Aalto University
School of Science
Bachelor's Programme in Science and Technology

Circadian rhythm in molecular data

Bachelor's Thesis

April 16, 2023

Paavo Nurminen

Aalto University
School of Science
Bachelor's Programme in Science and Technology

ABSTRACT OF
BACHELOR'S THESIS

Author:	Paavo Nurminen
Title of thesis:	Circadian rhythm in molecular data
Date:	April 16, 2023
Pages:	26
Major:	Computer Science
Code:	SCI3027
Supervisor:	Prof. Eero Hyvönen
Instructor:	PhD, Lu Cheng (Department of Computer Science)

In this bachelor's thesis, the effects of circadian rhythms on genetic data are examined. The circadian rhythm is an internal biological clock that regulates many physiological processes in organisms, such as sleep and metabolism. The objective of my bachelor's thesis is to determine which genes appear to be expressed periodically in accordance with the circadian rhythm. The samples have been obtained from public research material, in which the expression of thousands of different genes in mice liver samples has been recorded at various times of the day. This genetic expression data is analyzed using Gaussian processes to determine if gene expression follows the circadian rhythm.

The work reviews the basic mechanisms of circadian rhythms and their scientific and medical applications. The circadian rhythm regulates cellular function and intercellular communication. Gene expression is a key area in which the circadian rhythm affects cellular function. Gene expression refers to the process by which cells produce proteins based on the information contained within genes.

In this study, genetic expression data from mice were collected and analyzed using Gaussian process regression models. Gaussian processes are a set of mathematical models that can be used to model continuously changing phenomena. In this case, they identify periodically expressed genes that may be related to the circadian rhythm. As a result of the analysis, 297 and 206 genes out of 10246 were found for different age groups whose expression varied periodically in accordance with the circadian rhythm. This suggests that these genes are part of the circadian rhythm regulatory system or regulate other bodily functions under the guidance of this system. The identified genes were also compared to known circadian rhythm-regulating genes found in other literature. The result obtained in this study could be successfully validated, as the literature-known genes were identified to follow the circadian rhythm using the method employed in this study. Furthermore, the thesis examined how these findings can help to understand the functioning and applications of circadian rhythms in different scientific fields.

In summary, this bachelor's thesis provides a literature review and a re-analysis of public a dataset to study the effects of circadian rhythms on gene expression and its significance in the regulation of various biological processes. This study shows the usefulness of this Gaussian process-based approach by visual inspection of several example genes, which is further substantiated by identifying classic circadian genes.

Keywords:	Circadian rhythm, Gaussian Process, gene expression, gene analysis
Language:	English

Tekijä:	Paavo Nurminen
Työn nimi:	Circadian rhythm in molecular data
Päiväys:	16.4.2023
Sivumäärä:	26
Pääaine:	Computer Science
Koodi:	SCI3027
Vastuuopettaja:	Prof. Eero Hyvönen
Työn ohjaaja(t):	FT, Lu Cheng (Tietotekniikan laitos)

Tässä kandidaatintutkielmassa tarkastellaan vuorokausirytmin vaikutuksia geneettiseen dataan. Vuorokausirytmi on sisäinen biologinen kello, joka säätelee monia eliöiden fysiologisia prosesseja, kuten unta ja aineenvaihduntaa. Kandidaatintyöni tavoitteena on selvittää, mitkä geenit näyttävät ilmentyvän jaksollisesti vuorokausirytmin mukaisesti. Näytteet on saatu julkisesta tutkimusmateriaalista, jossa hiirten maksanäytteistä on kirjattu tuhansien eri geenien ilmentymä eri vuorokauden aikoina. Tätä geneettistä ilmentymistietoa analysoidaan gaussisten prosessien avulla, jotta voidaan todeta, seuraako geenien ilmentyminen vuorokausirytmiä.

Työssä käydään läpi vuorokausirytmin perusmekanismeja ja sen tieteellisiä sekä lääketieteellisiä sovelluksia. Vuorokausirytmi säätelee solujen toimintaa ja solujen välisiä viestintää. Geenien ilmentyminen on yksi keskeinen osa-alue, jossa vuorokausirytmi vaikuttaa solujen toimintaan. Geenien ilmentyminen tarkoittaa sitä, miten solut tuottavat proteiineja geenien sisältämän informaation perusteella.

Vuorokausirytmin tutkimus on tärkeää, sillä sen ymmärtäminen voi auttaa kehittämään uusia hoitomuotoja moniin sairauksiin, kuten unihäiriöihin ja metabolisiin häiriöihin. Lisäksi vuorokausirytmin merkitys vaikuttaa olevan suuri myös ikääntymisen kannalta. Yhteyksiä moniin eri toimintoihin on havaittu, mutta pääasiassa niiden taustalla olevat mekanismit ovat vielä tuntemattomia.

