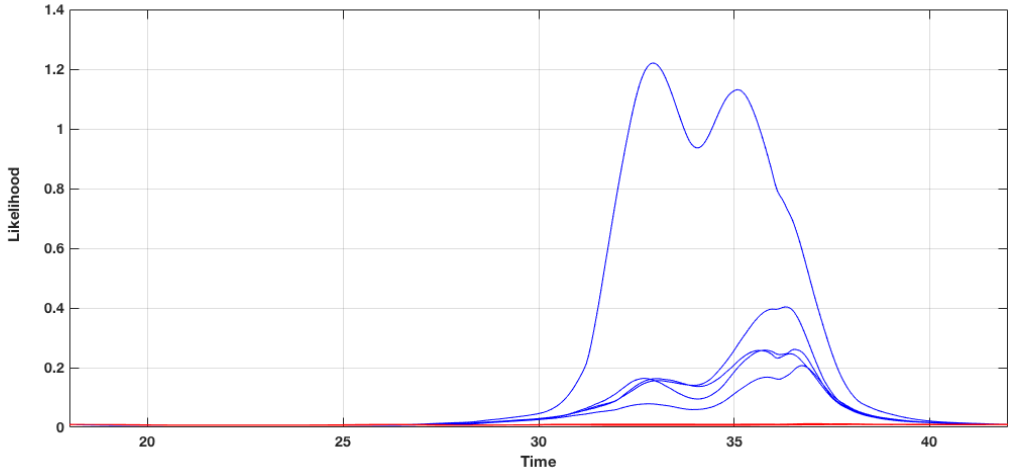


Figure 5.15: **Plot showing likelihoods for the Time of samples of the Fang data**. Blue curves represent WT samples with clear peaks around CT30-38, and the negligible amplitude curves in red for the KO data.

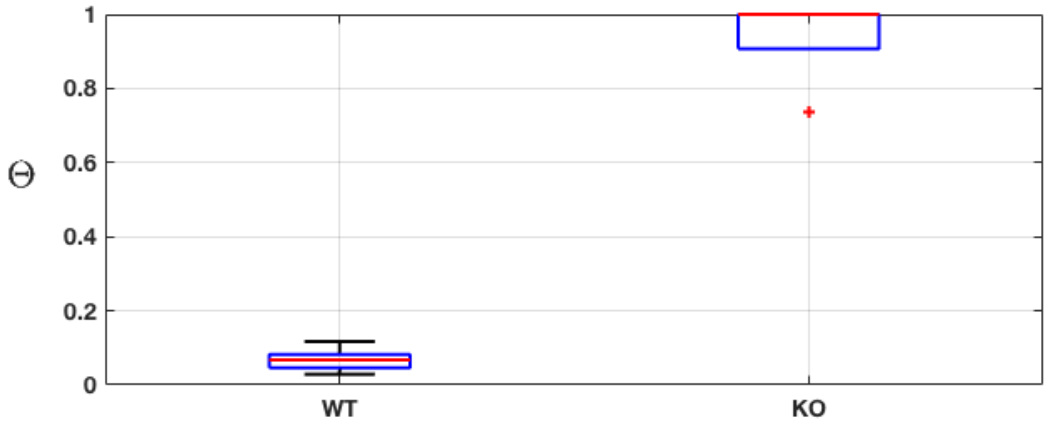


Figure 5.16: **Box plot of Θ values for WT and KO data**. There is a very significant difference between the groups - Wilcoxon's logrank test $p = 0$.

5.2.4 Barclay data: WT vs time-stressed

It is important to ask the question of whether a clock is dysfunctional or just perturbed from regular conditions, and to understand how Time-Teller model deals with this. Barclay *et al.* [106] published a dataset that uses all WT mice, but half of the mice were sleep deprived when they would usually be sleeping.

All mice were synchronised to LD cycles and given food *ad libitum*. The group of stressed mice were kept awake during the first 6 hours of light (ZT 0-6) on days 1 through 5 and days 8 through 12 using a gentle handling approach designed to minimize stress effects and intervention by the experimenter. At all other times mice were left undisturbed, and the other group of mice were left completely undisturbed. Liver and adipose samples were taken from 3 mice in each of the 2 conditions at ZT1, ZT7, ZT13, and ZT19.

Time-Teller very accurately estimated the time of the normal condition liver samples, shown as the blue circles on the left plot in figure 5.17 and shown to have a median Θ value around 0.06 in figure 5.18. There is also a high level of estimation accuracy within the normal sleep schedule group for the adipose data, with the exception of an $\sim$7 hour error for some final data points for the adipose tissue. The Θ values for these white fat CT19 data points are all $<$0.05 and all 3 replicates have the same result. The accuracy of the rest of the time estimations for the normal schedule data might well lead one to question the data rather than Time-Teller results.

For the time stressed group, some samples are inaccurately predicted, as shown on the right plot in figure 5.17. The Θ values for the time stressed group are generally higher, as shown in the boxplot in figure 5.18. As sleep stressed clocks are not technically classed as dysfunctional [145], we would not expect the Θ values to be close to 1, as with KO data. There is however, a notable difference of Θ values for the normal and sleep deprivation groups, which is significant in the liver data ($p = 0.02$) and less significant in the adipose date ($p = 0.1$).

This suggests that sleep stress actually changes circadian gene expression to a profile different than any gene expression profiles that occur in normal sleep-wake conditions. This provides evidence that a chronic change to the sleep-wake schedule results in abnormal circadian gene expression patterns, and not just a change in body time.

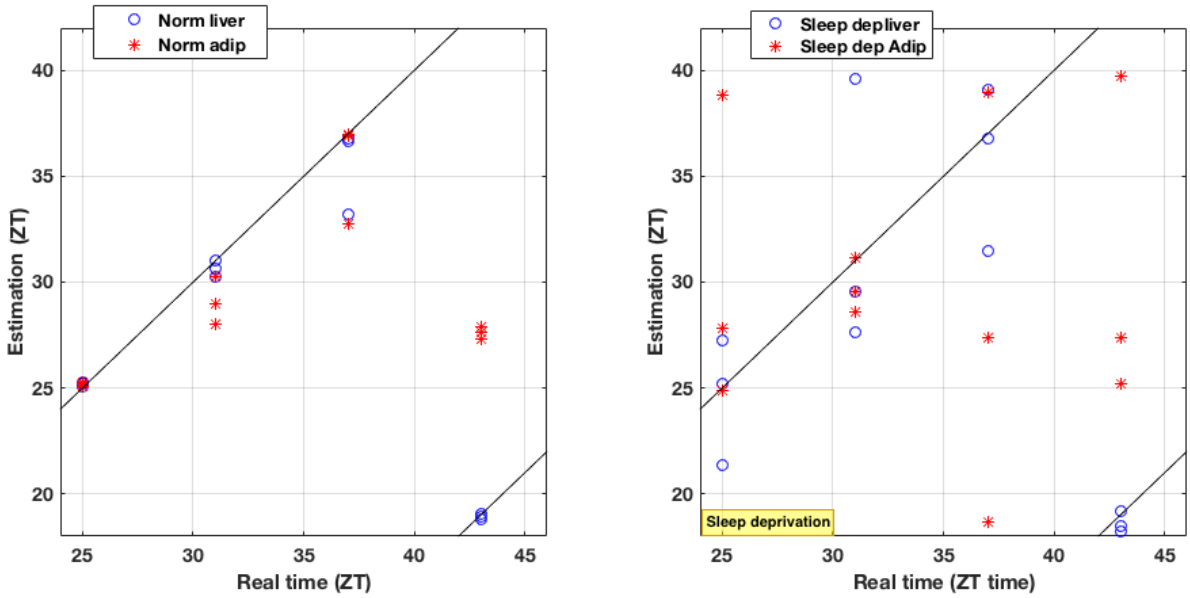


Figure 5.17: **Scatter plots showing real times vs estimations for the Barclay data.** Normal sleep schedule mice data points are shown on the left, and have less variance than the sleep deprived mice estimations in red.

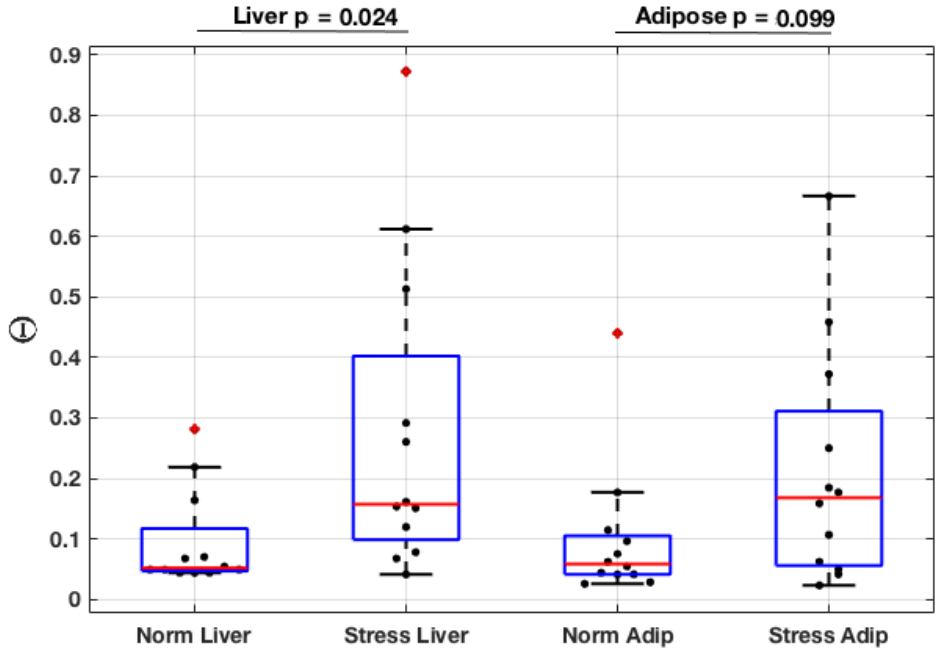


Figure 5.18: **Boxplots showing distributions of Θ metrics for normal and stressed, Liver and White adipose tissue samples.** The samples from mice in normal conditions have lower Θ metrics than their stressed counterparts.

5.3 Human Time-Teller applied to independent datasets

Working with human data raises many more issues. The training set is now much smaller, we are not working with genetically identical humans, and no data set exists with healthy human samples with annotated times the samples were taken, that also use the U133 2.0 GeneChip (that we know of). All of the data used in this thesis originated from live tissues, and not *in vitro* samples, due to the expectation that this would represent the true circadian clock due to the complex input signals from the rest of the body.

In order to attempt to keep the variance low, we start off the human validation of Time-Teller with datasets originating from oral mucosa, like the training data. We make the realistic assumption that all (live) tissue samples will be taken between 9am-5pm[2], as with most of the human data, the real timings of the samples are unknown. Correspondence with some of the authors of the studies presented below confirmed that this is an appropriate assumption.

Distribution of Θ for Bjarnason human data

Although there were 16 probes identified in chapter 3 as rhythmic and synchronised, and these 16 probes were used in the leave-one-out validation in chapter 4, we only use 15 probes going forward. The reason is that the Per1 probe 244677_at was found to have significant signal issues in many of the independent datasets, i.e. the signals values were very low. As there is another Per1 probe in this dataset, we can conclude that this is a probe issue, and not an issue with the Per1 gene expression.

Time-Teller was used to find the Θ values for the training data, using the full 10 individuals as the training data. This defines the Θ distribution for healthy functioning clocks and is shown in figure 5.19. For most human samples in the training set of the Human Time-Teller, $\Theta < 0.09$, with maximum values at $\Theta = 0.155$. These provide an approximate threshold for a "functioning clock" range when applying the human Time-Teller to independent datasets.

Figure 5.20 shows how (when using a leave-one-individual-out) approach as in the previous chapter, the Θ values for each individual are comparable. Male15 and female18, who are phase delayed and phase advanced, have normal Θ values for these time estimations, indicating that they have functioning internal clocks even though their body times are different to average.

[2]9am-5pm is the likely time interval that hospitals would work in.

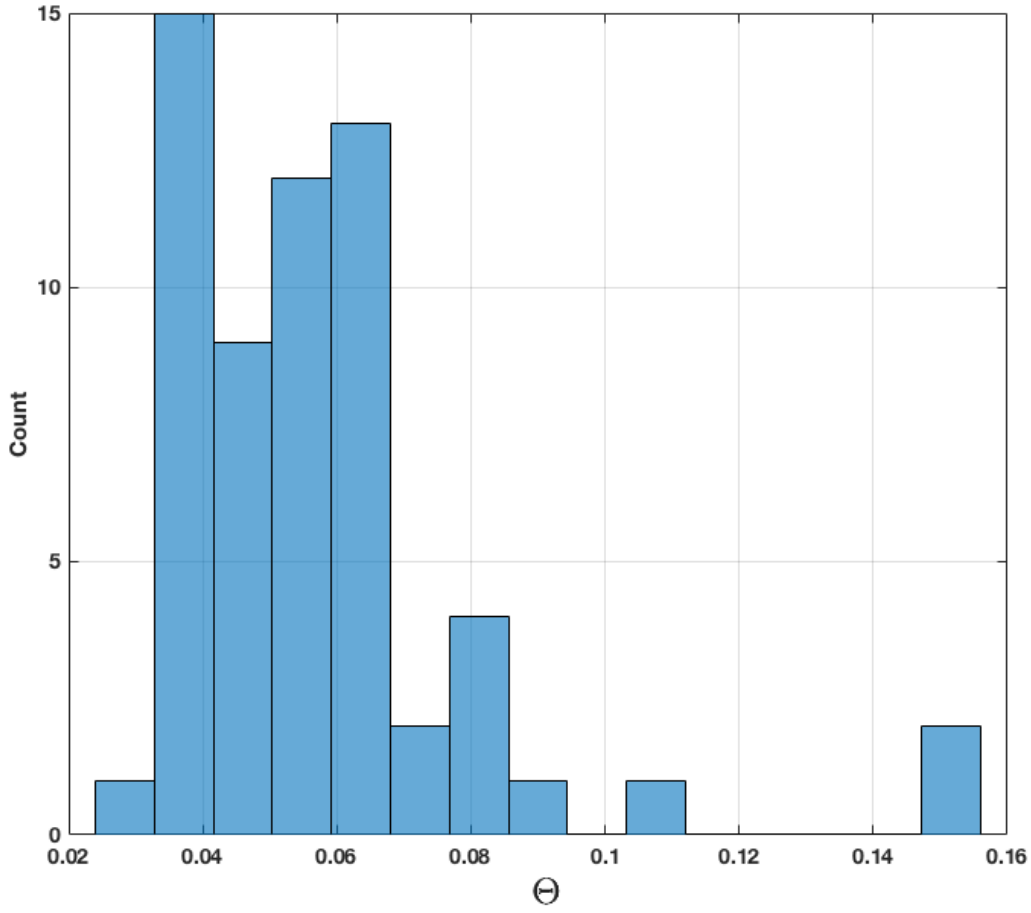


Figure 5.19: **The distribution of Θ values for the human training set.** This uses 10 genes (15 probes). The majority of the samples have $\Theta < 0.09$, but the maximum value is $\Theta = 0.155$. All 10 individuals were used in the training set to get these values.

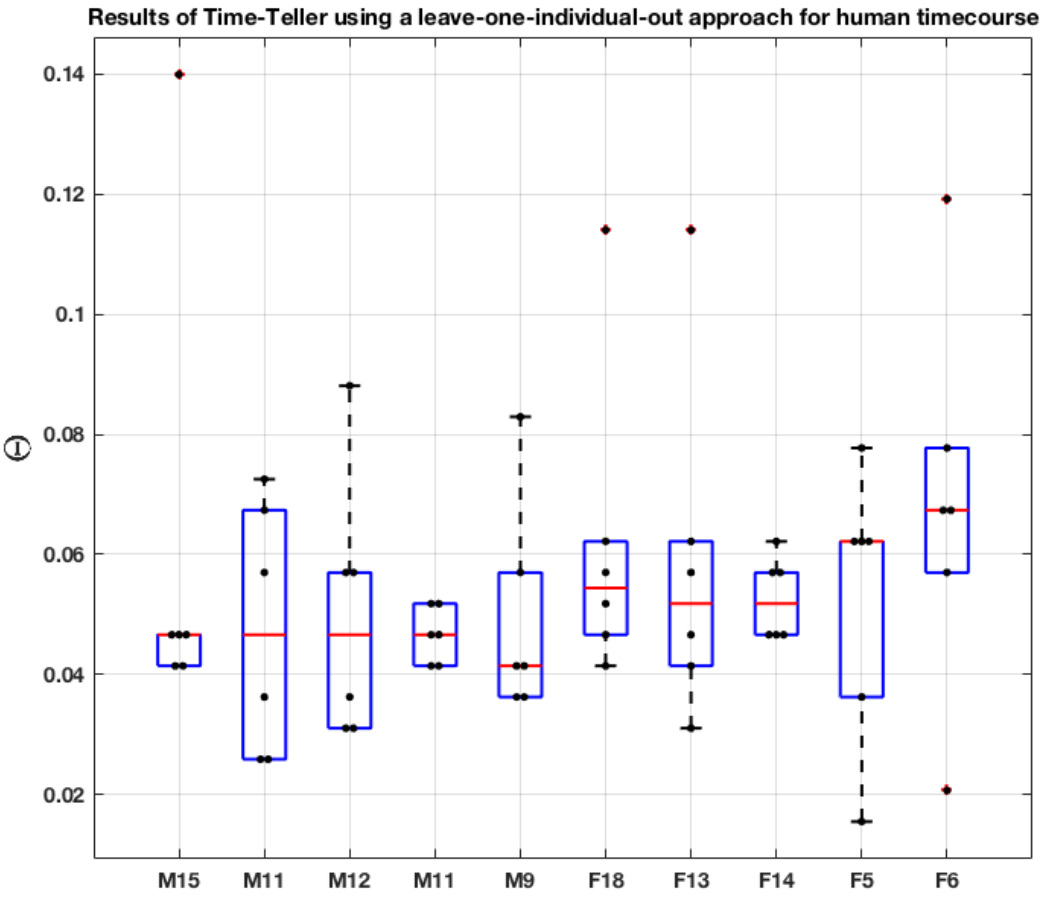


Figure 5.20: **The distribution of Θ values for the human training set, by individual.** This uses 10 genes (15 probes). A leave-one-individual-out approach was used to get these figures.

5.3.1 UK-Sri Lankan healthy data

Saeed *et al.* [108] conducted a study comparing UK and Sri Lankan oral squamous cell carcinomas. The control data for this study includes 5 healthy control oral mucosa samples from the UK and 3 healthy control samples from individuals in Sri Lanka. The study states that the samples were handled using identical protocols for tissue collection and processing[3]. All individuals were healthy, with low risk factors for cancers of the mouth, but there is no other information given on the individuals or their lifestyles. Time-Teller was used to estimate the time these samples were taken. The likelihood curves are shown in figure 5.21, where the red curves represent the UK samples. The likelihoods for the Sri Lankan samples are plotted in black and have far lower amplitudes.

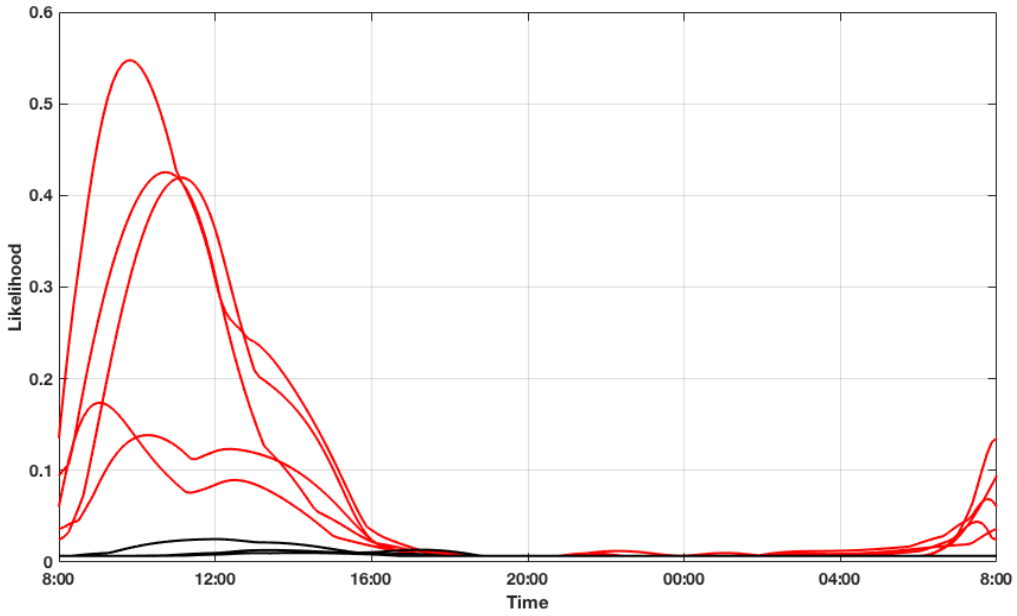


Figure 5.21: **Likelihoods for the time of sampling of the UK(red) and Sri Lankan (black) oral mucosa samples.**

The Θ values for the 5 UK samples are 0.05, 0.19 , 0.09, 0.08, and 0.07, with MLEs of 9:07am, 10:22am, 10:52, 11:15am, 10:00am. These appear to be realistic values reflecting our expectations. The Θ values for the 3 Sri Lankan samples are 0.66, 0.55, and 0.69 (with MLEs all around 2am), indicating that there is something wrong with these estimates. The reasons behind this can only be speculated upon. It could simply be that something went wrong during the sample handling in the many stages of the experiment from sample taking to loading to plate. Or, there could be a much more complex reason for this: the training data originated half a world away from Sri Lanka, in Canada, and very little is known about the circadian clock behaviour differences for people living in different

[3]Unfortunately, a different (custom) microarray was used to assess the tumour samples, which excluded the majority of the clock genes.

lifestyles, climates, cultures, or latitudes. In the future, if this kind of Time-Teller model would be used clinically, such differences (if in fact there are any) would have to be better understood.

5.3.2 Smoker and non smoker oral mucosa data

Oral mucosa transcriptome data was published by Boyle *et al.* [109] as part of a study on the effects of smoking on the oral mucosal transcriptome.

40 current smokers and 40 age and gender matched never smokers underwent buccal biopsies. Eligible subjects were healthy volunteers (except for the effects of smoking in the smoker cohort). One sample was excluded from the study based on a quality measure leaving 79 for analysis. Time-Teller's MLEs are summarised in the histograms in figure 5.22. Most of the estimated times are in the morning, between 6am-1pm, which are earlier than expected, but mostly realistic estimates. 15 of the estimates are in the very early morning, between 1am and 5am and are less realistic, but could reflect natural variations in body time, or the individuals could have been shift workers. All the estimations between 1am-4am have $\Theta > 0.14$

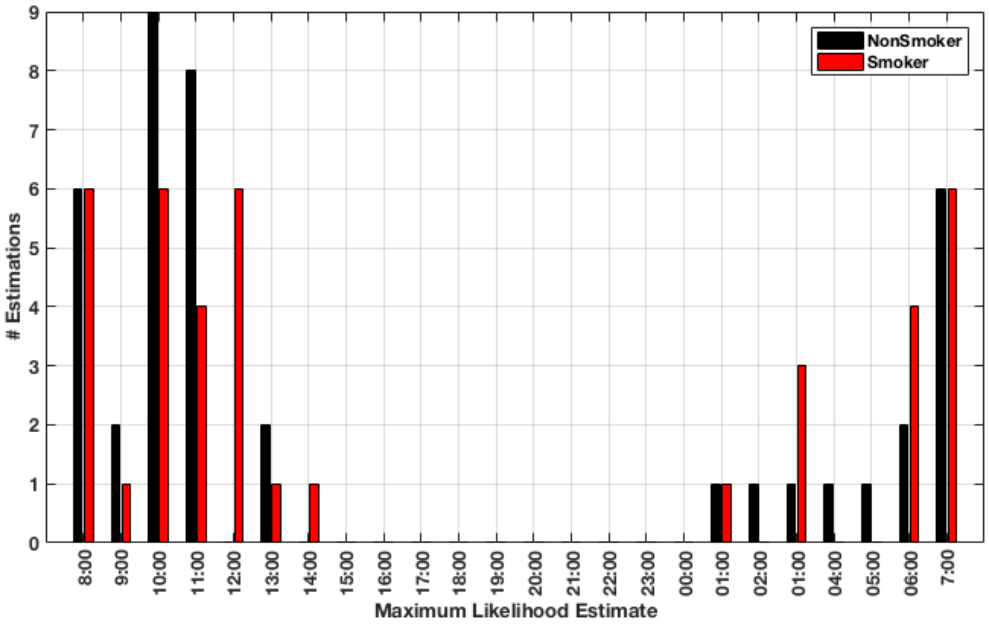


Figure 5.22: **Histogram summarising the MLEs for time of 40 non-smoker samples, and 39 smoker samples.**

The boxplot in figure 5.23 shows that the majority of the 79 non-smoker and smoker samples have $\Theta < 0.155$, which are in the "functioning clock" range. The median Θ for non-smoker samples is smaller than the median Θ for the smoker samples, but with an insignificant Wilcoxon test statistic of $p = 0.083$. This might suggest that the smoker group has more inter-subject variability for the functioning of the circadian clock.

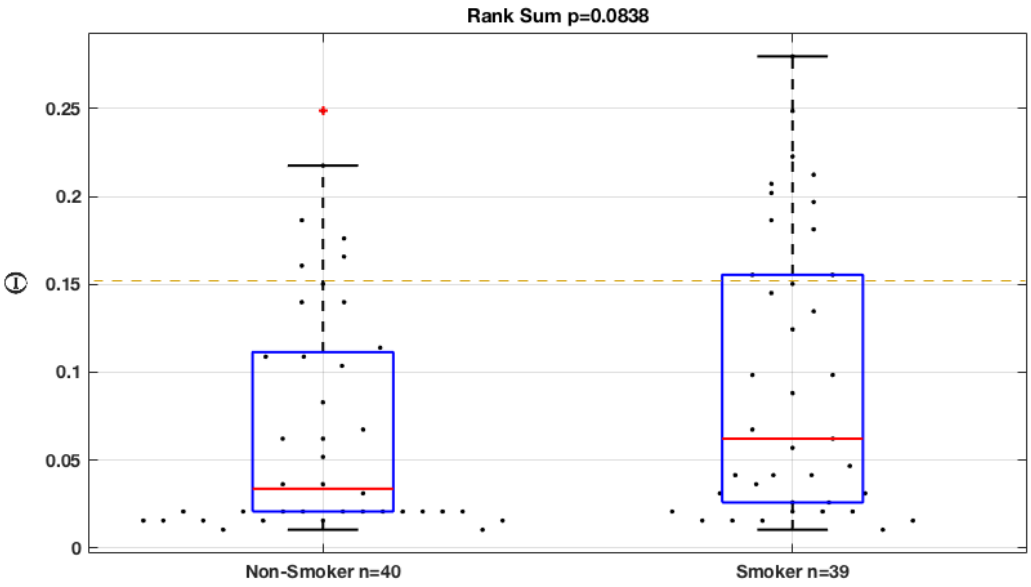


Figure 5.23: **Boxplot showing distribution of Θ values for the non-smoker and smoker samples.** Smokers generally have higher Θ values, but with an insignificant test statistic $p = 0.084$.

5.3.3 Autopsy data

The vast majority, if not all, live human tissue samples are taken in the daytime due to practical reasons. This makes it difficult to test the human Time-Teller model for estimating samples that were taken at night.

One very interesting dataset exists that can help with this validation step, but it does not come without some issues. The data published by Roth *et al.* [110] originates from autopsy samples of multiple organs of 10 individuals. The samples were taken up to 8.5 hours after death, so we do not expect that this data will perfectly show transcriptome fingerprints of well functioning clocks. We know very little about post mortem gene expression; it is unlikely that gene expression "freezes" at the moment of death, but it is also unlikely that gene expression continues as normal. These issues are likely to decrease the viability that Time-Teller can estimate the time of these samples (whether this is time sample was taken, or time-of-death, it is not defined here). Neither the time of death nor time sample was taken is published with this data. There is no reason to assume that all of the deaths occurred at the same time of day, or that they all died in the daytime.

The dataset is complex and is made up of 353 samples, but only the 184 non-brain samples will be used in this analysis. Datasets for individuals 1-4 also contain oral mucosa and head, neck and mouth samples, which are highlighted in this analysis. Time-Teller's MLEs for the 184 non-brain samples are shown in figure 5.24, where the maximum likelihood is shown on the y-axis, and the individual indicated on the x-axis. The variance of

estimations for each individual is high. Different organs probably have different mechanisms in place for gene expression and degradation after death, which probably accounts for some of the intra-individual variance seen in figure 5.24. The oral data is highlighted with black markers, and shows an "average" time estimate of all the organs in individuals 1-4, with less variance. From these results, we could hypothesise that individuals 1 and 2 died at similar times, and individuals 3 and 4 died at similar times, around 12 hours apart from each other.

This analysis has shown that Time-Teller can make estimations across the full 24 hours, without day-time bias.

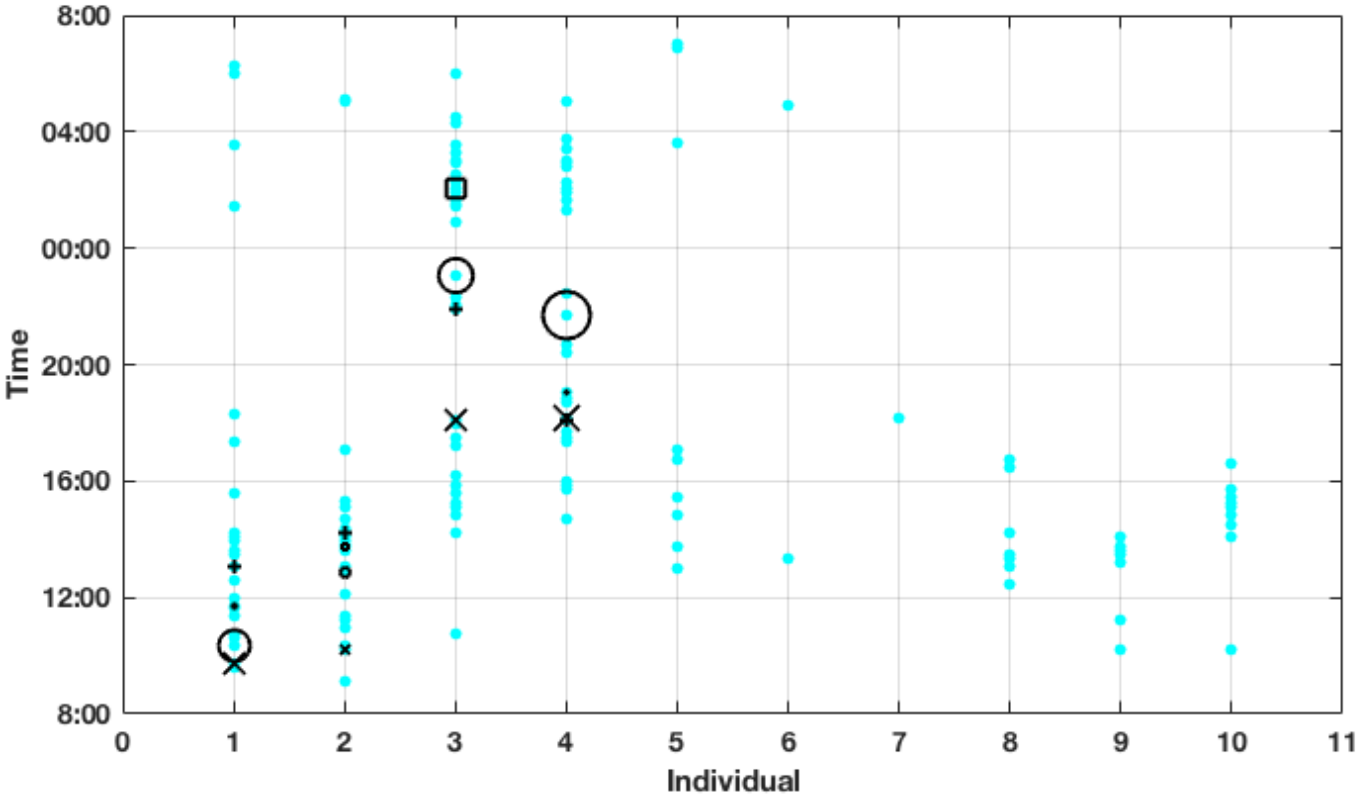


Figure 5.24: **Time-Teller results on autopsy data for 10 individuals across different organs.** Extra markers for Head/neck/throat samples for individuals 1-4 are marked as pharynx (plus), oral mucosa (cross), salivary gland (circle), tongue (square). The size of these black markers is scaled by $1/\Theta$. There is a lot of variation amongst individuals, but this is expected due to the low confidence in the data. This data shows that the Time-Teller can estimate times for samples during the night, and there is no obvious bias for day-time estimations.

5.4 Summary of Chapter

This chapter has presented the metric Θ as a measure of clock dysfunction, which is calculated using the likelihood functions produced by Time-Teller. Θ is calculated using a penalised likelihood ratio threshold.

The *in silico* Time-Teller was used to show how Θ is correlated with the level of an *in silico* KO, and how Θ is not sensitive to a change of the hyperparameters η and ϵ.

The mouse Time-Teller was able to accurately calculate the timings of the samples in the LeMartletot dataset, with "functional clock" Θ values. Time-Teller was able to detect KO samples with dysfunctional clocks, calculating $\Theta \approx 1$ for samples from full KO samples, and $\Theta < 0.155$ for WT samples. Sleep deprived mice were shown to have less functional clocks than their normal sleep schedule counterparts.

The human Time-Teller was able to provide realistic estimations with "functional clock" Θ values for 3 independent datasets. Problems were found with 3 samples from Sri Lankan individuals, but the lack of information about the individuals in the original study meant we can only speculate as to why this is. Data from autopsy samples were used to validate that Time-Teller could estimate that samples were taken during the night, with good confidence (i.e. low Θ values).

Time-Teller is an algorithm that provides an MLE for the time that a single sample was taken, along with a metric Θ which provides a measure of how functional the circadian clock is (and is a type of confidence measure for the MLE). The threshold for where Θ defines a "functional clock" is defined by the data set used to train Time-Teller.

Time-Teller could be used to tell if an individual's body time is in sync with clock time, i.e. if they are phase forward, or delayed with respect to some population average. This could be used in the clinic to inform on optimum timings for personalised chronotherapy regimes.

However, the clinical use of Time-Teller is not just in its ability to tell the time of a sample; it is in its ability to quantify the molecular circadian clock dysfunction from a single transcriptome sample. The next chapter explores how Time-Teller could provide insights into the circadian clock dysfunction in human cancers.

Chapter 6

Circadian Clock Dysfunction in Human Cancer

In 2007, the International Agency for Research on Cancer declared shift work that involved circadian disruption to be a "probable" carcinogen (group 2A), but noted that human evidence was limited [146]. However, the evidence to support circadian disruption as a carcinogen is growing, and is being researched from multiple, interrelated perspective. There are epidemiological studies that have tried to determine if shift workers have a higher incidence of cancer through statistic analyses of the gathered human metadata [147, 146, 148, 149]. Some studies use mice in jet-lagged conditions, with predispositions to cancer, to measure cancer incidence and growth in comparison to normal sleep schedule mice [16]. Other studies attempt to determine the mechanisms that drive the carcinogenic results of circadian disruption, but these are far more complex as the physiological, endocrine, and molecular rhythms must be tracked and related to tumourigenesis [17]. In parallel with this, a growing body of evidence supports the potential tumour suppressor role of the molecular circadian clock [150]. This research is being conducted on the regulating role of the molecular circadian clock for process involved in cancer. So far, we have evidence to support that:

- circadian rhythms are coupled to the cell cycle [66, 64] (dysfunctional cell cycle gating is a hallmark of cancer [151]),
- host genetic and functional circadian disruption promotes tumourigenesis [18, 150],
- and circadian rhythms are often disrupted in cancer [15].

Understanding of the role of the physiological inputs to the circadian clock, and the function of the molecular circadian clock in relation to cancer, could lead us to crucial information that allows the design of novel treatments. We already have evidence that optimal timings of administration of chemotherapy can increase efficacy and reduce toxity of cancer treatments; this is known as chronotherapy [152, 43, 153]. More information

could help in the design of chronotherapy strategies or the design of "clock resetting" treatments, in order to treat cancer.

Physiological circadian disruption promotes tumourigenesis

Here we discuss three categories of evidence that circadian disruption has an impact on both tumour risk and/or progression.

Evidence for shift work as a human carcinogen

There is a significantly higher prevalence of breast cancer in industrialized nations compared with developing countries [19]. This suggests that environmental aspects of modern society may play an important role in breast cancer risk [154]. A reason for this could be the western lifestyle resulting in disruptions of the internal body clock.

Multiple studies have been published that report an increased risk of cancer for nurses that work night shifts [147, 146, 148]. These are large scale, epidemiological studies using large complex datasets of human metadata, in some cases with inconclusive results due to the nature of the data [149]. For example, it is difficult to define what a night shift is, and different studies define night shift work as different things. Also, a lot of the data is a result of surveys filled out by women asked to think back about their schedules over their life times, which may not always be accurate.

Evidence for increased tumourigenesis in jet-lagged mice

There are various studies that investigate cancer prone mice which are kept under varying LD conditions. An example of this is the study published by Van Dycke *et al.* [16] where one group of cancer risk mice (with a p53 Li Fraumeni mutation) were kept in normal LD cycles, and another group were kept in alternating length LD cycles. Every week, mice were sacrificed, weighed, and tumours measured. The circadian disturbed mice gained more (relative) weight and developed tumours more quickly than those in a normal LD cycle. This is summarised in figure 6.1, where (A) shows that sleep disturbed mice gained more weight relative to food intake than normal LD scheduled mice. (B) shows a higher number of tumour baring mice in the disturbed scheduled mice with a Kolmogoroff Smirnoff Test statistic of p=0.0127 (a test for difference in distributions).

Papagiannakopoulos *et al.* [18] showed that both physiological and genetic circadian rhythm disruption accelerates lung tumourigenesis in mice. They genetically engineered mice so that lung tumours could be induced via p53 and K-ras pathways. After inducing these tumours to the genetically engineered mice, they kept one group of mice in normal LD conditions and one group in jet-lag conditions. They showed both a significant increase in tumour growth in the jet-lag group, and significantly reduced survival.

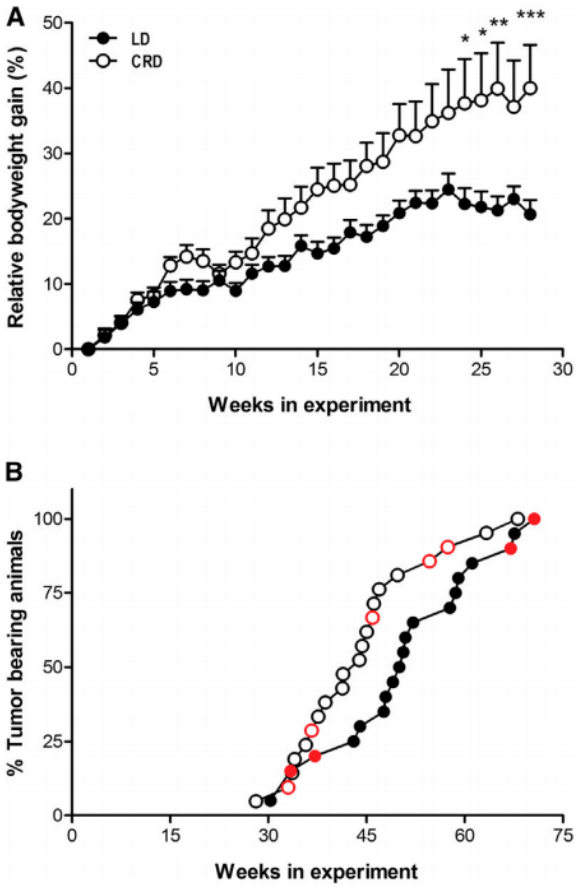


Figure 6.1: **Plots showing differences in weight gain and tumour numbers for normal and jet-lagged mice.** From [16]. (A) shows the relative body weight gain (normalised by food intake) and (B) The percentage of mice with tumours in normal LD cycles (n = 20; closed circles) or chronic jet-lagged conditions (n = 21; open circles). Black colour indicates mammary gland tumour, whereas red colour indicates other tumour types.

Evidence for decreased proliferation in synchronised cancer cell lines

Kiessling *et al.* [17] attempted to address whether enhancing circadian rhythmicity in tumour cells affects cell cycle progression and reduces proliferation. They used various cells lines and induced circadian synchronization with dexamethasone shock. Their main result is shown in figure 6.2, where the solid black line shows data representing the volume of the cells over a week that had been synchronised to a circadian rhythm. Both unsynchronised cells and Bmal1 KO cells (even with synchronisation) show much more growth. This result shows that proper expression Bmal1 has an inhibitory effect, when cells are shocked with dexamethasone, on tumour growth.

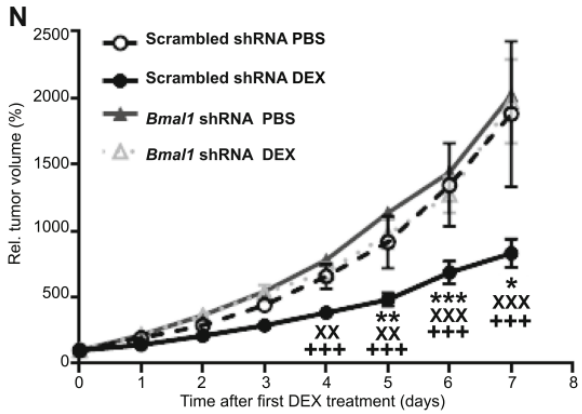


Figure 6.2: **Knock-down of Bmal1 prevents the inhibitory effect of dexamethasone on tumour growth.** From [17]. Cells with Bmal1 that are given a DEX shock have significantly lower relative tumour volume to cells with silenced Bmal1, or cells with Bmal1 that have not been synchronised with dexamethasone.

6.0.1 Genetic circadian disruption promotes tumourigenesis

Where physiological circadian clock disruption is studied with mice genetically engineered to have a high risk of cancer in changing LD schedules, genetic disruption is studied using mice with additional mutations in specific circadian clock genes (in normal LD conditions).

Papagiannakopoulos *et al.* [18] genetically engineered mice so that lung tumours could be induced via p53 and K-ras pathways and also engineered mice with a mutant Per2 or Bmal1 allele. After inducing tumourigenesis in the lungs, they found that mice whose circadian clock had been suppressed, had both more tumours, and lower survival. This is shown in figure 6.3 where (J) shows that the cancer prone mice developed more tumours with Per2 or Bmal1 knock-downs, and (K) shows significantly better survival in the mice with functional circadian clocks.

6.0.2 Mechanisms driving circadian disruption to promote tumourigenesis

The mechanisms by which circadian disruption acts as a carcinogen are not something that will be comprehensively addressed in this thesis, as this would require an expert understanding of molecular biology. However, as this is a very important area of research that is directly relevant the aims of this thesis, we present an overview of some of the key results in this research area. There is evidence that the mechanisms by which circadian disruption promotes tumourigenesis are consequences of:

- reduced production of melatonin or vitamin D [155],
- sleep deprivation resulting in the suppression of the immune system, metabolic changes favouring obesity, and the generation of pro-inflammatory cytokines [156],

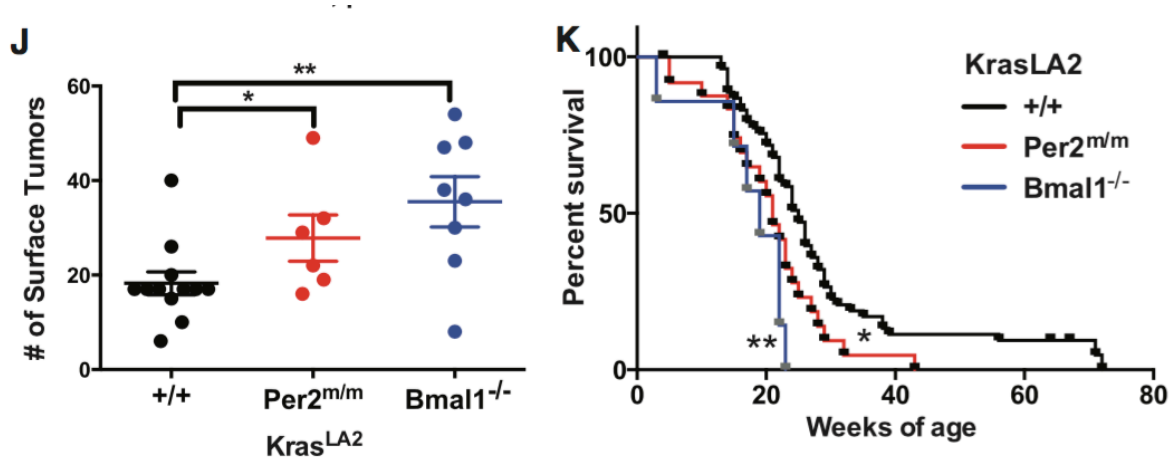


Figure 6.3: **Genetic disruption of circadian rhythms accelerates lung tumourigenesis** From [18]. (J) Number of Surface tumours in KrasLA2/+ WT (+/+) animals (n = 12), systemic loss of Per2 (Per2m/m) (n = 6), and Bmal1 loss (Bmal1-/-) (n = 8) are shown. (K) Kaplan-Meier survival analysis for KrasLA2/+ animals with WT (+/+) (n = 50), Per2m/m (n = 31), and Bmal1$-/-$ (n = 7).

- increased telomere shortening [157],
- and a predisposition to a natural variation of the circadian pathway [150].

Some of the mechanisms that have been proposed by Blakeman *et al.* [19] to link breast cancer and the dysfunction of the circadian clock are summarised in figure 6.4. They suggest that potential ways in which circadian disruption can drive breast cancer are through genetic defects, ageing, or shift work. A dysfunctional clock can directly disrupt the gating of the cell cycle and reduce apoptosis, and these effects can also be brought on indirectly through altered metabolism, in response to a broken circadian clock. They can also lead to elevated EMT (epithelialmesenchymal transition), driving the formation of lethal metastases. They also suggest that arrhythmic production of melatonin tips the balance towards tumourigenesis [19].

6.0.3 The circadian clock is coupled to the cell cycle

Cancer is a heterogeneous collection of diseases, but it is generally defined by the six hallmarks of cancer [151], the first being sustained and ungated proliferation resulting from the dysregulation of the signalling to the cell cycle. Due to this, the connection between the cell cycle and cancer is a widely studied field in oncology and molecular biology.

The circadian clock and cell cycle are two very different and very complex biological oscillators. This complexity and elaborate design probably arises from the criticality for proper control and timings, and the need for tight controls and backup mechanisms. Although the mechanisms of how the circadian clock and cell cycle communicate with each

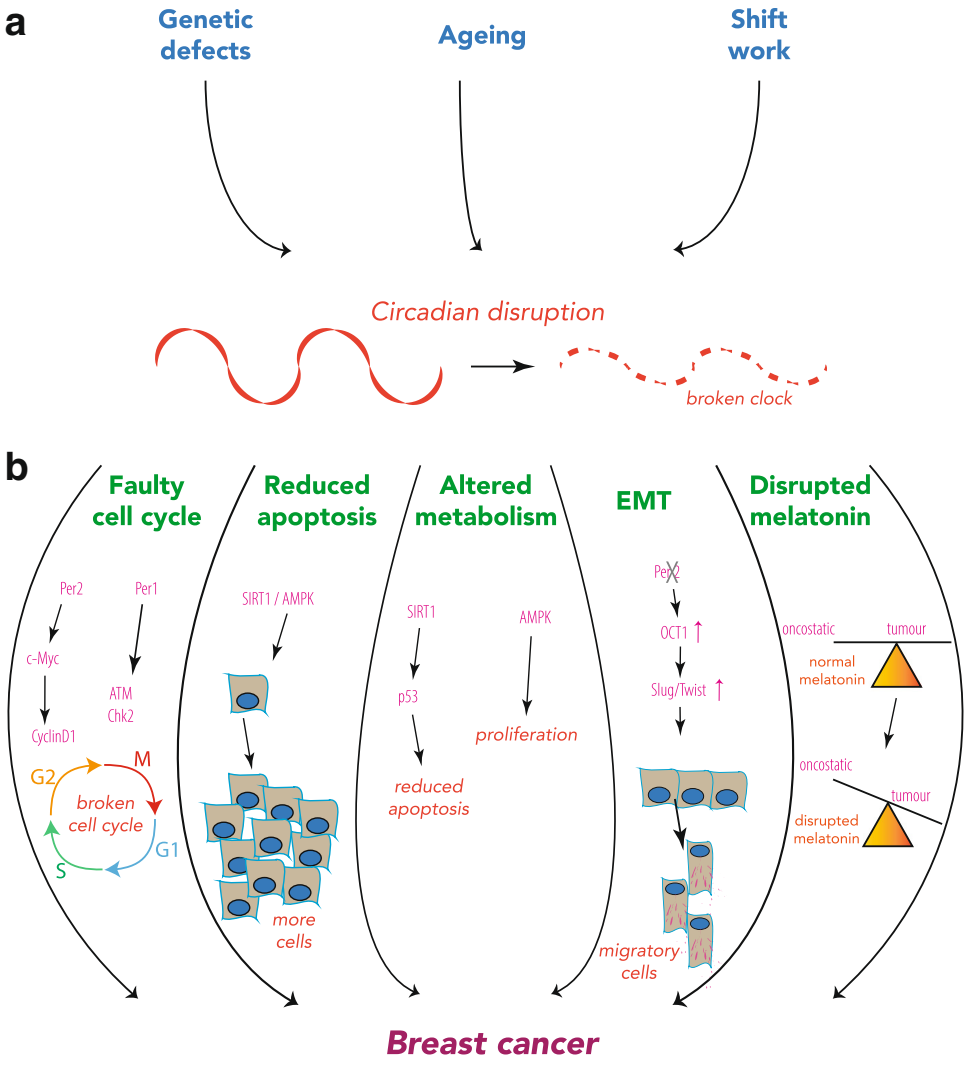


Figure 6.4: **Possible circadian disruption pathways to breast cancer.** From [19].

other are not yet fully understood, it is clear that they do. Indeed, Wee1 is a well-known circadian clock gene in mice (see figure 3.7). WEE1 kinase is responsible for a cell's entry into mitosis, by regulating the CYCLIN B1/CDC2 (cell division control protein 2 also known as cyclin-dependent kinase 1, CDK1) complex [68].

The circadian clock has been shown to control multiple aspects of the cell cycle, conferring gating mechanisms at key checkpoints of the cell cycle to ensure fidelity in DNA replication and cell division.

There is some evidence that the oncogene MYC interacts with core circadian clock genes [158]. The circadian control of many oncogenes and tumour suppressor genes are being investigated for other mechanisms that might link circadian disruption and cancer [159]. There is increasing amounts of evidence that the circadian clock and the cell-cycle are coupled oscillators [66, 64], which generally indicates that the coupling is bi-directional. Mathematical models have been developed in order to explore the dynamics of these coupled oscillators [66, 37, 160].

6.0.4 Chronotherapy

Research into the association between circadian clock dysfunction and cancer promises to be a successful area of research for new treatments and targeted therapies. The dose for conventional chemotherapies is usually found by dose increments until specific toxicities present [5]. Chronotherapy is the practise of administering treatments according to biological rhythms, in order to improve the efficacy and tolerability of treatment [152]. Circadian changes in tolerability are expected due to cells being more susceptible to damage at certain phases of the cell cycle, and hence more tolerable at others. Cancer cells are not expected to be maximally tolerant at the same time as healthy cells, as the circadian clock is likely to be dysfunctional in cancer. Hence it is possible to find the optimum time of day where healthy cells are shielded and cancer cells are maximally targeted.