Työssä kerättiin hiiriltä geneettistä ilmentymisdataa ja analysoitiin se gaussin prosessi regressiomallien avulla. Gaussiset prosessit ovat joukko matemaattisia malleja, joilla voidaan mallintaa jatkuvasti muuttuvia ilmiöitä. Tässä tapauksessa ne auttavat tunnistamaan jaksollisesti ekspressoituvia geenejä, joilla on mahdollinen yhteys vuorokausirytmiin. Analyysin tuloksena löydettiin useita geenejä, joiden ilmentyminen vaihteli jaksollisesti vuorokausirytmin mukaisesti. Tämä viittaa siihen, että nämä geenit ovat osa vuorokausirytmin säätelyjärjestelmää, tai säätelevät kehon muita toimintoja säätelyjärjestelmän ohjaamina. Löydettyjä geenejä verrattiin myös muussa kirjallisuudessa löydettyihin, tunnettuihin vuorokausirytmiä ohjaaviin geeneihin. Tässä työssä saatu tulos voitiin validoida onnistuneeksi, sillä kirjallisuudessa tunnetut geenit tunnistettiin seuraamaan vuorokausirytmiä myös tässä työssä käytetyllä menetelmällä. Lisäksi tutkielmassa tarkasteltiin, miten nämä löydökset voivat auttaa ymmärtämään vuorokausirytmin toimintaa ja sovelluksia eri tieteenaloilla.

Mahdollisen jatkotutkimuksen tulokset voivat olla hyödyllisiä niin perustutkimuksessa kuin lääketieteen sovelluksissa. Jatkotutkimuksissa voitaisiin esimerkiksi tarkastella, miten löydettyjen geenien toimintaa voitaisiin säädellä tai hyödyntää uusien hoitomuotojen kehittämisessä. Samalla voidaan myös tutkia vuorokausirytmin vaikutuksia erilaisiin sairauksiin ja niiden hoidossa käytettyihin lääkkeisiin.

Yhteenvetona tämä kandidaatintutkielma tarjoaa kirjallisuuskatsauksen ja itse tuotetun analyysin vuorokausirytmin vaikutuksista geenien ilmentymiseen sekä sen merkityksestä erilaisten biologisten prosessien säätelyssä. Tutkimuksen avulla voidaan laajentaa ymmärrystämme vuorokausirytmin monimutkaisista mekanismeista ja niiden merkityksestä elämälle.

Avainsanat:	Vuorokausirymi, gaussiset prosessit, geenianalyysi, geeni-ilmentymä
Kieli:	Suomi

Contents

Used abbreviations, symbols and terminology

Ad libitum	In the context of this thesis: to consume food at will
BMAL1	basic helix-loop-helix ARNT like 1, core clock gene
Cistrome	Collection of regulatory elements of a set of genes
CLOCK	Circadian Locomoter Output Cycles Protein Kaput, a core clock gene
Cryptochrome	Core clock gene that produces proteins CRY1, CRY2
Entrainment	Synchronization of processes based on zeitgebers
GP	Gaussian process
Ketogenesis	Biological process to produce ketones
LLR	Log-Likelihood-ratio
Midbrain raphe	Cluster of nuclei participating in regulation of hormones
Peripheral Clock	Cellular oscillators outside the central pacemaker
Period	Core clock gene that produces proteins PER1, PER2, PER3
Periodicity	Quality of being periodic, Tendency to appear periodically
Rev-erb	Core clock gene that produces proteins REV-ERBα and REV-ERBβ
Ror	Core clock gene that produces proteins RORα , RORβ and RORγ
SCN	Suprachiasmatic nucleus
Thalamic intergeniculate leaflet	A part of the brain that participates in regulation of circadian function
Transcriptome	set of all RNA transcripts
Zeitgeber	External or environmental cue that synchronizes biological processes

1 Introduction

The circadian rhythm is a 24-hour cycle that is influenced by numerous biochemical factors and the diurnal light cycle. This rhythm impacts a multitude of biological processes in organisms, spanning from bacteria to animals. Circadian oscillators, which are indicative genes that exhibit periodic patterns associated with the day-night cycle, emerge from intricate interactions between genetic elements that vary among different organisms (Saini et al., 2019). Despite these variations, there are some fundamental similarities across species. The most well-known mammalian circadian core gene pair, the CLOCK:BMAL1 cistrome, is further examined in this bachelor's thesis. Analogous core genes can be found in all organism domains. The circadian rhythm is connected to several crucial biological processes, including metabolism, behavior, and aging (Zhang et al., 2014).

The regulation of the circadian rhythm is governed by core molecular mechanisms, as depicted in Fig. 2. Researchers have devoted decades to elucidating this mechanism. As reported by Moreira et al. (2018), the understanding of circadian rhythms in mammals can be traced back to the 1940s. The same article reveals that the PER-protein cycle was identified by Nobel laureates Young, Hall, and Rosbash in the 1980s, while another research group discovered the CLOCK gene. Subsequently, the BMAL1 gene and more complex factors were identified.

Progress in genomics and molecular biology has furnished tools for examining circadian rhythms at the molecular level. The employment of sequencing technologies and bioinformatics methodologies has enabled researchers to pinpoint circadian oscillators and gain a deeper comprehension of their roles in sustaining biological processes (Hughes et al., 2017). Within the scope of this thesis, the utilization of Gaussian process-based techniques has demonstrated promise in detecting oscillations originating from circadian biological sources using publicly accessible molecular data (Wilkinson et al., 2019).

The aims of this thesis are (1) review the literature regarding circadian oscillators in human (2) try to identify circadian oscillators from public data using a Gaussian process based approach and (3) compare the analysis results with conclusions reported in the literature.

2 Background

This chapter provides necessary background information on circadian rhythm in general, circadian oscillators and their functions and lastly on the Gaussian method used for analysis

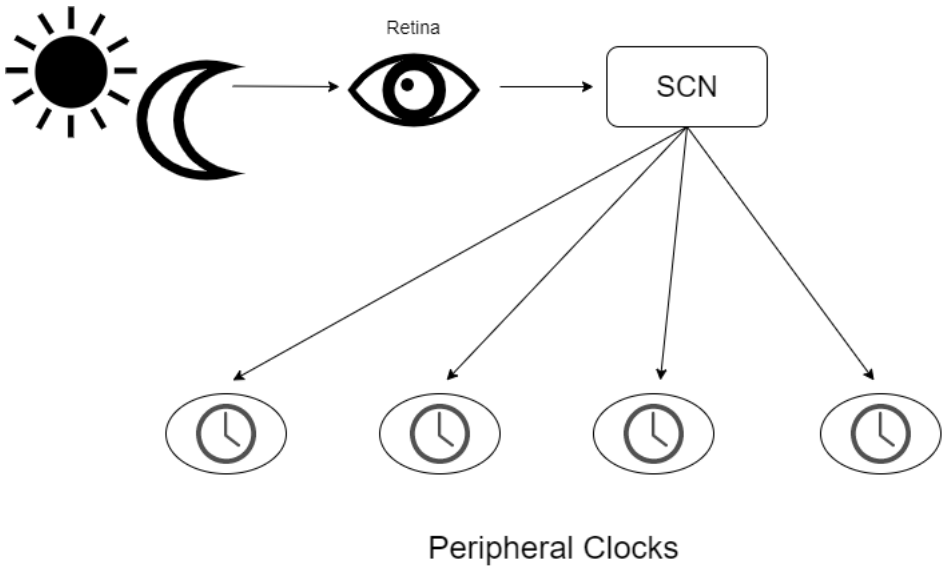


Figure 1: Simplified scheme of the hierarchical structure of complex circadian systems in mammals

2.1 Physiology of the circadian rhythm

Circadian rhythm is a 24-hour cycle in different organisms based on the day/light cycle of the surrounding environment. All eukaryotes share similar basic processes to upkeep the circadian rhythm. More complex creatures e.g. mammals, the processes get likewise more complex (O-Neill et al., 2011).

Mammalian circadian systems exhibit a hierarchical structure in which multiple layers of circadian oscillators constitute the comprehensive circadian framework (O-Neill et al., 2011). This hierarchical organization, illustrated in Fig. 1, introduces the notion of *leader* and *follower* oscillators, with the latter being synchronized by the former (Saini et al., 2019). In mammals the central oscillator, the suprachiasmatic nucleus (SCN), is located in the hypothalamus within the brain. The SCN synchronizes peripheral clocks either directly through autonomic neural connections and hormones, as depicted in Fig. 2, or indirectly by regulating body temperature and feeding patterns (Husse et al., 2015). These external or internal cues, e.g. sunlight or body temperature, are referred to as zeitgebers. The activity and processes of the SCN are influenced by light levels detected by the retinas and, indirectly, by non-photic zeitgebers from the thalamic intergeniculate leaflet and midbrain raphe. Peripheral clocks can also be affected by light levels and external zeitgebers without input from the SCN (Husse et al., 2015).