Chronopharmacokinetics

Chronopharmakokinetics is the field of study of the response to rhythmic exposure to a drug and its metabolites (chrono PK), and the rhythmic organization of the drug targets (chrono PD) via mathematical modelling [37]. If patient-specific parameters could be identified, this systems approach could enable optimised and personalised chronotheraputic drug regimes [37].

Evidence in mice

Li *et al.* [152] showed that optimal chemotherapy timing could be predicted by clock gene expression timecourse patterns. They found that the high amplitude mRNAs Rev-Erbα and Bmal1 clock markers were both critical determinants for optimal timing of tolerability.

Evidence in humans

Giacchetti *et al.* [161] report results from human clinical trials of chronotherapy on patients with metastatic colorectal cancer. Three international phase III clinical trials were conducted, where both chronotherapy and conventional chemotherapy were administered using oxaliplatin, 5-fluorouracil, or leucovorin. After 9 years the status of 345 females and 497 males was recorded. They reported that males who had been treated with a chronotheraphy regime had a significantly improved tumour response rate, progression-free survival, and overall survival as compared to conventional chemotherapy treatment regime. They report an opposite effect for females, suggesting that sex-dependent toxicities could be possible.

Tumours with dysfunctional clocks are more responsive to chemotherapy

One of the hallmarks of cancer is sustained, proliferative signalling [151]. Most chemotherapies target these dividing cells when they are at a vulnerable stage of the cell cycle.

Tumours with disrupted circadian clocks can respond better to chemotherapy because the cell cycle is less regulated. These tumours have more cells in the cell cycle because the clock disruption alleviates an inhibitory control on cell cycling. If more cells proliferate, more cells are in a susceptible state to chemotherapy, and response to chemotherapy is better. However, these tumours are also more aggressive, as we will show in the next section.

Lee & Sancar [162] published a study with the finding that circadian clock disruption improved the efficacy of chemotherapy. They saw that cells with p53 mutations were more sensitive to chemotherapy when Cry1 and Cry2 were also mutated.

Chronotherapy and Time-Teller

It would be very useful for studies in chronotherapy to be able to know ahead of treatment plans, if a tumour has a functional circadian clock. That way, treatment regimes could be tailored to the tumour. It is not viable to take timecourses of tumours. One time point sample is all that is likely to be available for use in these estimations. This is where Time-Teller provides a solution to better inform cancer therapeutics in the clinic, and not just in general research.

6.0.5 The circadian clock is dysfunctional in cancer

This chapter will now discuss current evidence whilst providing new evidence that the circadian clock is dysfunctional in many cancers.

The proposed mechanisms behind the association between the circadian clock and cancer genetics are not discussed in detail here, but a good review is given in [163].

This review also provides a summary of major findings in some tumour transcriptome studies, of whether any clock gene expression were significantly increased or decreased in tumours. Although most studies report a downregulation or decrease in expression levels, some report the opposite finding. This supports an assumption of this thesis: that dysregulation of rhythm does not necessarily mean that expression is decreased, and therefore a different metric is needed.

The circadian clock being dysfunctional in cancer is often assumed as this assumption is precursor to the fact that genetic circadian disruption promotes tumourigenesis (the genetic circadian disruption must first exist before it can drive tumourigenesis). The vast majority of the evidence for genetic circadian disruption promoting tumourigenesis involves a genetically engineered mouse whose circadian clock has been purposely disrupted.

Limited evidence has been produced to show that circadian rhythms are disrupted in a living tissue environment. There is evidence of faulty rhythms in the tumours of genetically engineered mice [164, 165], in cell lines [166], and in human cultured primary cells, that rhythms are weaker in cells with a diseased phenotype [167]. However, there is a lack of *in vivo* longitudinal time series data on molecular clock genes or proteins in solid tumours. There is a huge difficulty in acquiring this data, so sound evidence of the dysfunction of the circadian clock in cancer has not yet been generated by biological studies.

In fact, most of the evidence that exists towards showing that the circadian rhythms in cancer are faulty is contained in the mathematical studies discussed in chapter 4. CYCLOPS [13] and Zeitzeiger [12] both report a "different to normal" behaviour of their model's outputs when their algorithms are applied to cancer data. Shilts *et al.*'s ΔCCD study [15], *Evidence for widespread dysregulation of circadian clock progression in human cancer* (discussed in chapter 4) contains some of the most convincing evidence towards this hypothesis. Time-Teller can further substantiate this.

6.1 Time-Teller on human cancer datasets

Time-Teller provides a way to robustly estimate the circadian clock disruption in single samples which could help to design tailored therapies, and also provides evidence in support of the hypothesis that circadian clock is dysfunctional in many cancers. We begin using a data set of transcriptomes of oral cancers with some healthy controls.

6.1.1 Feng data: healthy and OSCC

Feng *et al.* [168] conducted comparative analysis of healthy oral mucosa transcriptome and oral squamous cell carcinoma (OSCC) transcriptome. The dataset contains 229 samples

in total, 167 of OSCCs, 45 of normal oral mucosa, and 17 samples are of dysplasic oral mucosa tissue. Dysplasic tissue is abnormal tissue that could signify early signs of cancer.

Time-Teller was applied to this data. The resulting MLEs are plotted against the Θ values in figure 6.5. The black pluses show normal samples with realistic estimated timings (9am-3pm), and the blue circles show a similar result. The red crosses represent the cancer data and show some unrealistic estimations during the night, but all with very high Θ values.

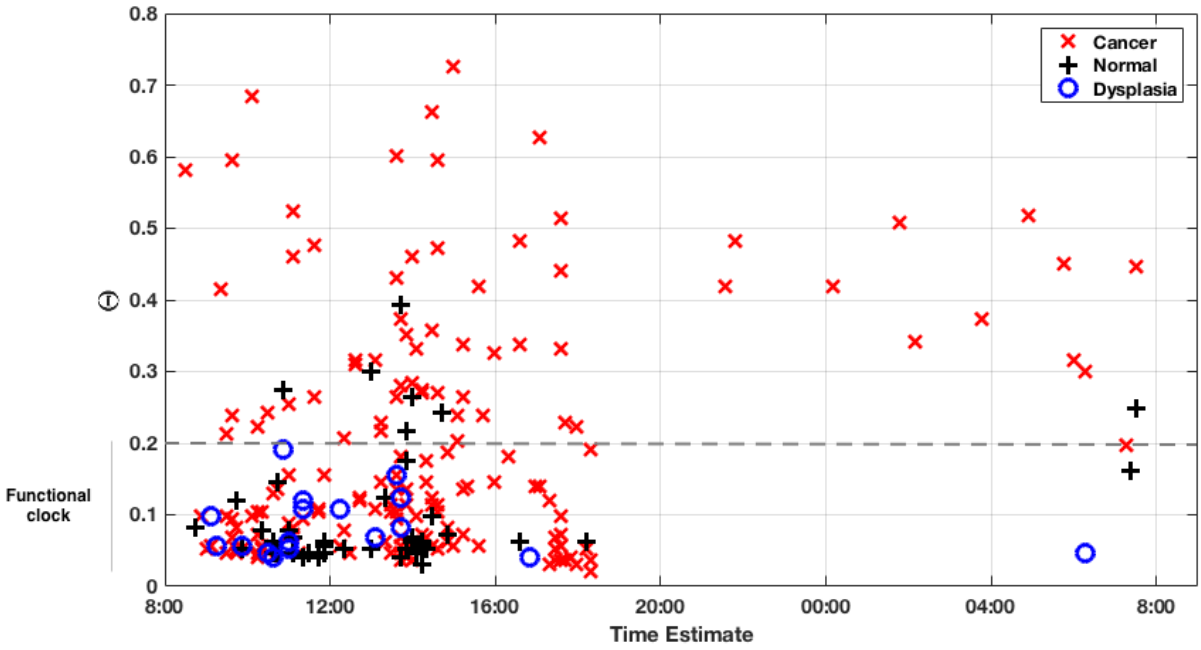


Figure 6.5: **The Maximum likelihood estimates plotted against Θ values for each prediction from Time-Teller for the Feng dataset.** The estimates are generally between 9 am and 6 pm, where the majority of the estimates outside of this time range are for cancer samples (red crosses).

The boxplots summarising the Θ distributions for the different groups are shown in figure 6.6. Where the normal and dysplasic samples are combined, the Wilcoxon Rank Sum test significance value is $p = 0.0003$. This represents the probabilty that the distribution of Θ for the cancer samples has a higher median than the normal and dysplasic samples. The majority of the normal and dysplasic samples have $\Theta < 0.1$, indicating a healthy clock. More than half of the cancer samples have $\Theta > 0.1$, indicating a larger variation of Θ values for cancer samples. This provides significant evidence for the disruption of the circadian clock in these OSCCs, compared with healthy samples.

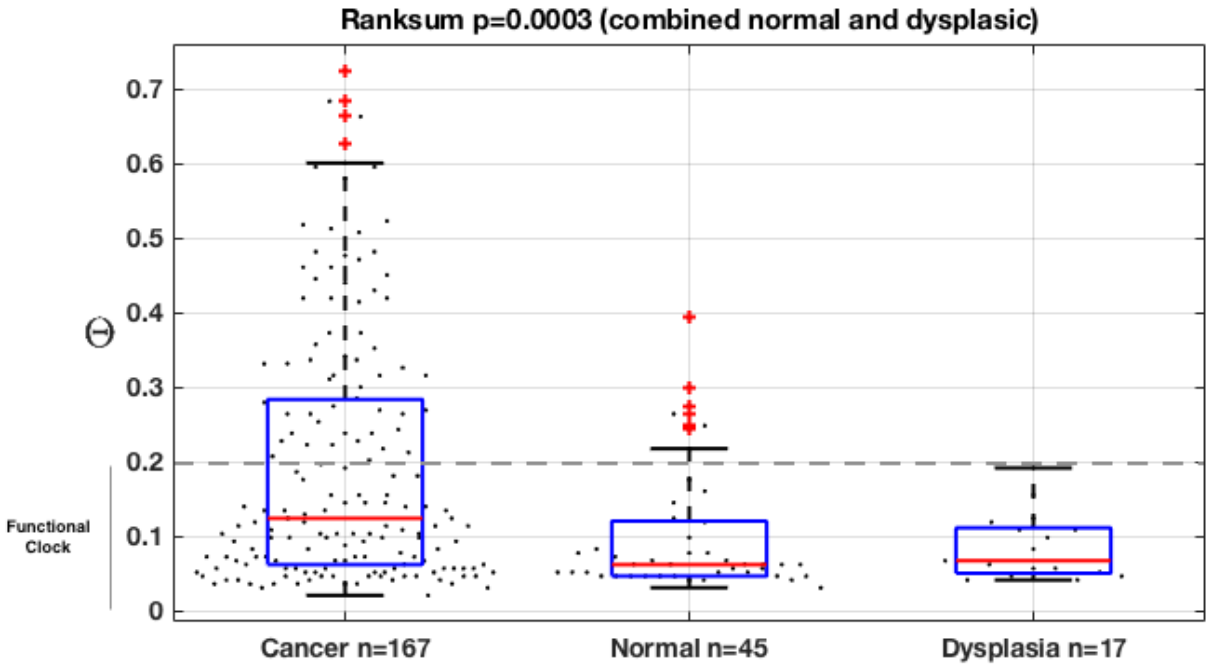


Figure 6.6: **The distribution of Θ metrics calculated by Time-Teller for the Feng data, by group.** Between the cancer group ($n = 167$) and normal/dysplasia group $n = 62$, the difference in median value is highly significant with $p = 0.00003$)

6.2 Breast cancer and circadian rhythms

The link between breast cancer and circadian clock dysfunction has been the central concept of many studies, review articles [19, 152], and has received much popular media attention. We focus on the literature reporting clock dysfunction links to breast cancer, and the results of Time-Teller applied to breast cancer data, for the rest of this chapter.

6.2.1 Summary of breast cancer prognostic markers

Breast cancer is a biologically heterogeneous disease, and the use of biomarkers ensures breast cancer patients receive optimal treatment. A very good summary of these prognostic markers is given by Taneja *et al.* [169], but we provide a simple overview here.

About 80% of all breast cancers are estrogen receptor positive (ER+), and 65% are progesterone receptor positive (PR+). The cancers grow in response to these hormones. Additionally, around 20% of breast cancers have a high number of HER2 (human epidermal growth factor) protein receptors, which stimulate the cells to divide and grow. These positive receptor cancers can be treated by targeted therapies to block the receptors and hence the signals that stimulate them to grow.

A tumour which is ER-, PR-, and HER2- is called triple negative breast cancer (TN), and accounts for around 15% of all breast cancers. Conventional chemotherapy would be used to treat these tumours. The histologic grade of a tumour, usually given a number 1-4, is assigned to tumours as a measure of cell abnormality, and how fast the tumour is likely to grow (4 being the most abnormal and aggressive). Tumour staging according to size are given the labels T2 for tumours greater than 2cm and less than 5cm in diameter, T3 for tumours larger than 5cm, and T4 for a tumour of any size that has spread beyond the tumour boundary into the chest wall or to the skin. Nodal status refers to a grading as to how many axillary lymph nodes the cancer has spread to; N0 being 0, N1 being 1-3, N2 being 4-9, and N3 being 10 or more. Nodal status and size of a tumour are often correlated [170]. Large tumours with a large number of axillary lymph nodes will often result in the worst outcomes.

The measurement of response to treatments involve tumour shrinkage, including complete disappearance of all clinically or radiologically detectable tumour deposits. Pathological complete response (pCR) is said to be achieved if there is an absence of infiltrating tumour in breast and lymph nodes after treatment [171].

Other endpoints are measured by survival outcomes. Disease free survival (DFS) is defined as the time from surgery to death, or cancer recurrence. Event free survival (EFS) is defined as the time from surgery to death, recurrence, or a complication from the treatment (e.g. pain). Overall survival (OS) is defined as the time from surgery to death [171].

Loss of circadian clock gene expression is associated with tumour progression in breast cancer

Cadenas *et al.* [172] examined the expression of 17 canonical clock genes in a collection of 766 breast cancer patients who were node-negative at the time of diagnosis and did not receive any chemotherapy. Clinical pathology factors such as MFS (metastasis free survival), grade, etc., were available. The microarray GeneChip used was HG-U133A, and unfortunately is not compatible with the training set used to train the human Time-Teller model presented in this thesis.

Cadenas *et al.* did not examine the clock function, but rather the transcriptional activity of each individual clock gene, and its relation with patient survival simple. They did simple comparative and correlation analysis of the dataset, and reported the main finding to be that "high expression of several clock genes (Clock, Per1, Per2, Per3, Cry2, Npas2, Rorc) was found to be associated with longer MFS". Other findings were reported associating higher levels of Per2, Cry2, and Per3 to lower grades of tumour, and ER- tumours. They claim that these results suggest a loss of expression of core clock genes, mainly those involved in Ebox regulation are associated with a worse prognosis in breast cancer.

In samples with functioning clocks, Npas2 should be maximally expressed 12 hours apart from Per1, Per2, Per3 and Cry2, with Rorc somewhere in between (see figure 3.4). In a microarray, amplitudes of circadian oscillation can span between 2^4 and 2^{12} (arbitrary expression units) so actual level of expression hugely depends on the time of the day. The circadian nature of these genes is not addressed in this study: the analysis was conducted as would any other gene expression analysis. There is no evidence in this study that the patients with more aggressive tumours were not taken at a different time of the day: which would invalidate all of the results presented.

The molecular circadian clock is such a complex network of oscillators, it is unlikely that clock dysfunction can be studied with such simplistic methods. This is a role that Time-Teller can fill.

Richardson data: different types of breast tissue

Richardson *et al.* [112] published a study looking at the gene expression signatures of different classes of breast tumours, with some healthy tissue as a control. The study was not looking at circadian rhythms and is a complex genomics study that will not be discussed here. The data however, is perfectly suited to be used by Time-Teller. 42 tissue samples of 3 different breast tumour types, and 7 normal tissues were processed using microarrays (Affymetrix HG U133 2.0, so is compatible with Time-Teller).

The tumour types were Basal-Like Carcinoma (BLC), BRCA1-, and non-BLC. BLCs

are aggressive, high grade, triple negative tumours, and BRCA1- tumours have a similar phenotype [112]. Without going into detail, we can assume that the non-BLC tumours refer to tumours which are ER+, most of which are PR+, and some also HER2+ [112].

Time-Teller estimated the MLE and the clock dysfunction metric Θ for each sample. The results are shown in the scatter plot in figure 6.7, showing that the 7 healthy samples were taken between 11 am and 5 pm, which is reasonable. The Θ values for the grouped tumours and the healthy samples are summarised in figure 6.8.

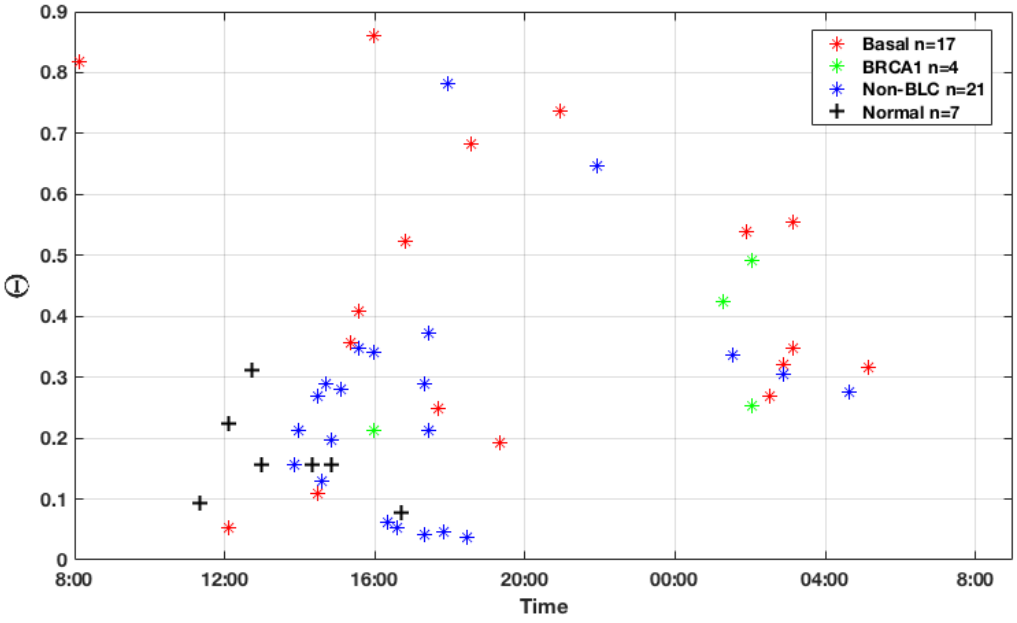


Figure 6.7: **A plot of estimated time vs Θ for the Richardson data**. All healthy data is estimated to have been taken between 11am and 5pm. The cancer data is scattered across all times.

Both figures show that the tumours are more likely to have a dysfunctional clock than the healthy samples. However, there are only 7 healthy samples, whose median Θ is 0.155, which is the upper limit of what the Bjarnason data suggests we should label a working clock. There is no mention in the paper of how the healthy breast tissue was attained, but it is likely that this tissue is from the resection boundary of tumours. Tissue from tumour micro-environment being labelled as healthy might explain why the Θ values for the normal samples are not as low as we would expect. Additionally, this may be because we are comparing oral mucosa to breast tissue. We showed in chapter 3 that the core clock genes of mice are synchronised across organs, and we use this to back up the assumption that we can compare oral muscosa and breast tissue.

Despite the limitations of the data, there is a significant difference in Θ distribution between groups, suggesting that the tumours are more likely to have dysfunctional clocks.

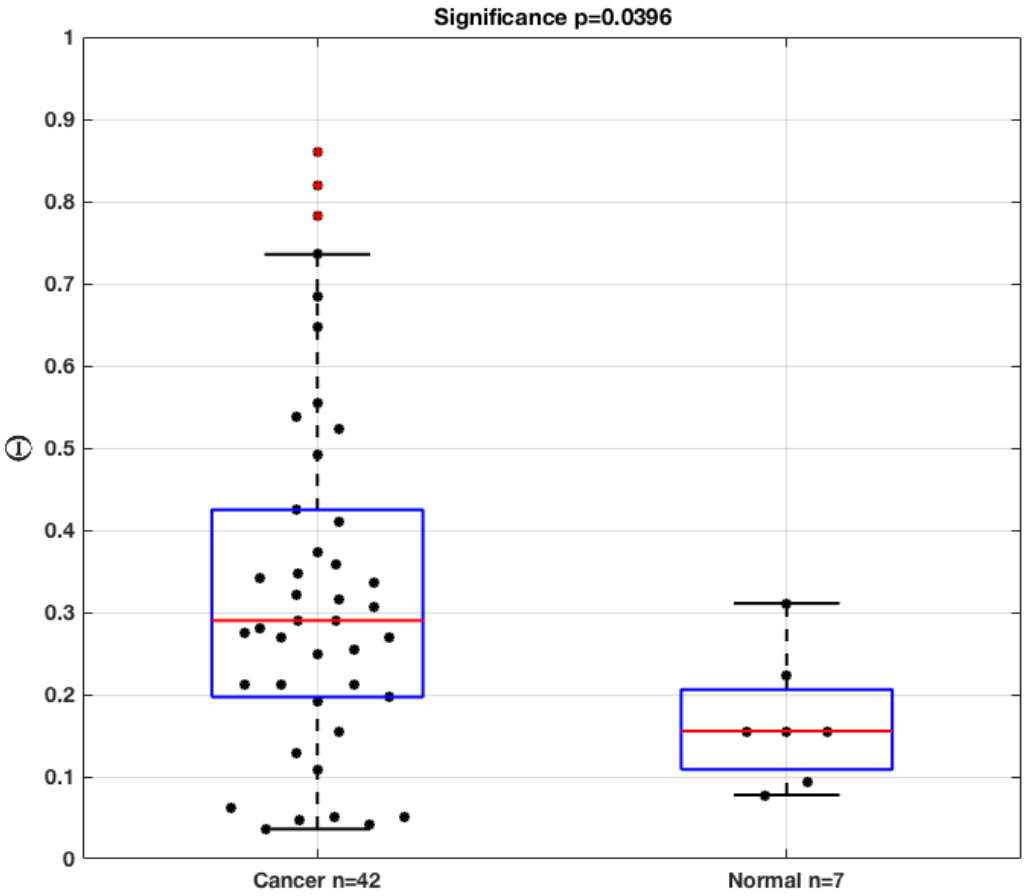


Figure 6.8: **The distribution of Θ values calculated by Time-Teller for the Richardson data.** The normal samples have a significantly lower median than the cancer samples, with Wilcoxon test statistic $p = 0.035$.

6.3 Circadian clock dysfunction as a novel prognostic factor

In this section, circadian clock dysfunction is incorporated into a study on cancer characteristics, using Θ. The data we will use as a case study is that of the REMAGUS clinical trial.

REMAGUS clinical trial

The REMAGUS multicenter randomised phase II clinical trial was carried out between 2004-2007, aiming to assess the response of primary breast cancer to different protocols of neoadjuvant chemotherapy, according to tumour hormonal receptor status and HER2 expression. Survival data was gathered up to 10 years after the completion of the study. Giacchetti *et al.* [171], Cremoux *et al.* [113], Valet *et al.* [173] and Pierga *et al.* [174] report the outcomes and clinical results of the trial. None of these results in these studies are specifically circadian clock related.

Briefly, excluding specific drug protocol results, the major findings reported from published works are:

- A set of 31 genes differentiate pCR or no pCR samples. Many of these were associated with estrogen receptor related genes.
- Negative hormonal receptor status and limited tumour size predicted for pCR,
- Triple negative breast cancers experience the highest pCR rate of 30%

6.3.1 Distribution of Θ for REMAGUS data

The transcriptome of 226 patients of the REMAGUS trial were analysed using Affymetrix U133 2.0 Arrays and is available on GEO under accession number GSE26639. Prognostic and survival data up to 10 years after the trial ended is available from the supplementary information of [171].

Time-Teller calculated time of samples and the Θ clock function metric for all 226 tumour transcriptome samples. The distribution of Θ values for the REMAGUS data are shown in the histogram in figure 6.9, with the corresponding Bjarnason data Θ values overlaid. The maximum Θ for the Bjarnason data is 0.155, but the majority of the data has $\Theta < 0.1$. Around half of the REMAGUS data has Θ values in the same range as the Bjarnason data.