Besides the central role of the SCN, peripheral clocks are located in various tissues and organs, playing a crucial part in maintaining the circadian rhythm (Bass and Takahashi, 2010). Although the peripheral clocks are synchronized by the SCN, they can also be

influenced by external factors such as feeding time, temperature, and hormonal changes (Mezhnina et al., 2022). Since most peripheral clocks are tissue-specific, they contribute to the regulation of local processes and metabolism, enabling the organism to better adapt to daily variations in environmental conditions and energy requirements (Bass and Takahashi, 2010).

The physiological significance of circadian rhythms is also intimately linked to their entrainment by zeitgebers. Entrainment refers to the synchronization of internal cycles to the external environment based on cues, zeitgebers. In mammals, light serves as the most potent zeitgeber; however, other factors such as temperature, social interactions, and feeding schedules can also impact the circadian system (Mezhnina et al., 2022). This adaptability allows organisms to respond to environmental shifts and maintain optimal physiological function under varying conditions (Froy, 2011).

2.2 Molecular mechanisms of the circadian rhythm

The circadian rhythm in mammals is underpinned by a core molecular mechanism, as depicted in Fig. 2. This mechanism essentially comprises feedback loops between various circadian oscillators, such as CLOCK, BMAL1, PER, CRY, among others. Within these feedback loops, the transcription and degradation of genes create a cycle that aligns with the circadian rhythm. These fundamental clocks are crucial to complex organisms, as they govern and maintain the circadian rhythm as peripheral oscillators in cells and tissues distant from the brain. The foundational circadian clock in mammals is the CLOCK:BMAL1 heterodimeric transcription factor complex (Trott and Menet, 2018). The CLOCK:BMAL1 cistrome accounts for approximately 15% of the rhythmic expression of the transcriptome and predominantly manages the regulation of daily biological processes (Trott and Menet, 2018).

According to Trott and Menet (2018) core functionality of CLOCK:BMAL1 functions as follows, also visualised in Fig. 2:

- CLOCK:BMAL1 binds to DNA rhythmically to transcribe core clock genes *Period, Cryptochrome, Rev-erb, Ror*
- After expression and maturation, CRYs and PERs rhythmically inhibit the CLOCK:BMAL1 transcription factor on-DNA and off-DNA by forming a repressive complex. This means that the formed complex prevents further transcription by CLOCK:BMAL1 and that the complex also binds to CLOCK-BMAL1 when not transcribing DNA.
- REV-ERBs and RORs regulate BMAL1 rhythmically by repressing or activating it's transcription cyclically