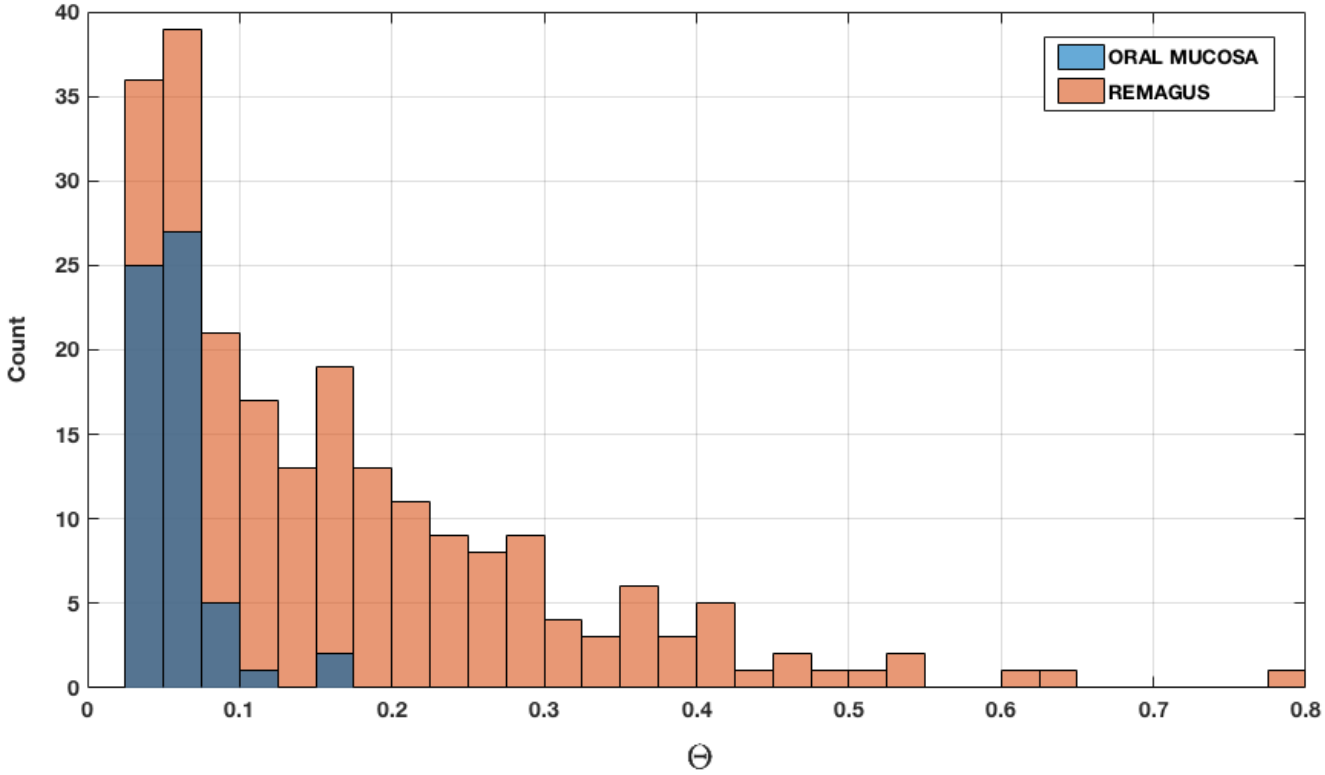


Figure 6.9: **Histogram showing the distribution of Θ values for oral mucosa and breast cancer samples.** The Θ distribution of the 60 healthy oral mucosa samples from the Bjarnason data is in blue, and the 226 REMAGUS tumour samples distributions are shown in orange. Histograms are overlaid, not stacked.

Estimated time vs real time

The times that the tumours were biopsied are available for 108 of the samples. All of the tumour biopsies were performed between 8:45 and 17:45. As we are now using cancer data, we do not expect that we can accurately predict the timings of all of these samples. The real versus estimated times from Time-Teller are plotted in figure 6.10 and the data points are coloured by Θ value. There is a clear cluster around midday on both axes where most of the estimates lie. Time estimates for samples with a functioning clock have a mean error of < 3 hours, but there is no correlation in this 9 hour time frame. All estimates outside of this "daytime cluster" have $\Theta \geq 0.1$, indicating the clock time was not estimated accurately because the circadian clock is dysfunctional in those tumour samples.

The absolute error of estimation versus the Θ values for the 108 breast cancer samples is plotted in figure 6.11. There is no correlation between estimation error and Θ. A correlation is not expected because, as Θ grows larger, the estimation for the time begins to be meaningless. The mean absolute error for time estimations for data with $\Theta < 0.1$ is 2hrs 45 mins. The mean absolute error for time estimations for data with $\Theta > 0.1$ is

3hrs 45 mins. Samples with an error < 3.5 (n=70) hours have a median Θ of 0.117, and samples with an error > 3.5(n=38) hours have a median Θ of 0.145. Error estimations > 7 (n=12) hours have a median Θ of 0.249. Hence, even though there is no simple correlation, there is still some significance to be found. All samples with > 7 hours error have $\Theta > 0.14$ and none of the estimates that have greater than 7 hours error have Θ values that would cause us to accept the estimation.

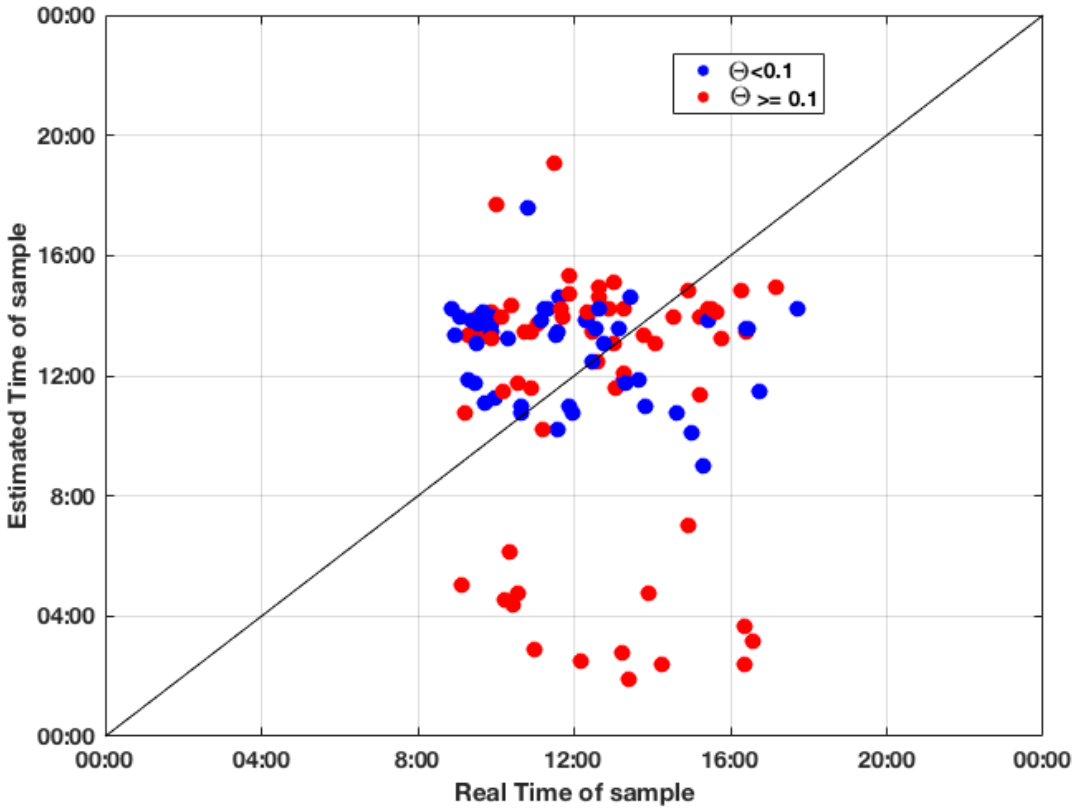


Figure 6.10: **Scatter plot showing the real vs estimated times for 108 breast cancer samples.** The markers are coloured by Θ, where $\Theta < 0.1$ is blue and $\Theta \geq 0.1$ is in red. There is no correlation in estimations over the 9 hour time frame or any group, but the mean times of estimations are very similar, around 1pm. All samples with > 7 hours error have $\Theta \geq 0.1$

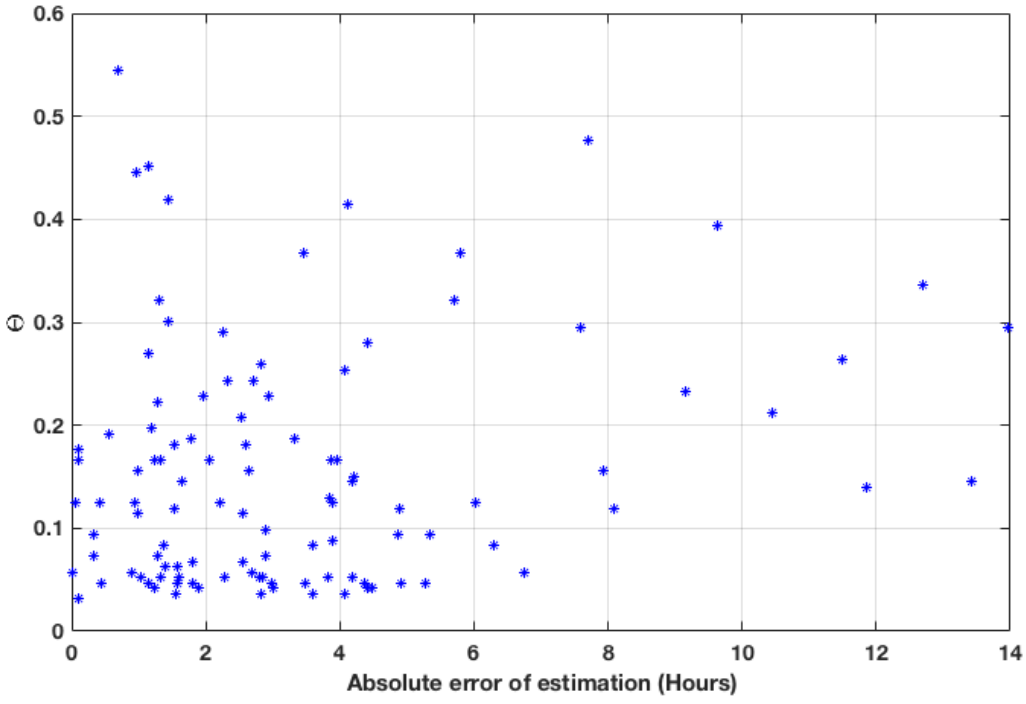


Figure 6.11: **Scatter plot showing the absolute error of estimation versus the Θ values for 108 breast cancer samples.** All samples with > 7 hours error have $\Theta \geq 0.1$

6.3.2 Clock function relation to prognostic markers

Θ values representing clock function were calculated for all 226 patient tumour transcriptome samples. Figure 6.12 shows the differences in distribution of Θ values for some of the prognostic factors for breast cancer. The Wilcoxon Rank sum test was used to calculate statistical significance in medians. In cases with more than 2 factors, factors were grouped.

Estrogen and progesterone positive tumours show a significantly lower Θ median value than ER- and PR- tumours (p=0.006 and p=0.007 respectively). HER2 positive tumours show an almost significantly higher median (p=0.063). The most significant difference is shown by Triple negative tumours, where the median value for the 52 triple negative tumours is around ~ 0.2, and the median for 171 non-triple negative tumours is ~ 0.1, and the significance is $p = 0.0005$.

The grade of the tumours also shows an increasing Θ as grade increases. The significance for a higher Θ median for grade 3 tumours, than combined grade 1&2 tumours, is p=0.014.

The majority of tumours had nodal status 0 or 1, where no difference between Θ medians was found. The 14 samples with nodal status 2 or 3 have significantly higher Θ values than the 200 samples with 0 or 1 nodal status, with significance p=0.040.

Tumour size does not show an increasing trend in Θ values as tumour size increases. However, as stated above, a tumour size of 4 does not measure actual size of tumour as a tumour size of 2 or 3 does, so this is not a discouraging result. A tumour size of 3 has a higher Θ median than a tumour size of 2. When 3 & 4 are grouped, the significance for difference in means is p=0.014.

pCR

Pathological complete response (pCR) is determined after treatment. In this case, after 8 courses of neoadjuvant chemotherapy, all the patients underwent breast surgery. pCR was defined as absence of residual invasive cancer cells in the breast and axillary lymph nodes (grade 1 and 2 of Chevallier's classification). The boxplots for pCR are boxed in figure 6.12 as these measurements are not determined at the same time as the others in the figure. Pathological Complete Response occurs if there is no sign of cancer after treatment. In the REMAGUS dataset, only 35 patients achieve pCR.

The Θ values for the pCR set have a statistically significant higher median than the non pCR set. At first thought this result may appear to contradict the highly significant result shown above that, in general, more aggressive tumours have more dysfunctional clocks. Intuitively, it may seem like less aggressive tumours (so tumours with better functioning clocks) are more likely to achieve pCR. However, as was discussed in section 6.0.4 we do have evidence to expect that tumours with dysfunctional circadian clocks will

respond better to chemotherapy. If it truly is the case that tumours with dysfunctional clocks are more susceptible to chemotherapy, as this evidence indicates, then the use of the circadian clock dysfunction metric Θ for chronotherapy strategies is a very promising path to follow.

Of the 35 patients that achieved pCR, 17 of these had TN tumours. It was reported in Giacchetti *et al.* [171] that tumours are more likely to achieve pCR if they are TN. We have shown that samples with a dysfunctional circadian clock are more likely to be TN, and more likely to achieve pCR. We hypothesise that tumours that were TN, and achieved pCR had the most dysfunctional clocks as they responded best to treatment. This analysis is shown in figure 6.13, and does show that the median Θ for the group with pCR and TN tumours is highest. However, the group sizes are very small and the Wilcoxon rank sum test does not report significance between the pCR, TN group and the pCR, not TN group ($p = 0.26$).

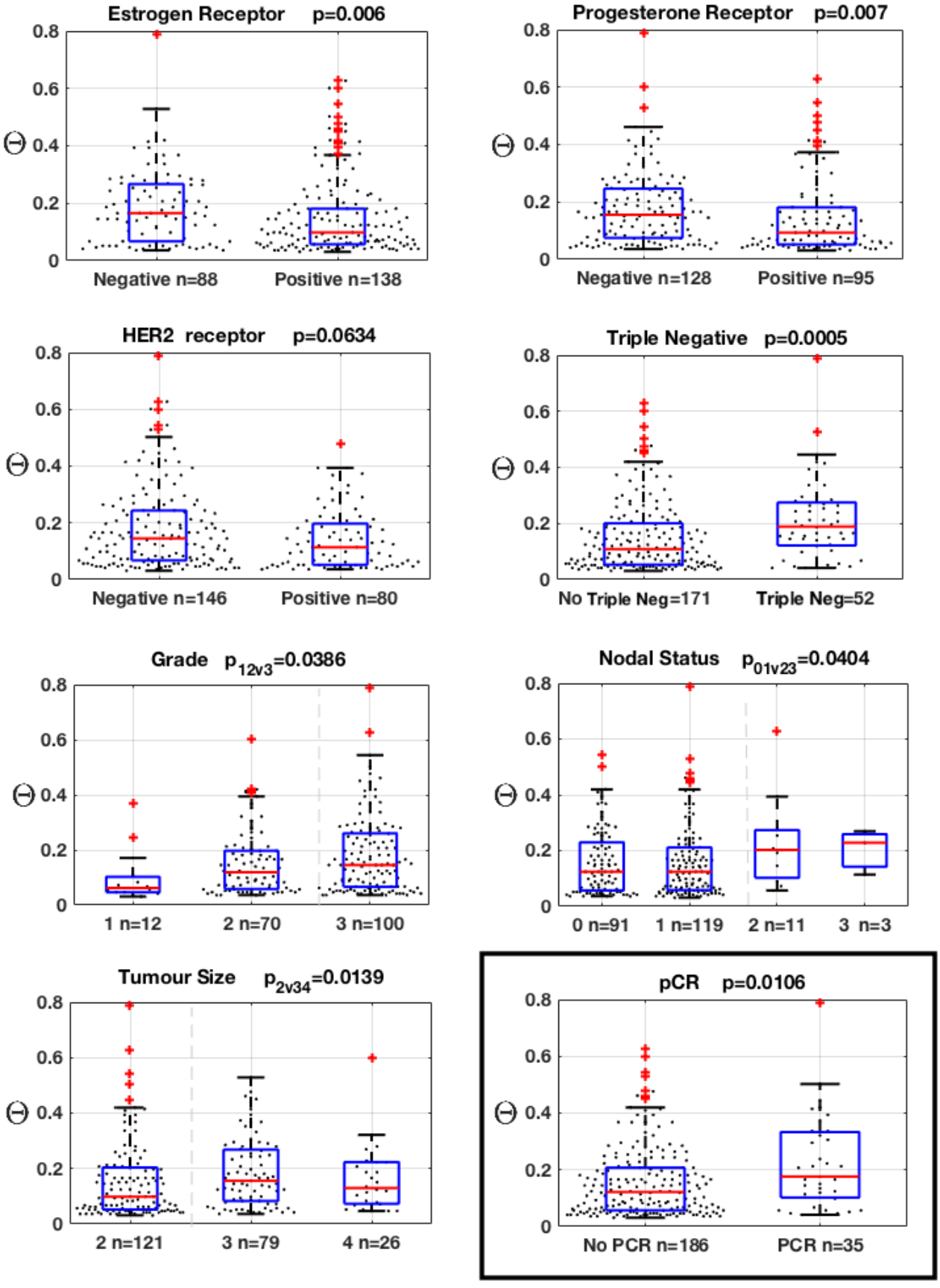


Figure 6.12: **Boxplots showing distributions of Θ values for Estrogen receptors, Progesterone receptors, HER2 receptors, Triple Negative status, Grade, Nodal Status, Tumour size, and pCR.**

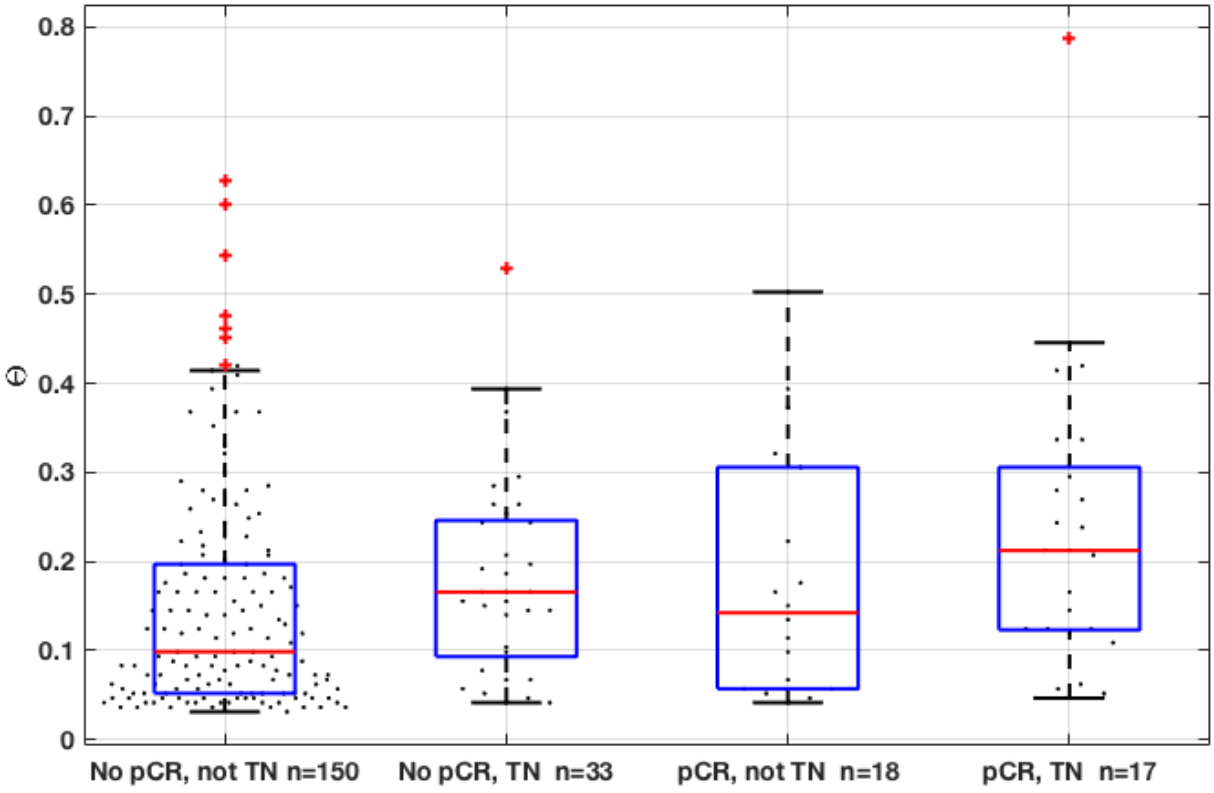


Figure 6.13: **Boxplots showing distributions of Θ values for tumours that did and did not achieve pCR after treatment sorted by tumours that were and were not triple negative at diagnosis.** Patients with the highest Θ values were those with TN tumours that achieved pCR, although numbers in the groups are too small to generate significant evidence.

6.3.3 Clock function relation to survival

Survival data is available for 224 patients for 10 years after the clinical trial. 3 types of survival are provided with this data; overall survival, disease free survival, and event free survival. Disease free survival reports whether the cancer returned, and the date this was reported. Event free survival can relate to the return of the disease, or any complications that occurred relating to the cancer, for example pain. Here we look at survival differences in relation to clock dysfunction. The relationship between decreased clock function and increased tumour aggression, as well as the relationship between decreased clock function and increased pCR were shown above. It should be noted here that the clock dysfunction was measured before treatment, and different treatment courses were given in this clinical trial, where outcomes were measured after treatment.

Overall survival

We set a Θ cut-off to be the maximum Θ measured for the healthy data $\Theta = 0.155$, as was shown in figure 6.9. A Kaplan-Meier survival analysis is plotted in figure 6.14 for overall survival separating the population with $\Theta \leq 0.155$ and $\Theta > 0.155$, which separates the 224 patients into groups of 128 and 96 respectively. The significance value for this analysis is tested with a *logrank* test in MATLAB (see appendix C), and provides a significance value of $p = 0.026$.

There is a clear separation according to measure of clock function, where the analysis suggests that 82% of patients with "good functioning" clocks survive for 10 years or more, whereas only 61% of patients with a "bad functioning" clock survive past 10 years.

This result is comparable to the logrank test results reported in Giacchetti *et al.* [171] for significant increases in disease free survival for PR+ status ($p = 0.02$) and HER+ ($p = 0.03$) (as clock function and receptor status are correlated)

If the "dysfunctional clock" data is split into 2 groups, we observe interesting and complex behaviour. Figure 6.15 shows samples with $\Theta > 0.3$ in green, having better overall survival than samples with $0.155 < \Theta < 0.3$. Excluding the "very bad" clocks from the analyses increases the significance for differences in survival between the good and bad clock. This might provide further evidence that tumours with dysfunctional clocks respond better to chemotherapy.

Disease and event free survival

Complex results were found for the link between clock function and disease free survival or event free survival. The Kaplan-Meier plots are shown in figures 6.16 -6.19. No significant results were found, but again, the group of 31 individuals with the worst clocks ($\Theta > 0.3$) showed increased DFS and EFS until the end of the 10 year trial window, even more so than the groups with good clocks.

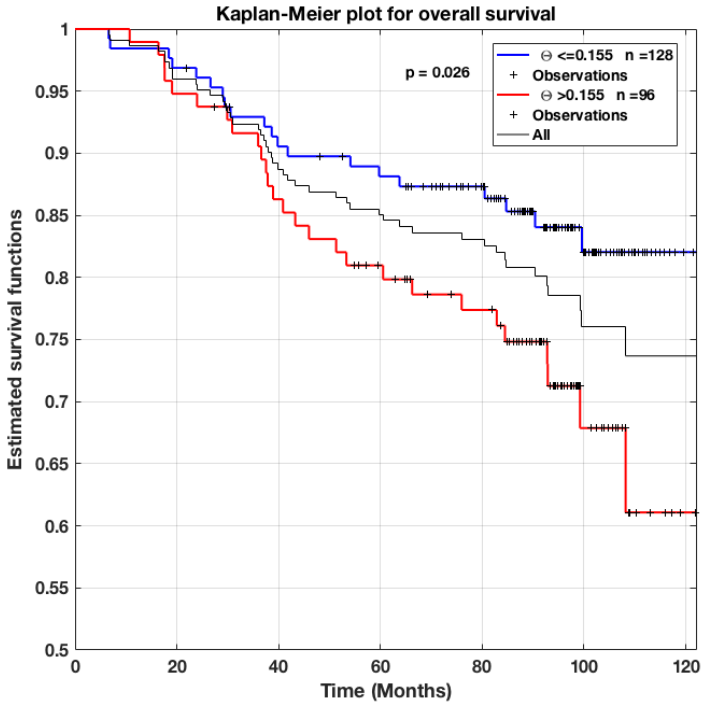


Figure 6.14: **Kaplan-Meier survival plot showing differences in survival for patients with "good clocks" an "bad clocks".** The log rank test produces a significance measure of $p = 0.026$, where samples are separated by $\Theta \leq 0.155$ (good clock, blue), and $\Theta > 0.155$ (bad clock, red).

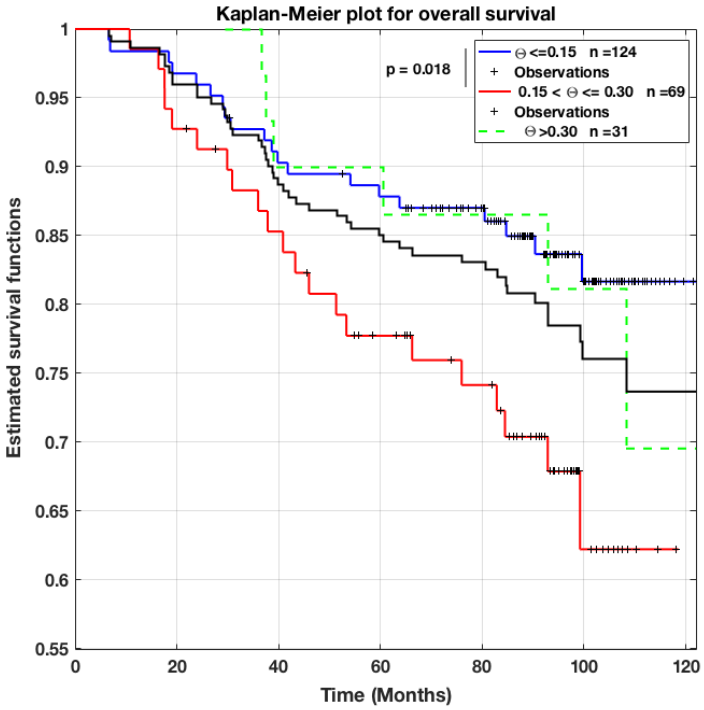


Figure 6.15: **Kaplan-Meier survival plot showing differences in survival for patients with "good clocks", "bad clocks" and "very bad" clocks.** The log rank test produces a significance measure of $p = 0.018$, where samples are separated by $\Theta \leq 0.155$ (good clock, blue), and $0.3 > \Theta > 0.155$ (bad clock, red). Green dashed line shows that "very bad" clocks have increased survival until 7-8 years after treatment.

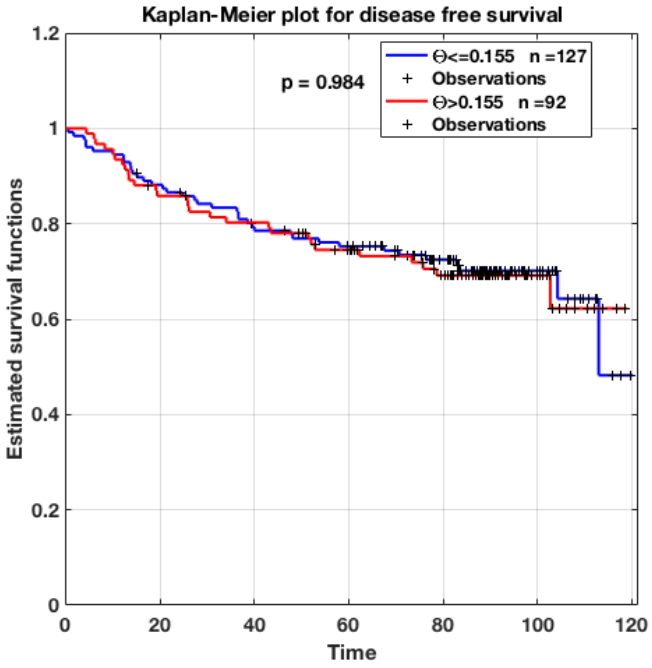


Figure 6.16: **Kaplan-Meier survival plot showing differences in dfs for patients with "good clocks" and "bad clocks".** The log rank test produces a significance measure of $p = 0.984$, where samples are separated by $\Theta \leq 0.155$ (good clock, blue), and $\Theta > 0.155$ (bad clock, red).

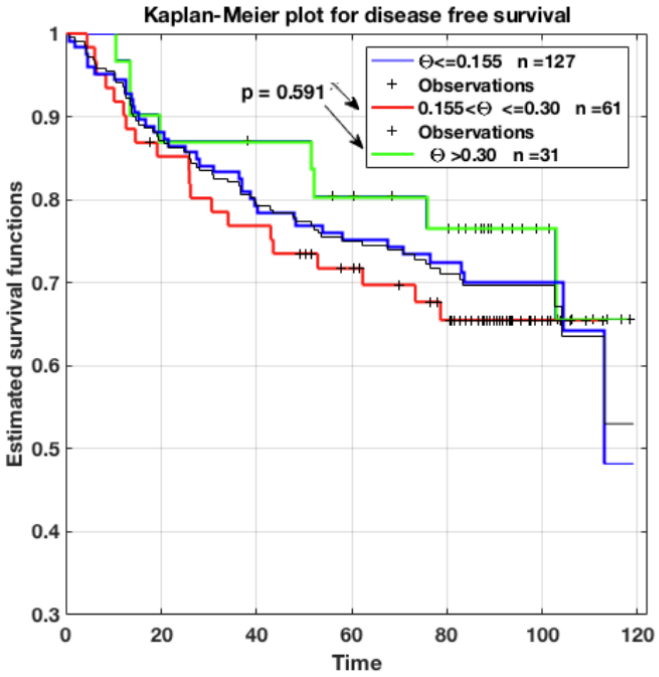


Figure 6.17: **Kaplan-Meier survival plot showing differences in dfs for patients with "good clocks", "bad clocks" and "very bad" clocks.** The log rank test produces a significance measure of $p = 0.591$, where samples are separated by $\Theta > 0.155$ (bad clock, red), and $0.3 > \Theta$ (very bad clock, green). Blue line shows that "good" clocks have average survival.

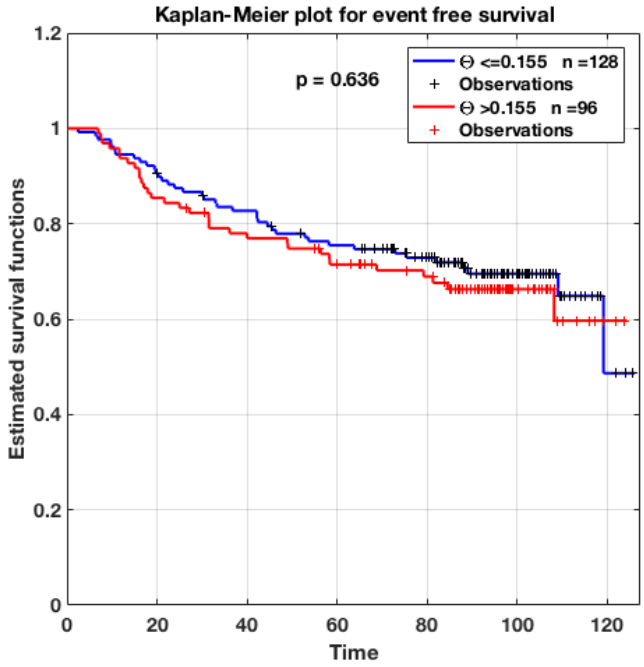


Figure 6.18: **Kaplan-Meier survival plot showing differences in efs for patients with "good clocks" and "bad clocks".** The log rank test produces a significance measure of $p = 0.636$, where samples are separated by $\Theta \leq 0.155$ (good clock, blue), and $\Theta > 0.155$ (bad clock, red).

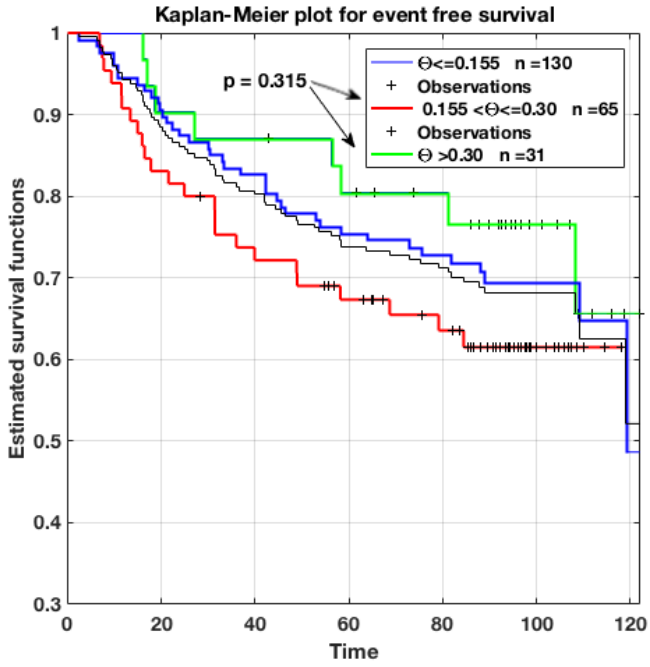


Figure 6.19: **Kaplan-Meier survival plot showing differences in EFS for patients with "good clocks", "bad clocks" and "very bad" clocks.** The log rank test produces a significance measure of $p = 0.315$, where samples are separated by $\Theta > 0.155$ (bad clock, red), and $0.3 > \Theta$ (very bad clock, green). Blue line shows that "good" clocks have average survival.

6.3.4 Summary of section

This analysis has showed how the metric for clock dysfunction Θ could be used as a novel prognostic factor. Further work could include a more in depth multivariate clinical statistical analysis, as was done in the REMAGUS follow up paper [171], but including Θ as another factor.

Some evidence in this section suggested that treatments are more effective for clock dysfunctional tumours, as pCR was more likely to occur in patients with clock dysfunctional tumours, and disease and event free survival appeared to be better until the end of the 10 year window. It could be that the results that we are seeing are a combination of clock dysfunction correlating with more aggressive tumours and worse outcomes, with the effect of chemotherapy working better for tumours with dysfunctional clocks.

This proves very difficult to explore in a robust manner, with only 35 patients achieving pCR, and only 31 individuals in this increased short term survival group. Many factors in this type of analysis are highly correlated, so robust statistical multivariate analysis, and the use of more data sets is necessary.

6.4 Comparison of all human data

We have shown so far that the distribution of the Θ metric shows significant differences when individual datasets are compared with the healthy oral mucosa samples from the Bjarnason data, or by groupings within each data set.

We now combine all non-cancer data sets: all smoker/non smoker data (n=79), normal and dysplasic samples from the Feng data (n=62), normal samples from the Richardson (n=7), and the UK samples from UK/Sri Lankan oral mucosa data (n=5).

We combine all the cancer datasets: all REMAGUS data (n=226), all the Richardson cancer data (n=55), and the Feng cancer data (n=62).

15 probes were used to train Time-Teller, and it was used to find the Θ metrics for the non-cancer and cancer datasets above, treating every sample as independent. The results of this are shown in the cumulative distribution plot in figure 6.20, where the significance for the difference in distribution is found by the two-sample Kolmogorov-Smirnoff test to be very significant. For these datasets, the Θ metric is significantly higher in cancer datasets, providing evidence that the level of circadian clock dysfunction in cancer is significantly higher than in non-cancer data.

When this plot is separated into the individual datasets, where the Θ distribution of the Bjarnason training data plotted too, the shift to higher Θ values is clear in the cancer datasets. Here, the small healthy datasets (Richardon breast healthy, Feng healthy & dysplasic, and UK oral mucosa data) are pooled to an "other healthy" data set.

The REMAGUS data Θ distribution is similar to that of the Θ distribution of the

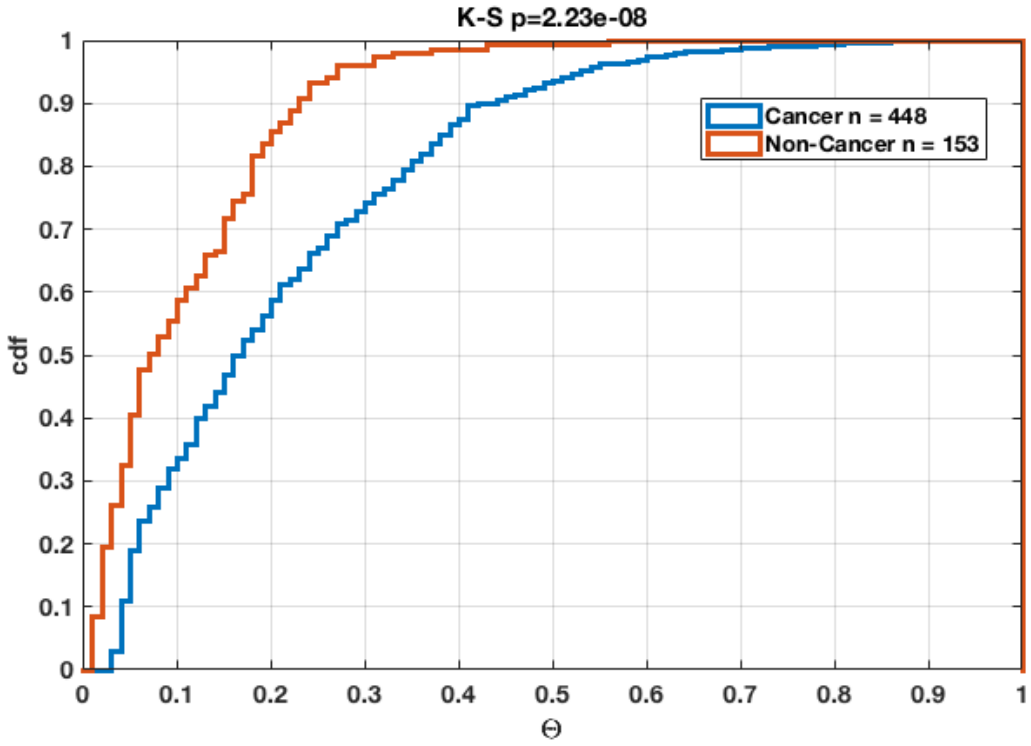


Figure 6.20: **Cumulative plot of Θ values for all cancer and non-cancer samples.** The two-sample Kolmogorov-Smirnov test for different distributions is very significant.

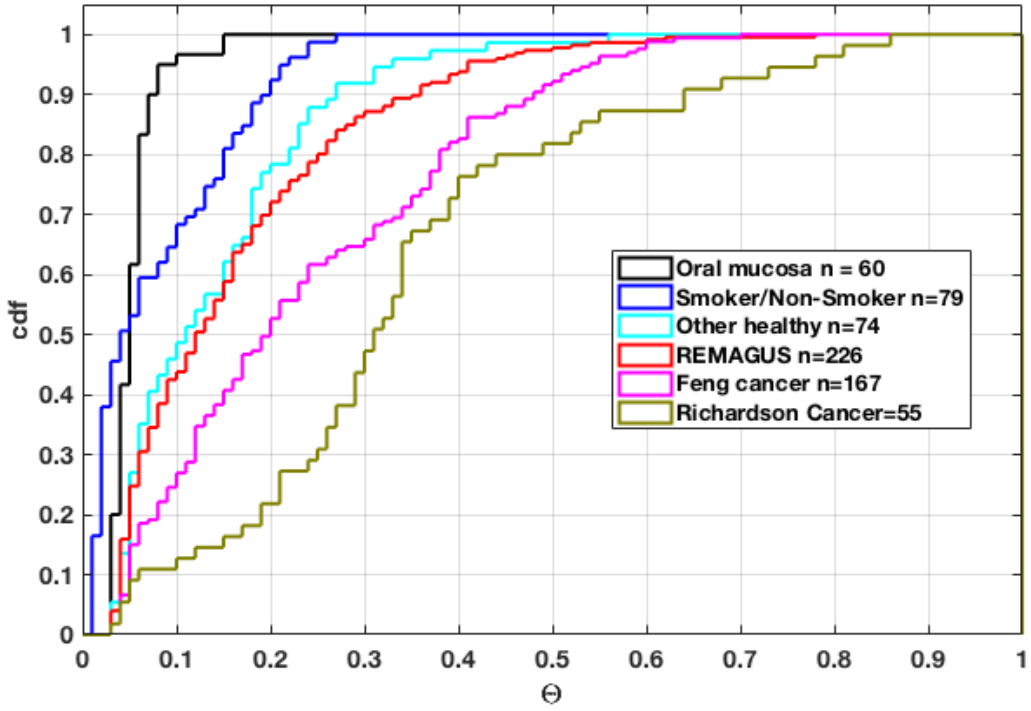


Figure 6.21: **Cumulative plot of Θ values for multiple human data sets.** The black plot shows the Θ distribution of the training set.

combined healthy data from the Feng dataset and Richardson breast tissue data set, and the UK samples from SriUK data.

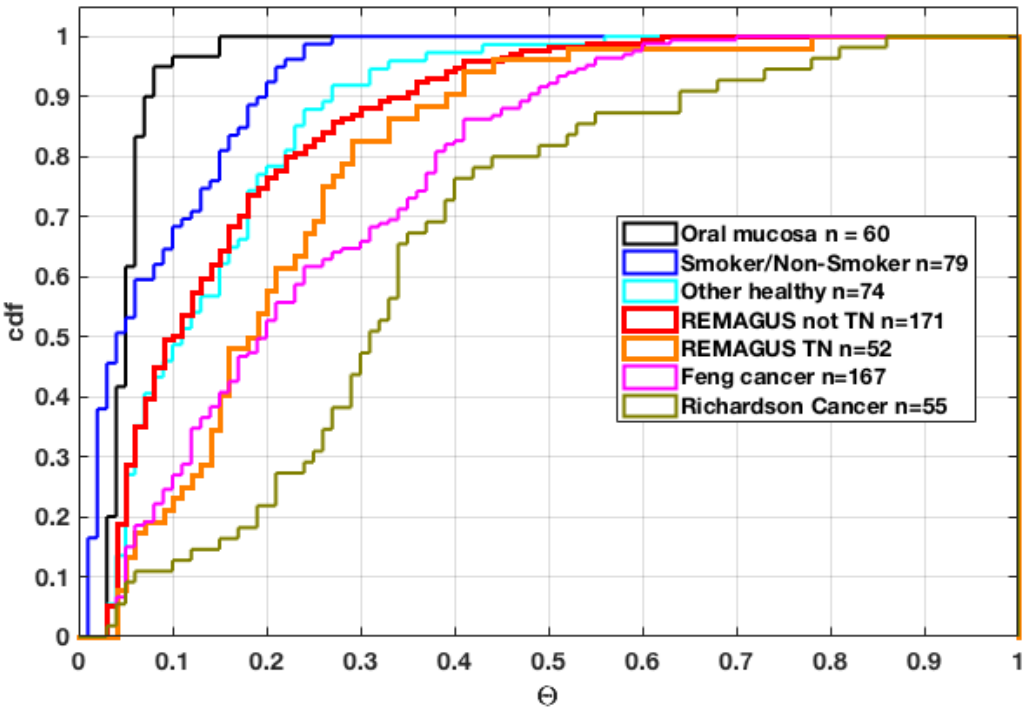


Figure 6.22: **Cumulative plot of Θ values for multiple human data sets, with the REMAGUS data separated by TN status.** The TN data shows a shift to the "dysfunctional clock" cancer region of the cumulative plot and the non TN data shifts to the "functioning clock" region with the healthy data.

Although there is a clear pattern in these plots that cancer data generally has higher Θ values, the divide is not a clear one. The REMAGUS data Θ distribution is very close to that of some of the pooled non-cancer data. This could be due to problems with the healthy data, or it could be due to many of the tumours in the REMAGUS data having functional clocks. For example, if the REMAGUS data were split into TN and non-TN samples, figure 6.22 shows that the TN data is clearly within the "cancer" region of the cumulative plot.

Many factors will be adding variation to this data; genetic differences, environmental differences, but there is still an obvious separation between cancer and non cancer data.

6.5 Summary of chapter

This chapter has presented an overview for some of the evidence towards circadian clock dysfunction in cancer. We discussed evidence that circadian clock dysfunction promotes tumourigenesis, and also that many tumours have dysfunctional clocks. We used 2 datasets that contained both cancer and healthy data, to show that clock function is significantly worse in cancer samples than in normal samples. The REMAGUS dataset was used to correlate clock function with cancer prognostic markers. There was significant evidence that clock function is generally worse for more aggressive tumours. There was also significant evidence that chances of achieving pCR are higher for tumours with more dysfunctional circadian clocks. This agrees with the evidence previously discussed, that chemotherapy is more effective when the circadian clock is dysfunctional.

We have presented how incorporating "clock function" as a new prognostic factor in tumour classification may help to provide new insights into the best courses of treatments.

Chapter 7

Discussion

7.1 Summary of thesis

The methods in this PhD project were developed towards the aim of developing a method to determine whether the circadian clock gene expression was different to "normal" circadian gene expression, in a single transcriptome from a tumour biopsy. Each chapter built up to this aim.

Through mathematical models, chapter 2 explored the robustness of the circadian clock in order to validate an assumption used in the Time-Teller model, that we can expect robust and reproducible rhythms in a healthy population. Using the Relogio ODE model and an equivalent stochastic model, the phase sensitivity to parameter perturbation was measured for both the deterministic and stochastic Relogio model. The results showed that both models are very robust to parameter perturbation, and only a small subset of parameter changes would have significant effects on the behaviour of the model. We provided evidence that stochastic circadian gene expression has the same characteristics of robustness as the deterministic characteristics that have been previously reported [75, 76, 77]. This *in silico* evidence allowed us to understand the reasoning behind an assumption that we use in the Time-Teller model; that there is a robust and synchronised behaviour of circadian clocks amongst independent samples or individuals.

In chapter 3, we explored this in real data. The use of the SVD was presented as a simple and computationally efficient way of finding rhythms by exploring synchronicity. The use of multiple rhythmicity algorithms in parallel with the SVD allowed interesting outliers to be identified, and allowed higher confidence in the final set of rhythmic and synchronised genes. We verified that the core clock genes in mice were very synchronised in expression profile for different tissues, and that the core clock genes in humans were synchronised across 10 individuals. It was shown that for human individuals, the only

genes that have synchronised timecourse expression are the circadian clock genes, and this was mostly true for mouse tissues. The identified rhythmic and synchronised genes of the mouse and human data were used as training genes for the mouse and human Time-Teller algorithms. These genes were used to define the expected behaviour of a healthy circadian clock in each Time-Teller model in chapter 4.

Chapter 4 presented the currently published time-telling algorithms, and highlighted that none of them can truly tell the time from one single independent sample of unknown time. The Time-Teller algorithm was presented as a method that can predict the time of one independent transcriptome, and the advantages of using a local PC approach were shown. Time-Teller was validated for *in silico* data with different levels of noise, and was shown to be very accurate when using a leave-one-organ-out approach for mouse data, and a leave-one-individual-out approach for the human data. Although we have shown that the circadian clock time (body time) is generally synchronised to time-of-day, the natural variation in body time was also shown. For example, Male15 was consistently phase delayed and female18 was consistently phase advanced, in comparison to the other individuals.

Chapter 5 presented the metric Θ as a measure of clock dysfunction, which was calculated using the likelihood functions produced by Time-Teller. The threshold for where Θ defines a "functional clock" was defined by each data set used to train Time-Teller. The *in silico* Time-Teller was used to show how Θ is correlated with the level of an *in silico* KO, and how Θ is not sensitive to a change of the hyperparameters η and ϵ.

The mouse Time-Teller was able to accurately calculate the timings of the samples in the Martletot dataset, with "functional clock" Θ values. Time-Teller was able to detect KO samples with dysfunction clocks, calculating $\Theta \approx 1$ for samples from full KO samples, and $\Theta < 0.155$ for WT samples. Sleep deprived mice were shown to have less functional clocks than their normal sleep schedule counterparts.

The human Time-Teller was able to provide realistic estimations with "functional clock" Θ values for 3 independent datasets. Problems were found with 3 samples from Sri Lankan individuals, but the lack of information about the individuals in the original study meant we can only speculate as to why this is. Data from autopsy samples were used to validate that Time-Teller could estimate that samples were taken during the night, with good confidence (i.e. low Θ values).

Chapter 6 presented an overview for some of the evidence towards circadian clock dysfunction in cancer. We discussed evidence that circadian clock dysfunction promotes tumourigenesis, and also that many tumours have dysfunctional clocks. We used 2 datasets that contained both cancer and healthy data, to show that clock function is significantly

worse in cancer samples than in normal samples. The REMAGUS dataset was used to correlate clock function with cancer prognostic markers. There was significant evidence that clock function is generally worse for more aggressive tumours. Interestingly, there was also significant evidence that chances of achieving pCR are higher for tumours with more dysfunctional circadian clocks. This agrees with the evidence previously discussed, that chemotherapy is more effective when the circadian clock is dysfunctional.

Survival analysis found that patients with tumours with more functional clocks had an 82% chance of living to 10 years, but patients with less functional clocks had a 61% chance of living to 10 years. It was also found that the tumours with the worst clocks do not have the worst survival. This suggested that there was a lot of complexity underlying the data: that bad clock function is related to more aggressive tumours and hence related to more deaths, but also that clock function is related to better chances of effective treatment.

7.2 Discussion of Time-Teller

7.2.1 Clinical uses and advantages

Time-Teller would be clinically useful as it is able to estimate the phase of a gene such as Bmal1 or Rev-Erbα (i.e. body time), from one transcriptome sample (if real clock time is known). This could be used in the clinic to inform on optimal timings for personalised chronotherapy regimes.

Time-Teller can also inform clinical strategies as to if a tumour has a functioning circadian clock from just one sample. This could be used in parallel to the standard cancer prognostic markers, in order to decide on optimum treatment regimes.

Time-Teller can also provide evidence that the circadian clock is dysfunctional in tumours, and add to the evidence that CYCLOPS [13] and ΔCCD [15] have begun to produce. Time-Teller can produce this evidence for individual samples, and small data sets (if a training set is available), where CYCLOPS and ΔCCD need substantially sized datasets to produce results.

7.2.2 Limitations

Time-Teller is dependent on the use of data from the same technology and GeneChips for the training and test sets. Time-Teller measures the transcriptome's time fingerprint by the comparative ordering of each gene expression value, so is very sensitive to absolute levels of expression. Microarrays use probes whose expression values are a function of transcript number, transcript binding affinity and other things, where different designs of probes can change these weightings. Using the same GeneChip and technology removes

some of the dangers of this sensitivity, but one must be mindful of other sources of experimental and biological variation when comparing independent datasets.

There was some evidence in this thesis that there are differences in circadian transcriptome for individuals in Sri Lanka, to those in Canada, but no obvious differences to individuals in the UK or US. Although these were only 3 samples, and they could have been compromised in some way, it may be true that there are differences in the circadian transcriptome for humans that live in different longitudes. There has not been any study (that we know of) into this, but as there is such a significant difference in length of daytime light, temperature, and perhaps even working cultures, it might be possible that the circadian transcriptome is different somehow for humans at different longitudes. If this were true, a comparable training set for Time-Teller would have to be used in order to produce proper results from some data sets.

Time-Teller cannot measure if a circadian clock period is changed, and assumes that all clocks have a 24 hour period. If a sample had a functioning clock, but with a non-24 hour period, Time-Teller would produce Θ metrics that indicate the functioning of the clock, but the time estimation would not be accurate.

We showed in chapter 3 that the core clock genes of mice are synchronised across organs, and we use this to back up the assumption that we can compare human oral mucosa and human breast tissue and expect them to have the same set of core clock genes. We chose genes for the training set of the human Time-Teller that are synchronised amongst 10 individual's oral mucosa, but it is possible that these genes are not organ wide synchronised.

The Bjarnason data arose from healthy human subjects that had been tracked so that the lab knew they had a good circadian rhythm. All of the independent human data has no such information on the individuals the samples are from. Additionally, the Zhang mouse dataset originated from a study designed to synchronise the mice. It could still be possible that chronodysruption effects tissues differently, so that a comparison between synchronised healthy human oral mucosa is not as easily comparable to other tissues.