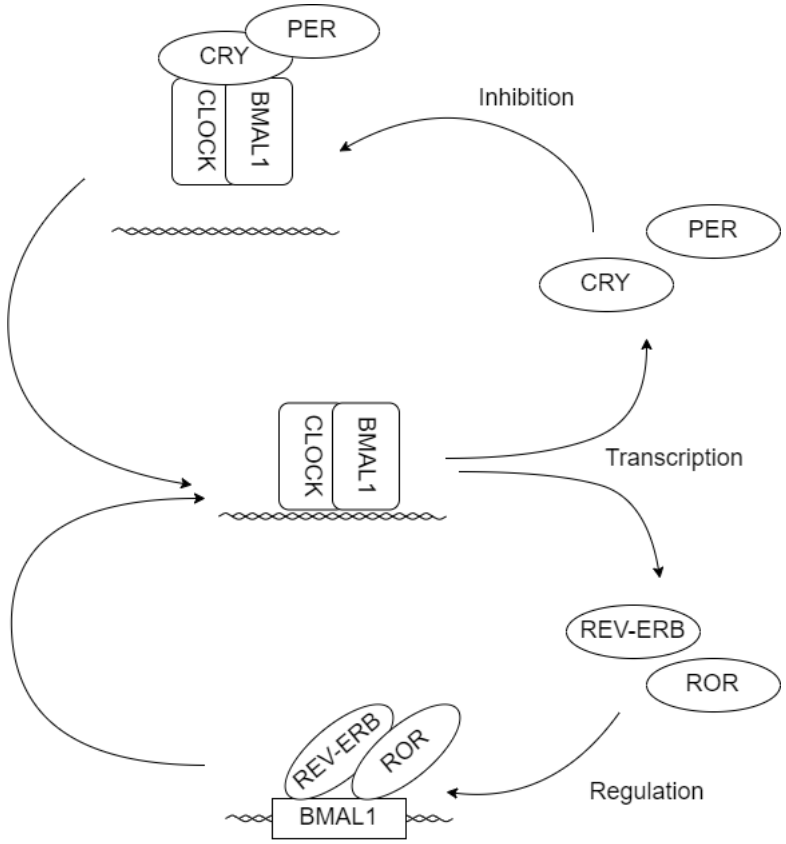


Figure 2: Simplified figure of the CLOCK:BMAL1 circadian clocks process

2.3 Applications of circadian rhythm

The circadian rhythm is crucial to numerous biological processes and diseases, including metabolism and aging. This section offers a brief overview of its influence on metabolism and aging. The genetic expression of circadian genes holds significant relevance in various research areas, as identifying the precise genes responsible for specific phenomena can facilitate a better understanding of their origin or possible treatments.

The circadian rhythm influences metabolism through the liver, the organ responsible for energy metabolism regulation. Research has been conducted on the effects of a calorie-restricted diet in mice, examining its impact on overall health and lifespan (Acosta-Rodríguez et al., 2022; Mezhnina et al., 2022). These studies discovered that aligning caloric restriction and subsequent fasting with the circadian rhythm enhanced the subjects' overall health and longevity. Mezhnina et al. (2022) reported that daily fasting and circadian synchronized feeding increased longevity in male mice by 35%. These effects on lifespan are visualized in Fig. 3. Mezhnina et al. (2022) also investigated the effects of aging and variously timed caloric restrictions on circadian gene expression in the mice's liver. The study revealed that overall circadian gene expression decreases with age, and during the aging process, some genes lose their circadian rhythm, while

some non-circadian genes acquire circadian rhythm. Findings from the Acosta-Rodríguez et al. (2022); Mezhnina et al. (2022) studies suggest that most circadian genes change over time, with only a few being expressed throughout the mouse's life. Similarly, a reduction in gene expression was observed in mice fed during daytime compared to those fed during nighttime.

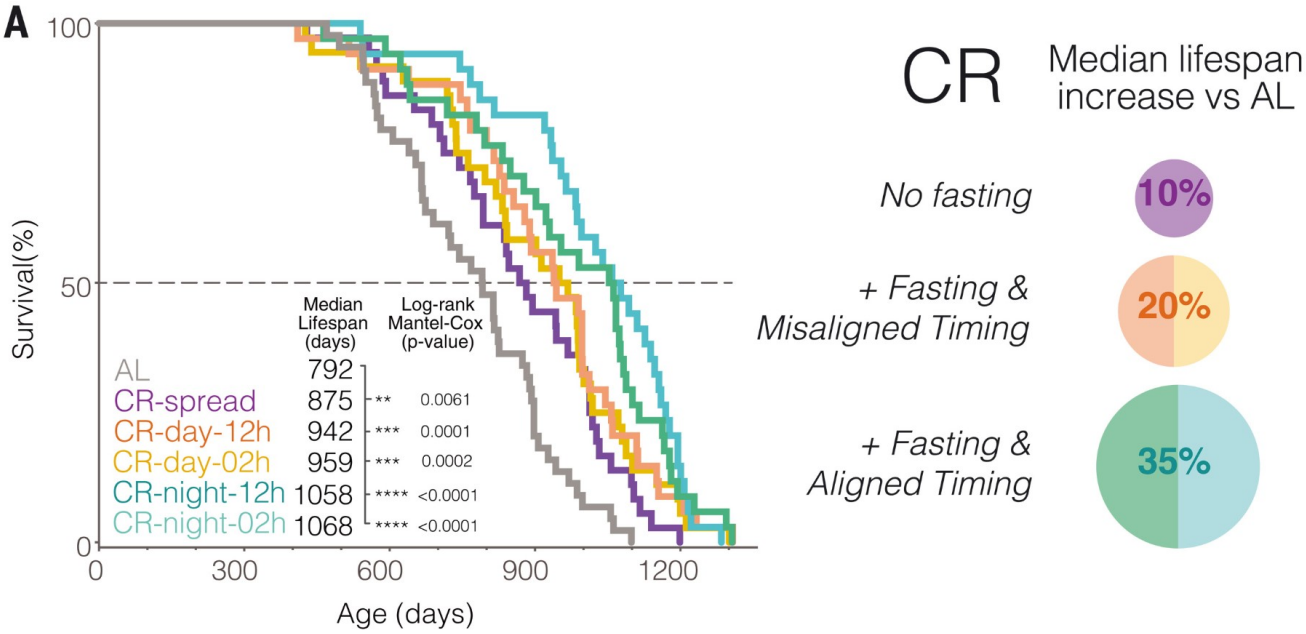


Figure 3: Lifespan of different groups of mice in different feeding conditions (Acosta-Rodríguez et al., 2022)

Acosta-Rodríguez et al. (2022) discovered that circadian genes induced a rhythmic presence of β-hydroxybutyrate in the blood through a transcriptional network. β-hydroxybutyrate, or βOHB, transports energy in the blood during fasting periods. This implies that by properly adjusting feeding and fasting times in accordance with the circadian rhythm, the body can utilize stored energy more effectively through ketogenesis and fatty acid oxidation. While β-OHB levels also depend on diet and caloric intake, the finding linking β-OHB levels to circadian rhythm further emphasizes the significance of the relationship between circadian rhythm and feeding patterns. Circadian rhythms are linked to various diseases, as circadian genes and the overall circadian rhythm are closely connected to metabolism and behavior regulation through hormones. For instance, mutations in circadian genes such as BMAL1 in the liver are associated with an increased risk of liver disease and obesity (Jouffe et al., 2022; Mukherji et al., 2019). As previously mentioned, metabolism and food intake are governed by numerous processes involving circadian genes. If the natural circadian rhythm is disrupted or circadian genes are mutated or inactivated, normal metabolic functions could be affected. Fig. 4 demonstrates that the body weight of BMAL1 knockout mice is significantly lower than

that of wild-type mice. Mice and other mammals have thousands of actively expressed circadian genes in the liver alone (Mezhnina et al., 2022). Thus, disruptions in the circadian rhythm could impact a large number of genes across different organs and be associated with multiple diseases.

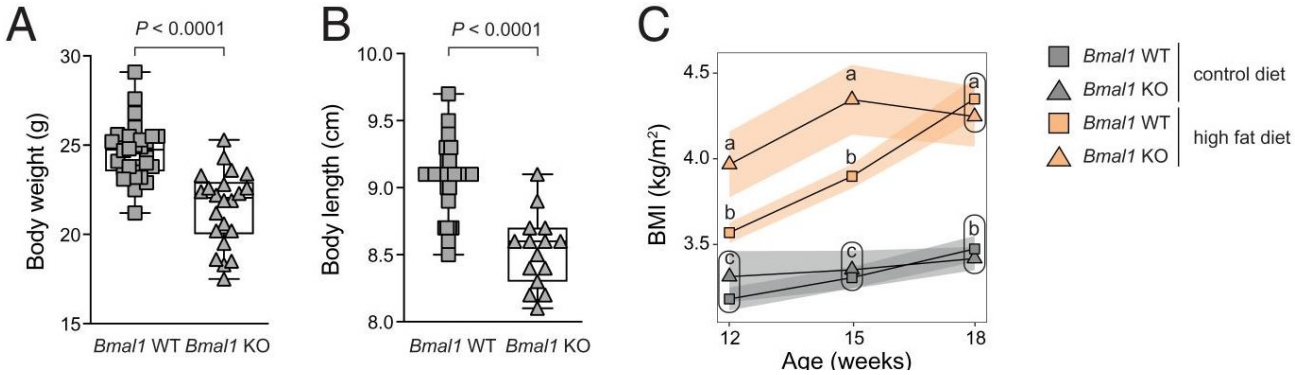


Figure 4: Growth of mice with and without BMAL1 gene knockout (Jouffe et al., 2022)

Aging influences circadian processes by altering the phase of the cycle, reducing the amplitude, and decreasing the number of active circadian genes (Froy, 2011). The reasons behind the decline in circadian activity during aging remain unclear. Changes in the SCN's neurochemical and electrophysiological output have been observed, which could account for some of the reduction (Froy, 2011). These alterations in the SCN's functionality affect its capacity to reset and synchronize peripheral clocks, thereby weakening the overall performance of peripheral clocks. As a result, if the expression of circadian genes could be genetically or medically re-established, it may help mitigate the effects of aging and improve the lifespan of mammals. Consequently, a more comprehensive understanding of circadian genes and their potential for pharmaceutical activation could provide a pathway to decelerate aging.

2.4 Gaussian processes for detecting periodicity

Nature often displays its phenomena in periodic behaviour. Gaussian processes could be used for analysis of data from natural sources with periodic tendencies. Gaussian processes are fundamentally non-linear regression method where the non-linear terms are coded by kernels (Bishop, 2006). Gaussian processes parameters are found by fitting the GP model to the observed data. To identify if a gene possesses circadian rhythm, both a periodic GP model and a null model are fitted to the time series data of a gene. The log-likelihoods of these two models are compared to determine the periodicity.