7.2.3 Future work

More data

There are many other data sets that could have been used to test the power of Time-Teller. Focussing only on human data, there are thousands of data sets posted to GEO that have been generated with the Affymetrix HG U133 2.0 arrays [1]. One must read studies and

[1] https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL571

identify data that is appropriate, for example identifying studies whose samples were taken from live tissue and frozen. The most robust datasets for validation are those that contain cancer data with corresponding healthy controls. Other ideal datasets would be those that are similar to the REMAGUS data, where cancer prognostic markers could be correlated with Θ.

Loadings of local PCs

As Time-Teller uses a combined local PC method, assessing the loadings of the principle components is not straight forward. Further work to be done to really understand Time-Teller, and the underlying dynamics of the data, would be to somehow analyse the changing weightings of the local PCs over time.

Possible Bias for particular times when the clock is dysfunctional

Figures 6.10 and 6.7 suggest that when the clock is dysfunctional, there is a bias to the MLE, suggesting a slight bias in the model that probably arises from the different covariances used at different times. This is unlikely to affect the results, as we do not care what the MLEs are when Θ is large. However, it would be be interesting to design a way to test this is a real bias, and what effect it does have.

RNA-seq data

RNA-seq more accurately measures transcript numbers, so may be a better technology to use for future models. However, there may still be some incompatibilities between different technologies. Time-Teller can be used for RNA-seq data with exactly the same methods as presented in this thesis for microarray data.

If and when multiple observations of timecourse RNA-seq data are available, Time-Teller will be able to be re-trained and used to assess clock time and function from single time point assays of RNA-seq data. There could be some issues to resolve concerning compatible RNA-seq experiments with slightly different designs (e.g. read depth), but this should generally be straight forward.

It it possible now to design a mouse RNA-seq Time-Teller using only the 6 hour resolution data in the Zhang data set [73]. This would result in a 4 knot periodic spline in 3D space when creating the distribution space when building Time-Teller, and it is possible that this could accurately represent the intermediate times. It would also be reasonable to use the microarray data to make estimates for the RNA-seq intermediate time data points.

Possible formalisation to Gaussian process regression

Time-Teller has similarities in its structure to some methods in Gaussian process regression. Homoscedastic Gaussian process regression with a specially designed periodic kernel could potentially summarise this Time-Teller model, but the practical aspects of this were not achieved for this thesis.

Optimising the microarray probes used in the training sets

The probes chosen for the training set were those that were most rhythmic and synchronised amongst observations. Attempts were made initially to optimise the probes chosen to be part of Time-Teller, by minimising the error of the real versus estimated time. However, this resulted in an over-fitting of the model to the training data, and the optimised set of probes for the training data were not necessarily the optimum set when comparing to independent datasets.

It was found when comparing the human Time-Teller to independent data sets, that one of the Per1 probes had much weaker signal for the independent data sets, and so was not used. Cross experiment comparisons will always have some issues, but it may be possible to understand more about the microarray probes themselves, in order to find the optimum, most robust training set.

7.3 Novel findings

Ciart

There was already some evidence that Ciart (or Chrono/Gm129) has an important role in the core circadian mechanism. This thesis has shown that Ciart has rhythmic and synchronised behaviour in both mice and humans, in a way comparable to genes that are only considered to be core circadian clock genes. The tight regulation and synchronicity of Ciart indicates that it is a major part of the central timing mechanism. Ciart was not identified in many past studies due to its low amplitude rhythms, where rhythmicity analysis was performed on non-logged, non-normalised data.

Male versus female differences

Although no male/female differences were discussed in this thesis, there is an interesting result that arose from the Θ values of the leave-one-out approach in figure 5.20. Although there are very few data points to be able to draw a strong conclusion from, the female Θs do appear to be higher than the male's. The Wilcoxon test statistic for males having lower Θs is 0.037, which is a significant difference, indicating that females do have more variable

clocks. Further analysis of female/male differences are being investigated by Bjarnason *et al.*. This indicates that Time-Teller is also useful for inter-data set exploration.

Mouse tissue synchronicity

Although it was shown in [73] that the core clock genes have similar phases across organs, this thesis is the first time (that we know of) that this synchronicity has been quantified. We provided quantitative evidence that a small set of genes are synchronised in expression profile in mouse peripheral circadian clocks.

Human individual synchronicity

It has always been expected that the human transcriptome is synchronised to time-of-day. However, we showed here the first quantitative transcriptomic evidence that a small set of genes are synchronised in expression profile in different human individual circadian clocks. Figure 3.13 showed that this synchronicity is extremely robust.

7.4 Future work in circadian rhythms

- **The connection between the dysregulation of the cell cycle and the circadian clock.** This may be enlightened through the use of single cell RNA-seq technologies.
- **Direct comparison of mouse and human core circadian mechanisms.** Comparisons of the behaviour of the human and mouse data have not been attempted this thesis, but would be a very interesting area to pursue.
- **Novel mathematical models of circadian clock expression.** Korencic *et al.* designed a delay differential equation model of 6 genes using promoter site interactions. Expanding this type of model may be very informative.
- **Raising awareness of circadian variation in biological experiments.** If biopsies were labelled with time of day in the meta-data of a clinical trial, we would be able to use so much more data in this thesis. Recording the time of biopsy, and some general lifestyle characteristics of individuals would not be difficult for studies to do, and this would hugely increase the scope of Time-Teller.

7.5 Final comments

The methods in this PhD project were developed towards the aims of developing a method to determine whether the circadian clock gene expression was different to "normal" circadian gene expression, in a single transcriptome from a tumour biopsy. The algorithm

Time-Teller was designed and validated in this thesis. Time-Teller can be used to generate the metric of clock dysfunction Θ from a single time point assay.

Appendices

Appendix A

SVD and PCA

A.1 Singular Value Decomposition (SVD)

This section has been written with guidance from [175].

The SVD decomposes a matrix into a weighted, ordered sum of separable matrices that describe the major axes of variation. The weightings are given by the singular values, σ_i, and are ordered by size, resulting in the first term of the decomposition explaining the matrix the most.

Eigenvalue problem

For an $n \times n$ square matrix B, with n linearly independent eigenvectors x_i for $i = 1, ..., n$ and n eigenvalues λ_i, the eigenvalue problem says the following is true;

$$Bx_i = \Lambda x_i \tag{A.1}$$

Eigenvalue problem:smymmetric matrices

If B is an $n \times n$ *symmetric* matrix, then its eigenvectors are all orthogonal so that for $i \neq j$ $v_i^T v_j = 0$. This means that for matrix $V = v_1, ..., v_n$, and diagonal matrix D containing all λ^2;

$$B = VDV^T \tag{A.2}$$

where V has the property that $V^{-1} = V^T$

SVD

The single value decomposition of the $m \times n$ matrix A, for m observations and and n features, and rank r, is defined as:

$$A = U\Sigma V^T = \mathbf{u_1}\sigma_1 \mathbf{v_1}^T + ... + \mathbf{u_r}\sigma_r \mathbf{v_r}^T \tag{A.3}$$

where

- $\mathbf{u}_i$ are the columns of $U^{(m\times m)}$ and are the orthonormal eigenvectors of the covariance matrix AA^T and the left singular vectors of A
- $\mathbf{v}_i$ are the columns of $V^{(n\times n)}$ and are the orthonormal eigenvectors of the covariance matrix A^TA and the right singular vectors of A
- $\sigma_i = \sqrt{\lambda_i} = \|A\mathbf{v_i}\|$ are the lengths of the vectors $A\mathbf{v_i}$
- $\Sigma^{(m\times n)}$ is made up of a diagonal matrix of σ_i's (where $\sigma_1 > \sigma_2 > ... > \sigma_r > 0$) in its upper left position, with 0's everywhere else.

Proof

Let A be a mean deviated $m \times n$ matrix, with rank r.

Let $V = v_1, ..., v_r$ be the eigenvectors of symmetric matrix A^TA, and let $\sigma_1^2, ..., \sigma_r^2$ be the associated eigenvalues. Then, by the eigenvalue problem it is true that

$$A^TAv_i = \sigma^2 v_i \tag{A.4}$$
$$AA^T(Av_i) = \sigma^2(Av_i) \tag{A.5}$$
$$\tag{A.6}$$

so that Av_i are the eigenvectors of AA^T.

We test the magnitude of Av_i,

$$(Av_i)^TAv_i = v_i^TA^TAv_i \tag{A.7}$$
$$= v_i^T\lambda_i v_i \tag{A.8}$$
$$= \lambda_i \tag{A.9}$$

Let $u_i, ..., u_n$ be the orthogonal vectors defined by $u_i = Av_i/\sigma_i$.

$$(Av_i/\sigma)^T Av_i = u_i^T Av_j \tag{A.10}$$
$$= v_i A^T Av_i/\sigma_j \tag{A.11}$$
$$= v_i \sigma^2 v_j/\sigma \tag{A.12}$$
$$= v_i^T v_j \sigma_i \tag{A.13}$$

Now, if $i \neq j$ then $v_i^T v_j = 0$, and if $i = j$, $v_i^T v_j = 1$ so that

$$u_i^T Av_i = \sigma_i \tag{A.14}$$
$$Av_i = \sigma_i u_i \tag{A.15}$$

In matrix form this is

$$A\begin{bmatrix} \mathbf{v_1}, ..., \mathbf{v_r} \end{bmatrix} = \begin{bmatrix} \mathbf{u_1}, ..., \mathbf{u_r} \end{bmatrix} \begin{bmatrix} \sigma_1 & & \\ & \ddots & \\ & & \sigma_r \end{bmatrix} \tag{A.16}$$

We need an additional $n - r$ $\mathbf{v}$'s and $m - r$ $\mathbf{u}$'s from the nullspace $\mathbf{N}(A)$ and the left nullspace $\mathbf{N}(A^T)$. They can be orthonormal bases for those two nullspaces (and then automatically orthogonal to the first r $\mathbf{v}$'s and $\mathbf{u}$'s). Now V and U are square.

The new Σ is $m \times n$, where the extra rows and columns are filled with zeros.

$$A\begin{bmatrix} \mathbf{v_1}, ..., \mathbf{v_r}, ..., \mathbf{v_n} \end{bmatrix} = \begin{bmatrix} \mathbf{u_1}, ..., \mathbf{u_r}, ..., \mathbf{u_m} \end{bmatrix} \begin{bmatrix} \sigma_1 & & & & 0 \\ & \ddots & & & \vdots \\ & & \sigma_r & ... & 0 \\ & & \vdots & \ddots & \vdots \\ 0 & ... & 0 & ... & 0 \end{bmatrix} \tag{A.17}$$

$v_1, ..., v_r$ are in the *row space* and $u_1, ..., u_r$ are in the *column space* of A. The singular values $\sigma_1, ...\sigma_r$ are all positive numbers.

The v_i's and u_i's go into the columns of V and U, where orthogonality gives $V^T V = I$ and $U^T U = I$. The σ's go into a diagonal matrix Σ.

u_i are eigenvectors of AA^T and v_i are eigenvectors of $A^T A$.

V is now a square orthogonal matrix, with inverse $V^{-1} = V^T$.

This is the Single Value Decomposition:

$$A = U\Sigma V^T = \mathbf{u_1}\sigma_1 \mathbf{v_1}^T + ... + \mathbf{u_r}\sigma_r \mathbf{v_r}^T \tag{A.18}$$

A.1.1 Principal component analysis

PCA uses the principle components (usually V but can be U depending on the orientation of A and if it is the row space or column space we are decomposing) to project observations into a new coordinate system.

We project an n-dimensional observation x into principle component space using V;

$$y_1 = xv_1 = x_1v_{1,1} + x_2v_{1,2} + ... + x_nv_{1,n} \tag{A.19}$$

$$\vdots \tag{A.20}$$

$$y_n = xv_n = x_1v_{n,1} + x_2v_{n,2} + ... + x_nv_{n,n}$$

Thus, $y = y_1, ..., y_n$ is a linear combination of the original variables $x_1, ..., x_n$, using the entries in the eigenvectors v_i as weights. Usually we take the first 2 or 3 of these to capture the most variance, whilst simplifying further analysis and visualisations. $y = Vx$ simply defines the principal component projection into the decomposed column space.

Appendix B

Relogio Model

B.1 ODE

$$
\begin{aligned}
\frac{\mathrm{d}CCK/BML}{\mathrm{d}t} &= kf_{x1}BML_N - kd_{x1}CCK/BML - d_{x1}CCK/BML \\
\frac{\mathrm{d}P_N^*/C_N}{\mathrm{d}t} &= ki_{z4}P_C^*/C_C - ke_{x2}P_N^*/C_N - d_{x2}P_N^*/C_N \\
\frac{\mathrm{d}P_N/C_N}{\mathrm{d}t} &= ki_{z5}P_C/C_C - ke_{x3}P_N/C_N - d_{x3}P_N/C_N \\
\frac{\mathrm{d}BML_N}{\mathrm{d}t} &= ki_{z8}BML_c + kd_{x1}CCK/BML - kf_{x1}BML_N - d_{x7}BML_N \\
\frac{\mathrm{d}Reverb_n}{\mathrm{d}t} &= ki_{z6}RVRB_c - d_{x5}Reverb_n \\
\frac{\mathrm{d}Ror_n}{\mathrm{d}t} &= ki_{z7}ROR_c - d_{x6}Ror_n \\
\frac{\mathrm{d}Per}{\mathrm{d}t} &= V_{1max}\left(\frac{(1+a\left(\frac{CCK/BML}{k_{t1}}\right)^b}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i1}}\right)^c\left(\frac{CCK/BML}{k_{t1}}\right)^b+\left(\frac{CCK/BML}{k_{t1}}\right)^b}\right) - d_{y1}Per \\
\frac{\mathrm{d}Cry}{\mathrm{d}t} &= V_{2max}\left(\frac{1+d\left(\frac{CCK/BML}{k_{t2}}\right)^e}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i2}}\right)^f\left(\frac{CCK/BML}{k_{t2}}\right)^e+\left(\frac{CCK/BML}{k_{t2}}\right)^e}\frac{1}{1+\left(\frac{Reverb_n}{k_{i21}}\right)^{f_1}}\right) - d_{y2}Cry \\
\frac{\mathrm{d}Reverb}{\mathrm{d}t} &= V_{3max}\left(\frac{1+g\left(\frac{CCK/BML}{k_{t3}}\right)^v}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i3}}\right)^w\left(\frac{CCK/BML}{k_{t3}}\right)^v+\left(\frac{CCK/BML}{k_{t3}}\right)^v}\right) - d_{y3}Reverb \\
\frac{\mathrm{d}Ror}{\mathrm{d}t} &= V_{4max}\left(\frac{1+h\left(\frac{CCK/BML}{k_{t4}}\right)^p)}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i4}}\right)^q\left(\frac{CCK/BML}{k_{t4}}\right)^p+\left(\frac{CCK/BML}{k_{t4}}\right)^p}\right) - d_{y4}Ror \\
\frac{\mathrm{d}Bmal}{\mathrm{d}t} &= V_{5max}\left(\frac{1+i\left(\frac{Ror_n}{k_{t5}}\right)^n}{1+\left(\frac{Reverb_n}{k_{i5}}\right)^m+\left(\frac{Ror_n}{k_{t5}}\right)^n}\right) - d_{y5}Bmal \\
\frac{\mathrm{d}Cry_C}{\mathrm{d}t} &= k_{p2}(Cry+y_{20}) + kd_{z4}P_C^*/C_C + kd_{z5}P_C/C_C - kf_{z5}Cry_CPer_C - kf_{z4}Cry_CPer_C^* - d_{z1}Cry_C \\
\frac{\mathrm{d}Per_C}{\mathrm{d}t} &= k_{p1}(Per+y_{10}) + kd_{z5}P_C/C_C + kd_{phz3}Per_C^* - kf_{z5}Per_CCry_C - kph_{z2}Per_C - d_{z2}Per_C \\
\frac{\mathrm{d}P_C^*/C_C}{\mathrm{d}t} &= kf_{z4}Cry_CPer_C^* + ke_{x2}P_N^*/C_N - ki_{z4}P_C^*/C_C - kd_{z4}P_C^*/C_C - d_{z3}P_C^*/C_C \\
\frac{\mathrm{d}Per{*}_C}{\mathrm{d}t} &= kph_{z2}Per_C + kd_{z4}P_C^*/C_C - kd_{phz3}Per_C^* - kf_{z4}Per_C^*Cry_C - d_{z4}Per_C^* \\
\frac{\mathrm{d}P_C/C_C}{\mathrm{d}t} &= kf_{z5}Cry_CPer_C + ke_{x3}P_N/C_N - ki_{z5}P_C/C_C - kd_{z5}P_C/C_C - d_{z5}P_C/C_C \\
\frac{\mathrm{d}RVRB_c}{\mathrm{d}t} &= k_{p3}(Reverb+y_{30}) - ki_{z6}RVRB_c - d_{z6}RVRB_c \\
\frac{\mathrm{d}ROR_c}{\mathrm{d}t} &= k_{p4}(Ror+y_{40}) - ki_{z7}ROR_c - d_{z7}ROR_c \\
\frac{\mathrm{d}BML_c}{\mathrm{d}t} &= k_{p5}(Bmal+y_{50}) - ki_{z8}BML_c - d_{z8}BML_c
\end{aligned}
$$

Stochastic model. 19 state variables. 44 reactions. 72 parameters.

$$a(1) = kf_{x1}BML_N$$

$$a(2) = kd_{x1}CCK/BML$$

$$a(3) = d_{x1}CCK/BML$$

$$a(4) = ki_{z4}P_C^*/C_C$$

$$a(5) = ke_{x2}P_N^*/C_N$$

$$a(6) = d_{x2}P_N^*/C_N$$

$$a(7) = ki_{z5}P_C/C_C$$

$$a(8) = ke_{x3}P_N/C_N$$

$$a(9) = d_{x3}P_N/C_N$$

$$a(10) = ki_{z8}BML_c$$

$$a(11) = d_{x7}BML_N$$

$$a(12) = ki_{z6}RVRB_c$$

$$a(13) = ki_{z7}ROR_c$$

$$a(15) = d_{x6}Ror_n$$

$$a(16) = V_{1max}\left(\frac{(1+a\left(\frac{CCK/BML}{k_{t1}}\right)^b}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i1}}\right)^c\left(\frac{CCK/BML}{k_{t1}}\right)^b+\left(\frac{CCK/BML}{k_{t1}}\right)^b}\right)$$

$$a(17) = d_{y1}Per$$

$$a(18) = V_{2max}\left(\frac{1+d\left(\frac{CCK/BML}{k_{t2}}\right)^e}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i2}}\right)^f\left(\frac{CCK/BML}{k_{t2}}\right)^e+\left(\frac{CCK/BML}{k_{t2}}\right)^e}\frac{1}{1+\left(\frac{Reverb_n}{k_{i21}}\right)^{f_1}}\right)$$

$$a(19) = d_{y2}Cry$$

$$a(20) = V_{3max}\left(\frac{1+g\left(\frac{CCK/BML}{k_{t3}}\right)^v}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i3}}\right)^w\left(\frac{CCK/BML}{k_{t3}}\right)^v+\left(\frac{CCK/BML}{k_{t3}}\right)^v}\right)$$

$$a(21) = d_{y3}Reverb$$

$$a(22) = V_{4max}\left(\frac{1+h\left(\frac{CCK/BML}{k_{t4}}\right)^p)}{1+\left(\frac{P_N^*/C_N+P_N/C_N}{k_{i4}}\right)^q\left(\frac{CCK/BML}{k_{t4}}\right)^p+\left(\frac{CCK/BML}{k_{t4}}\right)^p}\right)$$

$$a(23) = d_{y4}Ror$$

$$a(24) = V_{5max}\left(\frac{1+i\left(\frac{Ror_n}{k_{t5}}\right)^n}{1+\left(\frac{Reverb_n}{k_{i5}}\right)^m+\left(\frac{Ror_n}{k_{t5}}\right)^n}\right)$$

$$
\begin{aligned}
a(25) &= d_{y5} Bmal \\
a(26) &= k_{p2} Cry \\
a(27) &= kd_{z4} P_C^* / C_C \\
a(28) &= kd_{z5} P_C / C_C \\
a(29) &= kf_{z5} Cry_C Per_C \\
a(30) &= kf_{z4} Cry_C Per_C^* \\
a(31) &= d_{z1} Cry_C \\
a(32) &= k_{p1} Per \\
a(33) &= kd_{phz3} Per_C^* \\
a(34) &= kph_{z2} Per_C \\
a(35) &= d_{z2} Per_C \\
a(36) &= d_{z3} P_C^* / C_C \\
a(37) &= d_{z4} Per_C^* \\
a(38) &= d_{z5} P_C / C_C \\
a(39) &= k_{p3} Reverb \\
a(40) &= d_{z6} RVRB_c \\
a(41) &= k_{p4} Ror \\
a(42) &= d_{z7} ROR_c \\
a(43) &= k_{p5} Bmal \\
a(44) &= d_{z8} BML_c
\end{aligned}
$$

B.1.1 Parameters

Parameter	Value	Description
aa	12	Per
d	12	Cry
g	5	Rev-Erb
h	5	Ror
i	12	Bmal
b	5	Per-activation
c	7	Per-inhibition
e	6	Cry-activation rate
f	4	Cry-inhibition
$f1$	1	Cry-inhibition

v	6	Rev-Erb-activation
w	2	Rev-Erb-inhibition
p	6	Ror-activation
q	3	Ror-inhibition
n	2	Bmal-activation
m	5	Bmal-inhibition
y_{10}	0*omega	Per
y_{20}	0*omega	Cry
y_{30}	0*omega	Rev-Erb
y_{40}	0*omega	Ror
y_{50}	0*omega	Bmal
d_{z1}	0.23	CRYC
d_{z2}	0.25	PERC
d_{z3}	0.6	PERC*
d_{z4}	0.2	PERC*/CRYC
d_{z5}	0.2	PERC/CRYC
d_{z6}	0.31	REV-ERBC
d_{z7}	0.3	RORC
d_{z8}	0.73	BMALC
d_{y1}	0.3	Per
d_{y2}	0.2	Cry
d_{y3}	2	Rev-Erb
d_{y4}	0.2	Ror
d_{y5}	1.6	Bmal
d_{x1}	0.08	CLOCK/BMAL ** -23 PD Huge Phase deriv -38 PD
d_{x2}	0.06	PER*N/CRYN **
d_{x3}	0.09	PERN/CRYN
d_{x5}	0.17	REV-ERBN
d_{x6}	0.12	RORN ** -28 PD
d_{x7}	0.15	BMALN
kf_{x1}	2.3	CLOCK/BMAL-complex formation [hour-1]
kd_{x1}	0.01	CLOCK/BMAL-complex dissociation [hour-1]
kf_{z4}	1/omega	PERC*/CRYC-complex formation [(a.u.hour)-1]
kd_{z4}	1	PERC*/CRYC-complex dissociation [hour-1]
kf_{z5}	1/omega	PERC/CRYC-complex formation [(a.u.hour)-1]
kd_{z54}	1	PERC/CRYC-complex dissociation [hour-1]
ki_{z4}	0.2	PERC*/CRYC
ki_{z5}	0.1	PERC/CRYC
ki_{z6}	0.5	REV-ERBC

ki_{z7}	0.1	RORC ** 12.9 PD
ki_{z8}	0.1	BMALC ** 15.6 PD
ke_{x2}	0.02	PER*N/CRYN ** -21 PD
ke_{x3}	0.02	PERN/CRYN
kph_{z2}	2	PERC-phosphorylation rate
kd_{phz3}	0.05	PERC*-dephosphorylation rate
k_{p1}	0.4	PERC
k_{p2}	0.26	CRYC
k_{p3}	0.37	REV-ERBC
k_{p4}	0.76	RORC
k_{p5}	1.21	BMALC
k_{t1}	3*omega	Per-activation rate
k_{i1}	0.9*omega	Per-inhibition rate
k_{t2}	2.4*omega	Cry-activation rate PLOSGEN PERTUR 5.3 TO SIM RAS
k_{i2}	0.7*omega	Cry-inhibition rate
k_{i21}	5.2*omega	Cry-inhibition rate
k_{t3}	2.07*omega	Rev-Erb-activation rate
k_{i3}	3.3*omega	Rev-Erb-inhibition rate
k_{t4}	0.9*omega	Ror-activation rate
k_{i4}	0.4*omega	Ror-inhibition rate **15.5 PD
k_{t5}	8.35*omega	Bmal-activation rate
k_{i5}	1.94*omega	Bmal-inhibition rate
V_{1max}	1*omega	Per
V_{2max}	2.92*omega	Cry
V_{3max}	1.9*omega	Rev-Erb
V_{4max}	10.9*omega	Ror
V_{5max}	1*omega	Bmal

Appendix C

MATLAB and R functions and code

C.0.2 svd

https://uk.mathworks.com/help/matlab/ref/svd.html

C.0.3 Rhythmiciy detection

cosinor

https://uk.mathworks.com/matlabcentral/fileexchange/20329-cosinor-analysis
Cosinor analysis uses the least squares method to fit a sine wave to a time series. Inputs to cosinor are; t - the times of the observations, y - that observations, w - defined cycle length (here always 24), α - type I error used for confidence interval calculations. Here set to 0.05 which corresponds with 95% confidence intervals.

JTK_CYCLE

https://openwetware.org/wiki/HughesLab:JTK_Cycle

JTK_CYCLE runs in R. It is a non-parametric algorithm whose purpose is to identify rhythmic components in large, genome-scale data sets and estimate their period length, phase, and amplitude.

C.0.4 Gaussians fitting

fitgmdist

https://uk.mathworks.com/help/stats/fitgmdist.html
The MATLAB function *fitgmdist* is used to fit the 3D Gaussians to each set of projections $\mathcal{Q}$. The software optimizes the Gaussian mixture model likelihood using the iterative Expectation-Maximization (EM) algorithm. We set $k = 1$ for all uses in these works to only find one Gaussian cluster.