In the context of circadian rhythms, periodicity is a crucial aspect to consider as it reflects the underlying biological processes. By applying Gaussian processes to gene expression

data, scientists can gain knowledge of the molecular mechanisms governing circadian rhythms and identify potential targets for further investigation (Zhang et al., 2014). This is particularly important for understanding the intricate relationships between circadian expressions and metabolism and cell replication (Bass and Takahashi, 2010)

Kristjanson Duvenaud (2014) showcases an example of a Gaussian process model with a periodic covariance function as the kernel, as follows:

$$y(x) = y_p(x) + \epsilon \tag{1}$$

Here $y(x)$ are the target values, i.e. expression levels of a given gene, y_p denotes the non-linear periodic function, ϵ is the Gaussian noise variable with zero mean and variance of σ_n^2. y_p is defined by the following periodic kernel. :

$$k_p(x, x') = \sigma^2 exp(-\frac{2\sin^2(\pi|x - x'|/p)}{l^2}), \tag{2}$$

where x and x' denote the time points, l the length between periods for the periodicity, p is the period, σ^2 is the variance.

3 Methodology

This section further describes the methods used for analysis of the gathered data from the programming language used to the details of Gaussian process.

3.1 Data used for analysis

Gene expression data from Acosta-Rodríguez et al. (2022) is used in this thesis. The study provides a large dataset of genes collected from mice under different feeding conditions. Only data gathered from mice which were fed *ad libitum* is used, as caloric restriction or involuntary fasting can skew the data. As Illustrated in Fig. 5. the data consists of 6 months and 19 months old male mice, and datapoints were gathered from liver samples 4 hours apart from 2 separate mice, spanning 44 hours. In the Fig. 5. the A and B points denote the separate mice for each time point, from which the liver samples are taken. The whole data file contains 20496 rows of data, of which half are for the 6 months old mice and rest of the 19 month old. Thus 10246 different expressed genes are being analyzed.

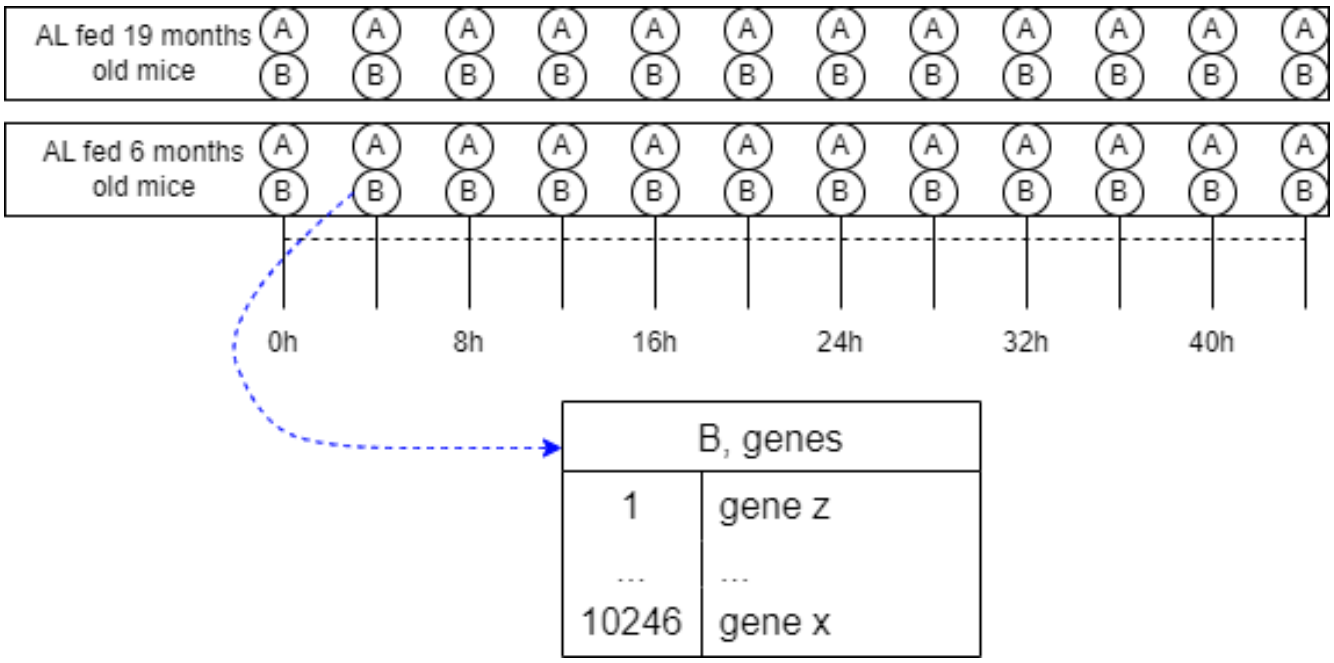


Figure 5: Visualisation of the data available from Acosta-Rodríguez et al. (2022) study