C.0.5 Splines

pchip

`https://uk.mathworks.com/help/matlab/ref/pchip.html`

Shape-Preserving Piecewise Cubic Interpolation pchip interpolates using a piecewise cubic polynomial $P(x)$ with these properties:

On each subinterval $x_k \leq x \leq x_{k+1}$, the polynomial P(x) is a cubic Hermite interpolating polynomial for the given data points with specified derivatives (slopes) at the interpolation points. P(x) interpolates y, that is, $P(x_j) = y_j$, and the first derivative $\frac{dP}{dx}$ is continuous. The second derivative $\frac{d^2P}{dx^2}$ is probably not continuous so jumps at the x_j are possible. The cubic interpolant P(x) is shape preserving. The slopes at the x_j are chosen in such a way that $P(x)$ preserves the shape of the data and respects monotonicity. Therefore, on intervals where the data is monotonic, so is $P(x)$, and at points where the data has a local extremum, so does $P(x)$.

csape

`https://uk.mathworks.com/help/curvefit/csape.html`

$pp = csape(x, y, conds)$ is the ppform of a cubic spline s with knot sequence x that satisfies $s(x(j)) = y(:, j)$ for all j, as well as an additional end condition at the ends (meaning the leftmost and at the rightmost data site), namely the default condition listed below. The data values $y(:, j)$ may be scalars, vectors, matrices, even ND-arrays. Data values at the same data site are averaged.

conds=periodic matches first and second derivatives at left end with those at right end.

csape returns a parametric cubic spline curve passing through the data, by setting up and solving a tridiagonal system. The accumulated square root of chord-length is used.

The relevant tridiagonal linear system is constructed and solved using the sparse matrix capabilities of MATLAB.

Covariance matrices must have non-zero eigenvalues. It is rare that any of the matrices will have 0 or very small eigenvalues, but as it is possible, the code generating the $\Sigma(t)$ contains an if statement checking this. If any of the eigenvalues of $\Sigma(t)$ are below a given threshold (here arbitrarily 0.0001), the diagonal matrix $I\epsilon$ is added to $\Sigma(t)$. $\epsilon = 0.001$ was set as it consistently ensures positive definiteness, and works within the working precision of MATLAB, whilst minimising the change to the covariance matrix itself.

C.1 Statistical tests

C.1.1 Wilcoxon Rank-sum test

ranksum

`https://uk.mathworks.com/help/stats/ranksum.html`

The Wilcoxon Rank-sum test, also known as the Mann -hitney U test is a non-parametric test that is similar to a t-test, but is a test for difference in medians, not means. Normality does not have to be assumed, as all numbers are ordered and turned into ranks.

The Wilcoxon rank sum test statistic is

$$W = R_1 - \frac{n_1(n_1+1))}{2} \tag{C.1}$$

where R_1 is the sum of the ranks in group 1, and n_1 represents the number of samples in group 1.

The null hypothesis H_0 is

$$\mu_W = \frac{n_1 n_2}{2} \tag{C.2}$$

$$\sigma_W = \sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}} \tag{C.3}$$

$$z = \frac{W - \mu_W}{\sigma_W} \sim \mathbb{N}(0,1) \tag{C.4}$$

The probability of H_0 is

$$p = \frac{\bar{R} - \frac{n_1+1}{2}}{n_2} \tag{C.5}$$

where $\bar{R}$ is the mean of the ranks in group 1. p is the probability that a random sample from one distribution is equally likely to be larger or smaller than a random sample for the other distribution.

C.1.2 Kolmogorov Smirnov 2 sided test

kstest2

`https://uk.mathworks.com/help/stats/kstest2`

The two-sample Kolmogorov-Smirnov test returns a test decision for the null hypothesis that the data in 2 provided vectors are from the same continuous distribution.

C.1.3 Kaplan-Meier survival analysis

Survival analysis is time-to-event analysis, that is, when the outcome of interest is the time until an event occurs.

kmplot

`https://uk.mathworks.com/matlabcentral/fileexchange/22293-kmplot`
Survival times are data that measure follow-up time from a defined starting point to the occurrence of a given event, for example the time from the beginning to the end of a remission period or the time from the diagnosis of a disease to death. Standard statistical techniques cannot usually be applied because the underlying distribution is rarely Normal and the data are often censored. A survival time is described as censored when there is a follow-up time but the event has not yet occurred or is not known to have occurred. The survival function $S(t)$ is defined as the probability of surviving at least to time t. The graph of $S(t)$ against t is called the survival curve. The Kaplan-Meier method can be used to estimate this curve from the observed survival times without the assumption of an underlying probability distribution. This function uses the vectorization technique.

logrank

`https://uk.mathworks.com/matlabcentral/fileexchange/22317-logrank`
The log rank test is a statistical hypothesis test comparing two survival curves. It is used to test the null hypothesis that there is no difference between the population survival curves (i.e. the probability of an event occurring at any time point is the same for each population). This function uses the Kaplan-Meier procedure to estimate the survival function.

C.2 fRMA normalisation

A generic work flow in R for a basic fRMA normalisation of mouse data is shown below;

```
source(``https://bioconductor.org/biocLite.R")
library(``affy")
library(``frma")
library(``mogene.1.0.st.v1frmavecs")
setwd(``directory_with_.CEL_files_and_summary")
summary=read_csv(``summary_of_files.csv")
rawData = read.celfiles(filenames=summary$Filename)
frmadata=frma(rawData)
write.exprs(frmadata, file=``eset_frma.txt")
```

Bibliography

[1] Y. Mrabet., "Circadian Rhythms. GFDL free licence image.."

[2] M. P. B. Utpal Bhadra, Utpal Bhadra, Nirav Thakkar, Paromita Das, "Evolution of circadian rhythms: from bacteria to human," *Sleep Medicine*, vol. 35, pp. 49–61, 2017.

[3] R. Lehmann, L. Childs, P. Thomas, M. Abreu, L. Fuhr, H. Herzel, U. Leser, and A. Relógio, "Assembly of a comprehensive regulatory network for the mammalian circadian clock: A bioinformatics approach," *PLoS ONE*, vol. 10, pp. 1–28, 2015.

[4] H. Ukai and H. R. Ueda, "Systems biology of mammalian circadian clocks.," *Annual review of physiology*, vol. 72, pp. 579–603, jan 2010.

[5] P. F. Innominato, V. P. Roche, O. G. Palesh, A. Ulusakarya, D. Spiegel, and A. L. Francis, "The circadian timing system in clinical oncology," *Annals of Medicine*, vol. 46, no. February, pp. 191–207, 2014.

[6] P. Dierickx, L. W. Van Laake, and N. Geijsen, "Circadian clocks: from stem cells to tissue homeostasis and regeneration," *EMBO reports*, p. e201745130, 2017.

[7] A. Relógio, P. O. Westermark, T. Wallach, K. Schellenberg, A. Kramer, and H. Herzel, "Tuning the mammalian circadian clock: robust synergy of two loops.," *PLoS computational biology*, vol. 7, p. e1002309, dec 2011.

[8] G. Minas and D. A. Rand, "Long-time analytic approximation of large stochastic oscillators: Simulation, analysis and inference," *PLoS Computational Biology*, vol. 13, no. 7, pp. 1–23, 2017.

[9] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, "A circadian gene expression atlas in mammals: implications for biology and medicine.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 45, pp. 16219–24, 2014.

[10] M. E. Hughes, L. DiTacchio, K. R. Hayes, C. Vollmers, S. Pulivarthy, J. E. Baggs, S. Panda, and J. B. Hogenesch, "Harmonics of circadian gene transcription in mammals.," *PLoS genetics*, vol. 5, p. e1000442, apr 2009.

[11] H. R. Ueda, W. Chen, Y. Minami, S. Honma, K. Honma, M. Iino, and S. Hashimoto, "Molecular-timetable methods for detection of body time and rhythm disorders from

single-time-point genome-wide expression profiles.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 31, pp. 11227–32, 2004.

[12] J. J. Hughey, T. Hastie, and A. J. Butte, "ZeitZeiger: supervised learning for high-dimensional data from an oscillatory system," *Nucleic Acids Research*, p. gkw030, 2016.

[13] R. C. Anafi, L. J. Francey, J. B. Hogenesch, and J. Kim, "CYCLOPS reveals human transcriptional rhythms in health and disease," *Proceedings of the National Academy of Sciences*, vol. 114, no. 20, pp. 5312–5317, 2017.

[14] E. E. Laing, C. S. Möller-Levet, N. Poh, N. Santhi, S. N. Archer, and D. J. Dijk, "Blood transcriptome based biomarkers for human circadian phase," *eLife*, vol. 6, pp. 1–26, 2017.

[15] J. Shilts, G. Chen, and J. J. Hughey, "Evidence for widespread dysregulation of circadian clock progression in human cancer," *PeerJ*, vol. 6, p. e4327, 2018.

[16] K. C. G. Van Dycke, W. Rodenburg, C. T. M. Van Oostrom, L. W. M. Van Kerkhof, J. L. A. Pennings, T. Roenneberg, H. Van Steeg, and G. T. J. Van Der Horst, "Chronically Alternating Light Cycles Increase Breast Cancer Risk in Mice," *Current Biology*, vol. 25, no. 14, pp. 1932–1937, 2015.

[17] S. Kiessling, L. Beaulieu-Laroche, I. D. Blum, D. Landgraf, D. K. Welsh, K. F. Storch, N. Labrecque, and N. Cermakian, "Enhancing circadian clock function in cancer cells inhibits tumor growth," *BMC Biology*, vol. 15, no. 1, pp. 1–18, 2017.

[18] M. R. Bauer, S. M. Davidson, J. Bartlebaugh, M. G. V. Heiden, and T. Jacks, "Short Article Circadian Rhythm Disruption Promotes Lung Short Article Circadian Rhythm Disruption," *Cell Metabolism*, vol. 24, no. 2, pp. 324–331, 2016.

[19] V. Blakeman, J. L. Williams, Q.-J. Meng, and C. H. Streuli, "Circadian clocks and breast cancer," *Breast Cancer Research*, vol. 18, no. 1, p. 89, 2016.

[20] T. Yamamoto, Y. Nakahata, H. Soma, M. Akashi, T. Mamine, and T. Takumi, "Transcriptional oscillation of canonical clock genes in mouse peripheral tissues," *BMC Molecular Biology*, vol. 5, pp. 1–9, 2004.

[21] A. Goriki, F. Hatanaka, J. Myung, J. K. Kim, T. Yoritaka, S. Tanoue, T. Abe, H. Kiyonari, K. Fujimoto, Y. Kato, T. Todo, A. Matsubara, D. Forger, and T. Takumi, "A Novel Protein, CHRONO, Functions as a Core Component of the Mammalian Circadian Clock," *PLoS Biology*, vol. 12, no. 4, 2014.

[22] A. Korenčič, G. Bordyugov, R. Košir, D. Rozman, M. Goličnik, and H. Herzel, "The Interplay of cis-Regulatory Elements Rules Circadian Rhythms in Mouse Liver," *PLoS ONE*, vol. 7, no. 11, 2012.

[23] H. R. Ueda, W. Chen, A. Adachi, and H. Wakamatsu, "A transcription factor response element for gene expression during circadian night," *Nature*, vol. 418, no. August, 2002.

[24] Y. Obi-Ioka, K. Ushijima, M. Kusama, E. Ishikawa-Kobayashi, and A. Fujimura, "Involvement of Wee1 in the Circadian RhythmDependent Intestinal Damage Induced by Docetaxel," *Journal of Pharmacology and Experimental Therapeutics*, vol. 347, no. 1, pp. 242–248, 2013.

[25] B. Lemmer, "Discoveries of rhythms in human biological functions: A historical review," *Chronobiology International*, vol. 26, no. 6, pp. 1019–1068, 2009.

[26] F. Halberg, "Some physiological and clinical aspects of 24-hour periodicity.," *J Lancet*, vol. 73, no. 1, pp. 20–32, 1951.

[27] G. Murtas and A. J. Millar, "How plants tell the time," *Current Opinion in Plant Biology*, vol. 3, no. 1, pp. 43–46, 2000.

[28] A. W. Kinsey and M. J. Ormsbee, "The health impact of nighttime eating: Old and new perspectives," *Nutrients*, vol. 7, no. 4, pp. 2648–2662, 2015.

[29] H. A. Raynor, M. R. Goff, S. A. Poole, G. Chen, K. C. Allison, A. Petra, C. Lutz, and H. A. Raynor, "Eating Frequency, Food Intake, and Weight: A Systematic Review of Human and Animal Experimental Studies," *Frontiers in Nutrition*, vol. 2, no. December, 2015.

[30] M. P. Mattson, D. B. Allison, L. Fontana, M. Harvie, V. D. Longo, and W. J. Malaisse, "Meal frequency and timing in health and disease," *PNAS*, vol. 111, no. 47, pp. 16647–16653, 2014.

[31] M. Hastings, J. S. O. Neill, and E. S. Maywood, "Circadian clocks : regulators of endocrine and metabolic rhythms," *Journal of Endocrinology (2007)*, vol. 195, pp. 187–198, 1999.

[32] G. Tosini, I. Ferguson, and K. Tsubota, "Effects of blue light on the circadian system and eye physiology," *Molecular Vision 2016;*, vol. 22, no. August 2015, pp. 61–72, 2016.

[33] J. F. D. Czeisler and C. A., "Effect of Light on Human Cicardian Physiology," *Sleep Med Clin*, vol. 4, no. 2, pp. 165–177, 2009.

[34] T. A. LeGates, D. C. Fernandez, and S. Hattar, "Light as a central modulator of circadian rhythms, sleep and affect," *Nature Reviews Neuroscience*, vol. 15, no. 7, pp. 443–454, 2014.

[35] Y. Hamaguchi, Y. Tahara, H. Kuroda, A. Haraguchi, and S. Shibata, "Entrainment of mouse peripheral circadian clocks to <24h feeding/fasting cycles under 24h light/dark conditions.," *Scientific reports*, vol. 5, no. August, p. 14207, 2015.

[36] T. C. Clarke, L. I. Black, B. J. Stussman, P. M. Barnes, and R. L. Nahin, "Trends in the use of complementary health approaches among adults: United States, 2002-2012.," *Natl. Health Stat. Report*, vol. 79, no. 79, pp. 1–16, 2015.

[37] A. Ballesta, P. F. Innominato, R. Dallmann, D. A. Rand, and F. A. Lévi, "Systems Chronotherapeutics," *Pharmacological Reviews*, vol. 69, no. 2, pp. 161–199, 2017.

[38] J. B. Nathaniel P. Hoyle, Estere Seinkmane, Marrit Putker, Kevin A. Feeney, Toke P. Krogager, Johanna E. Chesham, Liam K. Bray1 Justyn M. Thomas, Ken Dunn and J. S. O'Neill, "Circadian actin dynamics drive rhythmic fibroblast mobilization during wound healing," *Science Translational Medicine*, vol. 9, no. 415, 2017.

[39] A. S. Loudon, A. G. Semikhodskii, and S. K. Crosthwaite, "A brief history of circadian time," *Trends in Genetics*, vol. 16, no. 11, pp. 477–481, 2000.

[40] "No Title."

[41] L. Fuhr, M. Abreu, P. Pett, and A. Relogio, "Circadian systems biology: When time matters," *Computational and Structural Biotechnology Journal*, vol. 13, pp. 417–426, 2015.

[42] B. E. Thalén, B. F. Kjellman, L. Mørkrid, and L. Wetterberg, "Melatonin in light treatment of patients with seasonal and nonseasonal depression," *Acta Psychiatrica Scandinavica*, vol. 92, no. 4, pp. 274–284, 1995.

[43] F. Lévi, A. Okyar, S. Dulong, P. F. Innominato, and J. Clairambault, "Circadian timing in cancer treatments.," *Annual review of pharmacology and toxicology*, vol. 50, pp. 377–421, jan 2010.

[44] O. A. Podkolodnaya, N. N. Tverdokhleb, and N. L. Podkolodnyy, "Computational modeling of the cell-autonomous mammalian circadian oscillator," *BMC Systems Biology*, vol. 11, no. Suppl 1, 2017.

[45] J. O. Lipton, E. D. Yuan, L. M. Boyle, D. Ebrahimi-Fakhari, E. Kwiatkowski, A. Nathan, T. Güttler, F. Davis, J. M. Asara, and M. Sahin, "The circadian protein BMAL1 regulates translation in response to S6K1-mediated phosphorylation," *Cell*, vol. 161, no. 5, pp. 1138–1151, 2015.

[46] M. W. Young and S. a. Kay, "Time zones: a comparative genetics of circadian clocks.," *Nature reviews. Genetics*, vol. 2, no. 9, pp. 702–715, 2001.

[47] A. S. de Wit, R. Nijman, E. Destici, I. Chaves, and G. T. J. van der Horst, *Hepatotoxicity and the Circadian Clock: A Timely Matter*. Elsevier Inc., 2014.

[48] U. Albrecht, "Review Timing to Perfection : The Biology of Central and Peripheral Circadian Clocks," *Neuron*, vol. 74, no. 2, pp. 246–260, 2012.

[49] X. Zhao, H. Cho, R. T. Yu, A. R. Atkins, M. Downes, and R. M. Evans, "Nuclear receptors rock around the clock," *EMBO reports*, vol. 15, no. 5, pp. 518–528, 2014.

[50] J. S. Takahashi, "Molecular components of the circadian clock in mammals," 2015.

[51] Phillip L. Lowrey ; Joseph S.Takahashi, "Genetics of Circadian Rhythms in Mammalian Model Organisms," *Adv Genet*, vol. 74, pp. 175–230, 2011.

[52] J. Yan, H. Wang, Y. Liu, and C. Shao, "Analysis of gene regulatory networks in the mammalian circadian rhythm.," *PLoS computational biology*, vol. 4, p. e1000193, oct 2008.

[53] A. Bhargava, H. Herzel, and B. Ananthasubramaniam, "Mining for novel candidate clock genes in the circadian regulatory network," *BMC Systems Biology*, vol. 9, no. 1, p. 78, 2015.

[54] H. R. Ueda, S. Hayashi, W. Chen, M. Sano, M. Machida, Y. Shigeyoshi, M. Iino, and S. Hashimoto, "System-level identification of transcriptional circuits underlying mammalian circadian clocks.," *Nature genetics*, vol. 37, pp. 187–92, feb 2005.

[55] A. Korenčič, R. Košir, G. Bordyugov, R. Lehmann, D. Rozman, and H. Herzel, "Timing of circadian genes in mammalian tissues.," *Scientific reports*, vol. 4, p. 5782, jan 2014.

[56] S. Shi, A. Hida, O. P. McGuinness, D. H. Wasserman, S. Yamazaki, and C. H. Johnson, "Circadian Clock Gene Bmal1 Is Not Essential; Functional Replacement with its Paralog, Bmal2," *Current Biology*, vol. 20, no. 4, pp. 316–321, 2010.

[57] J. P. DeBruyne, D. R. Weaver, and S. M. Reppert, "CLOCK and NPAS2 have overlapping roles in the suprachiasmatic circadian clock," *Nature Neuroscience*, vol. 10, no. 5, pp. 543–545, 2007.

[58] Y. Annayev, S. Adar, Y. Y. Chiou, J. D. Lieb, A. Sancar, and R. Ye, "Gene model 129 (Gm129) encodes a novel transcriptional repressor that modulates circadian gene expression," *Journal of Biological Chemistry*, vol. 289, pp. 5013–5024, 2014.

[59] R. C. Anafi, Y. Lee, T. K. Sato, A. Venkataraman, C. Ramanathan, I. H. Kavakli, M. E. Hughes, J. E. Baggs, J. Growe, A. C. Liu, J. Kim, and J. B. Hogenesch, "Machine learning helps identify CHRONO as a circadian clock component.," *PLoS biology*, vol. 12, p. e1001840, apr 2014.

[60] S. Honma, T. Kawamoto, Y. Takagi, K. Fujimoto, F. Sato, M. Noshiro, Y. Kato, and K. I. Honma, "Dec1 and Dec2 are regulators of the mammalian molecular clock," *Nature*, vol. 419, no. 6909, pp. 841–844, 2002.

[61] A. Nakashima, T. Kawamoto, K. K. Honda, T. Ueshima, M. Noshiro, T. Iwata, K. Fujimoto, H. Kubo, S. Honma, N. Yorioka, N. Kohno, and Y. Kato, "DEC1 Modulates the Circadian Phase of Clock Gene Expression ," *MOLECULAR AND CELLULAR BIOLOGY*, vol. 28, no. 12, pp. 4080–4092, 2008.

[62] S. Yamaguchi, S. Mitsui, L. Yan, K. Yagita, S. Miyake, and H. Okamura, "Role of DBP in the Circadian Oscillatory Mechanism.," *Molecular and cellular biology*, vol. 20, no. 13, pp. 4773–81, 2000.

[63] M. Stratmann, D. M. Suter, N. Molina, F. Naef, and U. Schibler, "Circadian Dbp Transcription Relies on Highly Dynamic BMAL1-CLOCK Interaction with E Boxes and Requires the Proteasome," *Molecular Cell*, vol. 48, no. 2, pp. 277–287, 2012.

[64] S. Masri, M. Cervantes, P. Sassone-corsi, and I. Irvine, "Circuits," *Curr Opin Cell Biol. 2013*, vol. 25, no. 6, pp. 730–734, 2015.

[65] C. Gérard and A. Goldbeter, "Entrainment of the mammalian cell cycle by the circadian clock: modeling two coupled cellular rhythms.," *PLoS computational biology*, vol. 8, p. e1002516, may 2012.

[66] C. Feillet, P. Krusche, F. Tamanini, R. Janssens, M. J. Downey, P. Martin, M. Teboul, S. Saito, F. Lévi, T. Bretschneider, G. T. J. van der Horst, F. Delaunay, and D. Rand, "Phase locking and multiple oscillating attractors for the coupled mammalian clock and cell cycle.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, pp. 9828–33, jul 2014.

[67] E. Engelen, R. C. Janssens, K. Yagita, V. a. J. Smits, G. T. J. van der Horst, and F. Tamanini, "Mammalian TIMELESS is involved in period determination and DNA damage-dependent phase advancing of the circadian clock.," *PloS one*, vol. 8, p. e56623, jan 2013.

[68] T. Matsuo, S. Yamaguchi, S. Mitsui, A. Emi, F. Shimoda, and H. Okamura, "Control mechanism of the circadian clock for timing of cell division in vivo.," *Science (New York, N.Y.)*, vol. 302, pp. 255–9, oct 2003.

[69] C. H. Ko and J. S. Takahashi, "Molecular components of the mammalian circadian clock.," *Human molecular genetics*, vol. 15 Spec No, pp. R271–7, oct 2006.

[70] C. Dibner and U. Schibler, "Circadian timing of metabolism in animal models and humans," *Journal of Internal Medicine*, vol. 277, no. 5, pp. 513–527, 2015.

[71] J.-C. Leloup and A. Goldbeter, "Critical phase shifts slow down circadian clock recovery: implications for jet lag.," *Journal of theoretical biology*, vol. 333, pp. 47–57, sep 2013.

[72] T. a. LeGates and C. M. Altimus, "Measuring circadian and acute light responses in mice using wheel running activity.," *Journal of visualized experiments : JoVE*, no. 48, pp. 1–7, 2011.

[73] R. Zhang, N. F. Lahens, H. I. Ballance, M. E. Hughes, and J. B. Hogenesch, "A circadian gene expression atlas in mammals: Implications for biology and medicine," *Proceedings of the National Academy of Sciences*, pp. 2–7, oct 2014.

[74] R. L. Perlman, "Mouse Models of Human Disease: An Evolutionary Perspective," *Evolution, Medicine, and Public Health*, p. eow014, 2016.

[75] D. A. Rand, B. V. Shulgin, D. Salazar, and A. J. Millar, "Design principles underlying circadian clocks.," *Journal of the Royal Society, Interface / the Royal Society*, vol. 1, no. 1, pp. 119–130, 2004.

[76] D. A. Rand, B. V. Shulgin, J. D. Salazar, and A. J. Millar, "Uncovering the design principles of circadian clocks: Mathematical analysis of flexibility and evolutionary goals," *Journal of Theoretical Biology*, vol. 238, no. 3, pp. 616–635, 2006.

[77] D. a. Rand, "Mapping global sensitivity of cellular network dynamics: sensitivity heat maps and a global summation law.," *Journal of the Royal Society, Interface / the Royal Society*, vol. 5 Suppl 1, no. May, pp. S59–S69, 2008.

[78] M. Domijan, P. E. Brown, B. V. Shulgin, and D. A. Rand, "PeTTSy: a computational tool for perturbation analysis of complex systems biology models," *BMC Bioinformatics*, vol. 17, no. 1, p. 124, 2016.

[79] B. Kepsutlu, R. Kizilel, and S. Kizilel, "Quantification of interactions among circadian clock proteins via surface plasmon resonance," *Journal of Molecular Recognition*, vol. 27, no. 7, pp. 458–469, 2014.

[80] F. Dyson, "A meeting with Enrico Fermi: How one intuitive physicist rescued a team from fruitless research.," *Nature*, vol. 427, no. January, p. 8540, 2004.

[81] J.-c. Leloup and A. Goldbeter, "Toward a detailed computational model for the mammalian circadian clock," *PNAS*, vol. 100, no. 12, 2003.

[82] J. C. Leloup and A. Goldbeter, "Modeling the mammalian circadian clock: Sensitivity analysis and multiplicity of oscillatory mechanisms," *Journal of Theoretical Biology*, vol. 230, no. 4 SPEC. ISS., pp. 541–562, 2004.

[83] D. B. Forger and C. S. Peskin, "A detailed predictive model of the mammalian circadian clock.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 25, pp. 14806–11, 2003.

[84] J. K. Kim and D. B. Forger, "A mechanism for robust circadian timekeeping via stoichiometric balance.," *Molecular systems biology*, vol. 8, p. 630, jan 2012.

[85] H. P. Mirsky, A. C. Liu, D. K. Welsh, S. A. Kay, and F. J. Doyle, "A model of the cell-autonomous mammalian circadian clock," *Proceedings of the National Academy of Sciences*, vol. 106, no. 27, pp. 11107–11112, 2009.

[86] A. Woller, H. Duez, B. Staels, and M. Lefranc, "A Mathematical Model of the Liver Circadian Clock Linking Feeding and Fasting Cycles to Clock Function," *Cell Reports*, vol. 17, no. 4, pp. 1087–1097, 2016.

[87] A. J. Davidson, O. Castanon-Cervantes, T. L. Leise, P. C. Molyneux, and M. E. Harrington, "Visualizing jet lag in the mouse suprachiasmatic nucleus and peripheral circadian timing system," *European Journal of Neuroscience*, vol. 29, no. 1, pp. 171–180, 2009.

[88] L. Yan and R. Silver, "Resetting the brain clock: Time course and localization of mPER1 and mPER2 protein expression in suprachiasmatic nuclei during phase shifts," *European Journal of Neuroscience*, vol. 19, no. 4, pp. 1105–1109, 2004.

[89] E. D. Buhr and R. N. Van Gelder, "Local photic entrainment of the retinal circadian oscillator in the absence of rods, cones, and melanopsin," *Proceedings of the National Academy of Sciences*, vol. 111, no. 23, pp. 8625–8630, 2014.

[90] V. Jakubcakova, H. Oster, F. Tamanini, C. Cadenas, M. Leitges, G. T. van der Horst, and G. Eichele, "Light Entrainment of the Mammalian Circadian Clock by a PRKCA-Dependent Posttranslational Mechanism," *Neuron*, vol. 54, no. 5, pp. 831–843, 2007.

[91] R. Chen, A. Schirmer, Y. Lee, H. Lee, V. Kumar, S. H. Yoo, J. S. Takahashi, and C. Lee, "Rhythmic PER Abundance Defines a Critical Nodal Point for Negative Feedback within the Circadian Clock Mechanism," *Molecular Cell*, vol. 36, no. 3, pp. 417–430, 2009.

[92] C. Dibner, D. Sage, M. Unser, C. Bauer, T. D 'eysmond, F. Naef, and U. Schibler, "Circadian gene expression is resilient to large fluctuations in overall transcription rates," *The EMBO Journal*, vol. 28262, no. November 2008, pp. 123–134, 2009.

[93] D. T. Gillespie, "Stochastic Simulation of Chemical Kinetics," *Annual Review of Physical Chemistry*, vol. 58, no. 1, pp. 35–55, 2007.

[94] M. L. Guerriero, O. E. Akman, G. V. Ooijen, and D. Chiarugi, "Stochastic models of cellular circadian rhythms in plants help to understand the impact of noise on robustness and clock structure," *Frontiers in Plant Science*, vol. 5, no. October, pp. 1–6, 2014.

[95] D. Gonze, J. Halloy, and a. Goldbeter, "Deterministic and stochastic models for circadian rhythms," *Pathologie-biologie*, vol. 51, no. 4, pp. 227–230, 2003.

[96] D. Gonze, J. Halloy, and A. Goldbeter, "Robustness of circadian rhythms with respect to molecular noise.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 2, pp. 673–678, 2002.

[97] D. J. Higham, "Modeling and Simulating Chemical Reactions," *SIAM Review*, vol. 50, no. 2, pp. 347–368, 2008.

[98] Kurt and T. Z, "Approximation of Population Processes. Regional Conference Series in Applied Mathematics," *Society for Industrial and Applied Mathematics.*, vol. 36, 1981.

[99] N. G. Van Kampen, *Stochastic processes in physics and chemistry*, vol. 11. Elsevier, 1992.

[100] D. B. Forger and C. S. Peskin, "Stochastic simulation of the mammalian circadian clock.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 2, pp. 321–324, 2005.

[101] J. Li, G. R. Grant, J. B. Hogenesch, and M. E. Hughes, "Considerations for RNA-seq analysis of circadian rhythms.," *Methods in enzymology*, vol. 551, no. JANUARY, pp. 349–67, 2015.

[102] D. D. Dalma-Weiszhausz, J. Warrington, E. Y. Tanimoto, and C. G. Miyada, "[1] The Affymetrix GeneChip® Platform: An Overview," *Methods in Enzymology*, vol. 410, no. 06, pp. 3–28, 2006.

[103] T. Park, S. G. Yi, S. H. Kang, S. Y. Lee, Y. S. Lee, and R. Simon, "Evaluation of normalization methods for microarray data," *BMC Bioinformatics*, vol. 4, pp. 1–13, 2003.

[104] M. N. McCall, B. M. Bolstad, and R. A. Irizarry, "Frozen robust multiarray analysis (fRMA)," *Biostatistics*, vol. 11, no. 2, pp. 242–253, 2010.

[105] B. Fang, L. J. Everett, J. Jager, E. Briggs, S. M. Armour, D. Feng, A. Roy, Z. Gerhart-Hines, Z. Sun, and M. A. Lazar, "Circadian enhancers coordinate multiple phases of rhythmic gene transcription in vivo," *Cell*, vol. 159, no. 5, pp. 1140–1152, 2014.

[106] J. L. Barclay, J. Husse, B. Bode, N. Naujokat, J. Meyer-Kovac, S. M. Schmid, H. Lehnert, and H. Oster, "Circadian desynchrony promotes metabolic disruption in a mouse model of shiftwork," *PLoS ONE*, vol. 7, no. 5, 2012.

[107] G. Le Martelot, D. Canella, L. Symul, E. Migliavacca, F. Gilardi, R. Liechti, O. Martin, K. Harshman, M. Delorenzi, B. Desvergne, W. Herr, B. Deplancke, U. Schibler, J. Rougemont, N. Guex, N. Hernandez, and F. Naef, "Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles," *PLoS Biology*, vol. 10, no. 11, 2012.

[108] A. A. Saeed, A. H. Sims, S. S. Prime, I. Paterson, P. G. Murray, and V. R. Lopes, "Gene expression profiling reveals biological pathways responsible for phenotypic heterogeneity between UK and Sri Lankan oral squamous cell carcinomas," *Oral Oncology*, vol. 51, no. 3, pp. 237–246, 2015.

[109] J. O. Boyle, Z. H. Gümü, A. Kacker, V. L. Choksi, J. M. Bocker, X. K. Zhou, R. K. Yantiss, D. B. Hughes, B. Du, B. L. Judson, K. Subbaramaiah, and A. J. Dannenberg, "Effects of cigarette smoke on the human oral mucosal transcriptome," *Cancer Prevention Research*, vol. 3, no. 3, pp. 266–278, 2010.

[110] F. P. Roth, G. A. Troia, C. K. Worthington, and D. Handy, "Promoting Awareness of Sounds in Speech (PASS): The effects of intervention and stimulus characteristics on the blending performance of preschool children with communication impairments," *Learning Disability Quarterly*, vol. 29, no. 2, pp. 67–88, 2006.

[111] C. Chen, E. Méndez, J. Houck, W. Fan, P. Lohavanichbutr, D. Doody, B. Yueh, N. D. Futran, M. Upton, D. G. Farwell, S. M. Schwartz, and L. P. Zhao, "Gene expression profiling identifies genes predictive of oral squamous cell carcinoma.," *Cancer Epidemiol Biomarkers Prev*, vol. 17, no. 8, pp. 2152–62, 2008.

[112] A. L. Richardson, Z. C. Wang, A. De Nicolo, X. Lu, M. Brown, A. Miron, X. Liao, J. D. Iglehart, D. M. Livingston, and S. Ganesan, "X chromosomal abnormalities in basal-like human breast cancer," *Cancer Cell*, vol. 9, no. 2, pp. 121–132, 2006.

[113] P. de Cremoux, F. Valet, D. Gentien, J. Lehmann-Che, V. Scott, C. Tran-Perennou, C. Barbaroux, N. Servant, S. Vacher, B. Sigal-Zafrani, M.-C. Mathieu, P. Bertheau, J.-M. Guinebretière, B. Asselain, M. Marty, and F. Spyratos, "Importance of pre-analytical steps for transcriptome and RT-qPCR analyses in the context of the phase II randomised multicentre trial REMAGUS02 of neoadjuvant chemotherapy in breast cancer patients," *BMC Cancer*, vol. 11, no. 1, p. 215, 2011.

[114] M. E. Hughes, J. B. Hogenesch, and K. Kornacker, "JTK_CYCLE: an efficient non-parametric alorithm for detecting rhythmic components in genome-scale datasets," *Journal of biological rhythms*, vol. 25, no. 5, pp. 372–380, 2011.

[115] H. Kitamura, T. Ishino, Y. Shimamoto, J. Okabe, T. Miyamoto, E. Takahashi, and I. Miyoshi, "Ubiquitin-Specific Protease 2 Modulates the Lipopolysaccharide-Elicited Expression of Proinflammatory Cytokines in Macrophage-like HL-60 Cells," *Mediators of Inflammation*, vol. 2017, 2017.

[116] G. Wu, J. Zhu, J. Yu, L. Zhou, J. Z. Huang, and Z. Zhang, "Evaluation of five methods for genome-wide circadian gene identification," *Journal of Biological Rhythms*, vol. 29, no. 4, pp. 231–242, 2014.

[117] G. Cornelissen, "Cosinor-based rhythmometry," *Theoretical Biology and Medical Modelling*, vol. 11, no. 1, pp. 1–24, 2014.

[118] H. Matsumae, R. Ishiwata, T. Minamoto, N. Ishida, S. Ogishima, and H. Tanaka, "Detection of periodic patterns in microarray data reveals novel oscillating transcripts of biological rhythms in Ciona intestinalis," *Artificial Life and Robotics*, vol. 20, no. 4, pp. 347–352, 2015.

[119] S. Panda, M. P. Antoch, B. H. Miller, A. I. Su, A. B. Schook, M. Straume, P. G. Schultz, S. A. Kay, J. S. Takahashi, and J. B. Hogenesch, "Coordinated transcription of key pathways in the mouse by the circadian clock," *Cell*, vol. 109, no. 3, pp. 307–320, 2002.

[120] S. Zhao, W. P. Fung-Leung, A. Bittner, K. Ngo, and X. Liu, "Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells," *PLoS ONE*, vol. 9, no. 1, 2014.

[121] A. Flis, A. P. Fernández, T. Zielinski, V. Mengin, R. Sulpice, K. Stratford, A. Hume, A. Pokhilko, M. M. Southern, D. D. Seaton, H. G. McWatters, M. Stitt, K. J. Halliday, and A. J. Millar, "Defining the robust behaviour of the plant clock gene circuit with absolute RNA timeseries and open infrastructure," *Open Biology*, vol. 5, no. 10, p. 150042, 2015.

[122] A. Atwood, R. DeConde, S. S. Wang, T. C. Mockler, J. S. M. Sabir, T. Ideker, and S. A. Kay, "Cell-autonomous circadian clock of hepatocytes drives rhythms in transcription and polyamine synthesis," *Proceedings of the National Academy of Sciences*, vol. 108, no. 45, pp. 18560–18565, 2011.

[123] S. W. Cain, C. F. Dennison, J. M. Zeitzer, A. M. Guzik, S. B. S. Khalsa, N. Santhi, M. W. Schoen, C. A. Czeisler, and J. F. Duffy, "Sex differences in phase angle of entrainment and melatonin amplitude in humans," *Journal of Biological Rhythms*, vol. 25, no. 4, pp. 288–296, 2010.

[124] "Affymetrix website,"

[125] K. Eckel-mahan and P. Sassone-corsi, "HHS Public Access," *Curr Protoc Mouse Biol.*, vol. 5, no. 3, pp. 271–281, 2016.

[126] J. R. Biller, H. Elajaili, V. Meyer, G. M. Rosen, S. S. Eaton, and G. R. Eaton, "NIH Public Access," *Journal of MAgnetic Resonance*, vol. 236, no. 46, pp. 47–56, 2013.

[127] R. B. Toomey and K. J. Mitchell, "HHS Public Access," *Circulation*, vol. 51, no. 1, pp. 87–100, 2016.

[128] H. Hamzeiy, J. Allmer, and M. Yousef, *Computational Methods for MicroRNA Target Prediction.*, vol. 1107. 2014.

[129] L. V. D. Maaten and G. Hinton, "Visualizing Data using t-SNE," vol. 9, pp. 2579–2605, 2008.

[130] E. Bair, T. Hastie, D. Paul, and R. Tibshirani, "Prediction by supervised principal components," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 119–137, 2006.

[131] J. J. Hughey, "Machine learning identifies a compact gene set for monitoring the circadian clock in human blood," *Genome Medicine*, vol. 9, no. 1, pp. 1–11, 2017.

[132] F. Agostinelli, N. Ceglia, B. Shahbaba, P. Sassone-Corsi, and P. Baldi, "What time is it? Deep learning approaches for circadian rhythms," *Bioinformatics*, vol. 32, no. 12, pp. i8–i17, 2016.

[133] O. Alter, P. O. Brown, and D. Botstein, "Singular value decomposition for genome-wide expression data processing and modeling," *Proceedings of the National Academy of Sciences*, vol. 97, no. 18, pp. 10101–10106, 2000.

[134] A.-L. Boulesteix and K. Strimmer, "Partial least squares: a versatile tool for the analysis of high-dimensional genomic data," *Briefings in Bioinformatics*, vol. 8, no. 1, pp. 32–44, 2006.

[135] D. Yang, Z. Ma, and A. Buja, "A Sparse SVD Method for High-dimensional Data," *arXiv*, vol. 19104, pp. 1–37, 2011.

[136] W. E. Johnson, C. Li, and A. Rabinovic, "Adjusting batch effects in microarray expression data using empirical Bayes methods," *Biostatistics*, vol. 8, no. 1, pp. 118–127, 2007.

[137] N. Leng, L. F. Chu, C. Barry, Y. Li, J. Choi, X. Li, P. Jiang, R. M. Stewart, J. A. Thomson, and C. Kendziorski, "Oscope identifies oscillatory genes in unsynchronized single-cell RNA-seq experiments," *Nature Methods*, vol. 12, no. 10, pp. 947–950, 2015.

[138] C.-Y. Chen, R. W. Logan, T. Ma, D. A. Lewis, G. C. Tseng, E. Sibille, and C. A. McClung, "Effects of aging on circadian patterns of gene expression in the human prefrontal cortex," *Proceedings of the National Academy of Sciences*, vol. 113, no. 1, pp. 206–211, 2016.

[139] R. E. Kerwin, J. M. Jimenez-Gomez, D. Fulop, S. L. Harmer, J. N. Maloof, and D. J. Kliebenstein, "Network Quantitative Trait Loci Mapping of Circadian Clock Outputs Identifies Metabolic Pathway-to-Clock Linkages in Arabidopsis," *The Plant Cell*, vol. 23, no. 2, pp. 471–485, 2011.

[140] J. Matsuzaki, Y. Kawahara, and T. Izawa, "Punctual Transcriptional Regulation by the Rice Circadian Clock under Fluctuating Field Conditions," *The Plant Cell*, vol. 27, no. 3, pp. 633–648, 2015.

[141] C. de Boor, *A Practical Guide to Splines*. Springer-Verlag, 1978.

[142] S. R. Cole, H. Chu, and S. Greenland, "Maximum likelihood, profile likelihood, and penalized likelihood: A primer," *American Journal of Epidemiology*, vol. 179, no. 2, pp. 252–260, 2014.

[143] D. S. Broomhead, R. Indik, A. C. Newell, and D. A. Rand, "Local Adaptive Galerkin Bases for Large Dimensional Dynamical Systems," *Nonlinearity*, vol. 4, pp. 159–197, 1991.

[144] E. D. Buhr and J. S. Takahashi, "Molecular components of the mammalian circadian clock," *Handbook of Experimental Pharmacology*, vol. 217, no. 217, pp. 3–27, 2013.

[145] E. Filipski, P. F. Innominato, M. W. Wu, X. M. Li, S. Iacobelli, L. J. Xian, and F. Lévi, "Effects of light and food schedules on liver and tumor molecular clocks in mice," *Journal of the National Cancer Institute*, vol. 97, no. 7, pp. 507–517, 2005.

[146] K. Straif, R. Baan, Y. Grosse, B. Secretan, F. E. Ghissassi, V. Bouvard, A. Altieri, L. Benbrahim-Tallaa, and V. Cogliano, "Carcinogenicity of shift-work, painting, and firefighting," *The Lancet Oncology*, vol. 8, pp. 1065–1066, dec 2007.

[147] T. C. Erren, P. Falaturi, P. Morfeld, P. Knauth, R. J. Reiter, and C. Piekarski, "Shift work and cancer: the evidence and the challenge," *Deutsches Arzteblatt International*, vol. 107, no. 38, pp. 657–662, 2010.

[148] L. R. Wegrzyn, R. M. Tamimi, B. A. Rosner, S. B. Brown, R. G. Stevens, A. H. Eliassen, F. Laden, W. C. Willett, S. E. Hankinson, and E. S. Schernhammer, "Original Contribution Rotating Night-Shift Work and the Risk of Breast Cancer in the Nurses ' Health Studies," *American Journal of Epidemiology*, vol. 186, no. 5, pp. 532–540, 2018.

[149] R. C. Travis, A. Balkwill, G. K. Fensom, P. N. Appleby, G. K. Reeves, X.-s. Wang, A. W. Roddam, T. Gathani, R. Peto, J. Green, T. J. Key, and V. Beral, "Night Shift Work and Breast Cancer Incidence : Three Prospective Studies and Meta-analysis of Published Studies," *NCI J Natl Cancer Inst (2016)*, vol. 108, no. December, pp. 1–9, 2017.

[150] S. Mocellin, S. Tropea, C. Benna, and C. R. Rossi, "Circadian pathway genetic variation and cancer risk : evidence from genome-wide association studies," *BMC Medicine*, vol. 16, no. 20, pp. 1–8, 2018.

[151] D. Hanahan and R. Weinberg, "Hallmarks of cancer: The next generation," *Cell*, vol. 144, pp. 646–674, mar 2011.

[152] X.-M. Li, A. Mohammad-Djafari, M. Dumitru, S. Dulong, E. Filipski, S. Siffroi-Fernandez, A. Mteyrek, F. Scaglione, C. Guettier, F. Delaunay, and F. Lévi, "A circadian clock transcription model for the personalization of cancer chronotherapy.," *Cancer research*, vol. 73, pp. 7176–88, dec 2013.

[153] F. Lévi, A. Altinok, and A. Goldbeter, "Cancer Systems Biology, Bioinformatics and Medicine," in *Cancer Systems Biology, Bioinformatics and Medicine* (A. Cesario and F. Marcus, eds.), ch. 15, pp. 381–408, Dordrecht: Springer Netherlands, 2011.

[154] R. G. Stevens, G. C. Brainard, D. E. Blask, S. W. Lockley, and M. E. Motta, "Breast Cancer and Circadian Disruption From Electric Lighting in the Modern World," *CA CANCER J CLIN*, vol. 64, no. 3, pp. 207–218, 2014.

[155] L. Fritschi, D. C. Glass, J. S. Heyworth, K. Aronson, J. Girschik, T. Boyle, A. Grundy, and T. C. Erren, "Hypotheses for mechanisms linking shiftwork and cancer," *Medical Hypotheses*, vol. 77, no. 3, pp. 430–436, 2011.

[156] E. L. Haus and M. H. Smolensky, "Shift work and cancer risk: Potential mechanistic roles of circadian disruption, light at night, and sleep deprivation," *Sleep Medicine Reviews*, vol. 17, no. 4, pp. 273–284, 2013.

[157] J. Samulin Erdem, H. Ø. Notø, Ø. Skare, J. A. S. Lie, M. Petersen-Øverleir, E. Reszka, B. Pepłońska, and S. Zienolddiny, "Mechanisms of breast cancer risk in shift workers: Association of telomere shortening with the duration and intensity of night work," *Cancer Medicine*, 2017.

[158] B. J. Altman, A. L. Hsieh, A. Sengupta, J. B. Hogenesch, A. M. Weljie, C. V. Dang, B. J. Altman, A. L. Hsieh, A. Sengupta, S. Y. Krishnanaiah, and Z. E. Stine, "Article MYC Disrupts the Circadian Clock and Metabolism in Cancer Cells Article MYC Disrupts the Circadian Clock and Metabolism in Cancer Cells," *Cell Metabolism*, vol. 22, pp. 1009–1019, 2015.

[159] A. Salavaty, N. Mohammadi, M. Shahmoradi, and M. N. Soorki, "Bioinformatic Analysis of Circadian Expression of Oncogenes and Tumor Suppressor Genes," *Bioinformatics and Biology Insights*, vol. 11, pp. 1–9, 2017.

[160] C. Gérard and A. Goldbeter, "From quiescence to proliferation: Cdk oscillations drive the mammalian cell cycle.," *Frontiers in physiology*, vol. 3, p. 413, jan 2012.

[161] S. Giacchetti, P. A. Dugué, P. F. Innominato, G. A. Bjarnason, C. Focan, C. Garufi, S. Tumolo, B. Coudert, S. Iacobelli, R. Smaaland, M. Tampellini, R. Adam, T. Moreau, and F. Lévi, "Sex moderates circadian chemotherapy effects on survival of patients with metastatic colorectal cancer: A meta-analysis," *Annals of Oncology*, vol. 23, no. 12, pp. 3110–3116, 2012.

[162] J. H. Lee and A. Sancar, "Circadian clock disruption improves the efficacy of chemotherapy through p73-mediated apoptosis," *Proceedings of the National Academy of Sciences*, vol. 108, no. 26, pp. 10668–10672, 2011.

[163] C. Savvidis and M. Koutsilieris, "Circadian Rhythm Disruption in Cancer Biology," *Molecular Medicine*, vol. 18, no. 9, p. 1, 2012.

[164] F. Okazaki, N. Matsunaga, H. Okazaki, H. Azuma, K. Hamamura, A. Tsuruta, Y. Tsurudome, T. Ogino, Y. Hara, T. Suzuki, K. Hyodo, H. Ishihara, H. Kikuchi, H. To, H. Aramaki, S. Koyanagi, and S. Ohdo, "Circadian clock in a mouse colon tumor regulates intracellular iron levels to promote tumor progression," *Journal of Biological Chemistry*, vol. 291, no. 13, pp. 7017–7028, 2016.

[165] B. J. Altman, "Cancer Clocks Out for Lunch: Disruption of Circadian Rhythm and Metabolic Oscillation in Cancer," *Frontiers in Cell and Developmental Biology*, vol. 4, no. June, pp. 1–9, 2016.

[166] A. Relógio, P. Thomas, P. Medina-Pérez, S. Reischl, S. Bervoets, E. Gloc, P. Riemer, S. Mang-Fatehi, B. Maier, R. Schäfer, U. Leser, H. Herzel, A. Kramer, and C. Sers, "Ras-mediated deregulation of the circadian clock in cancer.," *PLoS genetics*, vol. 10, p. e1004338, may 2014.

[167] C. Saini, S. A. Brown, and C. Dibner, "Human peripheral clocks: Applications for studying circadian phenotypes in physiology and pathophysiology," *Frontiers in Neurology*, vol. 6, no. MAY, pp. 1–8, 2015.

[168] L. Feng, J. R. Houck, P. Lohavanichbutr, and C. Chen, "Transcriptome analysis reveals differentially expressed lncRNAs between oral squamous cell carcinoma and healthy oral mucosa.," *Oncotarget*, vol. 8, no. 19, pp. 31521–31531, 2017.

[169] P. Taneja, D. Maglic, F. Kai, S. Zhu, R. D. Kendig, E. A. Fry, and K. Inoue, "Classical and Novel Prognostic Markers for Breast Cancer and their Clinical Significance," *Clin Med Insights Oncol*, vol. 4, pp. 15–34, 2010.

[170] W. D. Foulkes, "Size surprise? Tumour size, nodal status, and outcome after breast cancer," *Current Oncology*, vol. 19, no. 5, pp. 241–243, 2012.

[171] S. Giacchetti, A. S. Hamy, S. Delaloge, E. Brain, F. Berger, B. Sigal-Zafrani, M. C. Mathieu, P. Bertheau, J. M. Guinebretière, M. Saghatchian, F. Lerebours, chafouny Mazouni, O. Tembo, M. Espié, F. Reyal, M. Marty, B. Asselain, and J. Y. Pierga, "Long-term outcome of the REMAGUS 02 trial, a multicenter randomised phase II trial in locally advanced breast cancerpatients treated with neoadjuvant chemotherapy with or without celecoxib or trastuzumab according to HER2 status," *European Journal of Cancer*, vol. 75, no. January, pp. 323–332, 2017.

[172] C. Cadenas, L. van de Sandt, K. Edlund, M. Lohr, B. Hellwig, R. Marchan, M. Schmidt, J. Rahnenführer, H. Oster, and J. G. Hengstler, "Loss of circadian clock gene expression

is associated with tumor progression in breast cancer.," *Cell cycle (Georgetown, Tex.)*, vol. 13, pp. 3282–91, oct 2014.

[173] F. Valet, P. de Cremoux, F. Spyratos, N. Servant, M. E. Dujaric, D. Gentien, J. Lehmann-Che, V. Scott, B. Sigal-Zafrani, M. C. Mathieu, P. Bertheau, J. M. Guinebretière, J. Y. Pierga, S. Delaloge, S. Giacchetti, E. Brain, O. Tembo, M. Marty, and B. Asselain, "Challenging single- and multi-probesets gene expression signatures of pathological complete response to neoadjuvant chemotherapy in breast cancer: experience of the REMAGUS 02 phase II trial.," *Breast (Edinburgh, Scotland)*, vol. 22, pp. 1052–9, dec 2013.

[174] J. Y. Pierga, S. Delaloge, M. Espié, E. Brain, B. Sigal-Zafrani, M. C. Mathieu, P. Bertheau, J. M. Guinebretiére, M. Spielmann, A. Savignoni, and M. Marty, "A multicenter randomized phase II study of sequential epirubicin/ cyclophosphamide followed by docetaxel with or without celecoxib or trastuzumab according to HER2 status, as primary chemotherapy for localized invasive breast cancer patients," *Breast Cancer Research and Treatment*, vol. 122, no. 2, pp. 429–437, 2010.

[175] D. C. Lay, "Linear Algebra and its Applications," *Linear Algebra and Its Applications*, vol. 437, no. 11, pp. 2755–2772, 2012.