Data is prepared for analysis as follows:

- The original data is in this format, where each row represent a gene. Each data point of a row is the amount of expression of the respective gene. The program also automatically normalizes each gene so, that variance is 1 and mean is 0. A and B denote the different mice liver samples for each time point.

$$0h_A, 0h_b, 4h_A, 4h_B ... 44h_A, 44h_B \tag{3}$$

- This is then formatted so, that timepoint pairs are separated and B timepoints come after all A timepoints.

$$0h_A, 4h_A, 8h_A, ... 44h_A, 0h_B, ... 40h_B, 44h_B \tag{4}$$

- This leaves the final amount of datapoints per gene to 24 and the time range to 0h - 92h.

3.2 Periodic model and null model

The Gaussian process code is provided by thesis instructor Dr. Lu Cheng. The model for running analysis on individual genes is setup as follows:

- The starting parameters for analysis are defined as tensors:

- p, is predefined as 24 for 24h periodicity that this model aims to find.
- l, the length-scale as a random float ranging from 0 to 1, divided by $p/8$.
- σ_f^2, as follows: $x\ Unif(0,1)$, for $(x * 0.5 + 0.5)^2$
- σ_n^2 , as follows: $x\ Unif(0,1)$, for $(x * 0.09 + 0.01)^2$

Parameters were chosen based on prior knowledge.

- A Periodic model is used:

$$\sigma_f^2 * exp(-2 * \frac{sin^2(\pi * |(x - x')/p|)}{l^2}) \tag{5}$$

- The model is iterated for 100 epochs by maximizing the models Log-Likelihood
- The steps above are repeated for 10 times, and at last the best possible result is chosen and then saved as the result for the gene in question.

A null model is setup in a similar manner as described above, but the p is set as a very high value so the result produced cannot find any peridocity in that large range. This null model is used as a comparison for the periodic model to determine its accuracy.

3.3 Analysis with gaussian process

The analysis is ran on eahc gene with the models setup as described above. As previously mentioned, the null model is also produced with a periodic model, but with the period set so high, it simulates a random noise. Each genes analysed are displayed as graphs and their key values are printed and stored in a summary file. The accuracy of analyses can be interpreted from the Log-Likelihood-Ratio, or LLR.

The value of LLR is calculated as follows, N denotes the number of time points for a gene, L_n the Log-likelihood of the null model and L_p the log-likelihood for the periodic kernel:

$$\frac{L_p - L_n}{N} \tag{6}$$

LLR compares the circadian model and the null model for each genes expression. For positive LLR values the periodic model is preferred, vice versa for negative the null model is favoured. Thus as genes that have negative LLR's favour the null model, they can't be deemed as circadian. These genes may still show periodic patterns, but they are not strong enough to be accepted as circadian.

4 Results

This chapter outlines the outcomes of the analysis, which entailed scrutinizing 10,246 genes from a dataset comprising expression data for two separate age groups: 6 months old and 19 months old mice. We applied the Gaussian model analysis techniques mentioned earlier to assess the periodicity of each gene's expression and employed the log-likelihood ratio to categorize genes as circadian or non-circadian. These results are compared with those reported in the study by Acosta-Rodríguez et al. (2022). The subsequent sections delineate the findings in depth, explicating the outcomes for each age group and discussing the implications of the identified trends.

4.1 Circadian genes in the test data

The findings reveal that, of the genes analyzed, 2.9% (298 out of 10,246) can be considered circadian genes for the 6-month-old male mice, while 2% (206 out of 10,246) are circadian genes for the 19-month-old mice. This indicates that these genes have positive LLR values, suggesting a circadian pattern in their expression.

Genes that exhibit no periodicity, and are therefore not classified as circadian genes, have LLR values below zero. For instance, the gene LGSF5, as depicted in Fig. 6, demonstrates that the data points do not adhere to any periodic function. Nonetheless, the model attempts to identify the closest possible periodic function that corresponds to the given data. The LLR value, which is relatively negative in this case, further confirms this observation. This additional insight highlights the importance of considering both the visual representation and LLR values in determining whether a gene exhibits circadian characteristics or not.

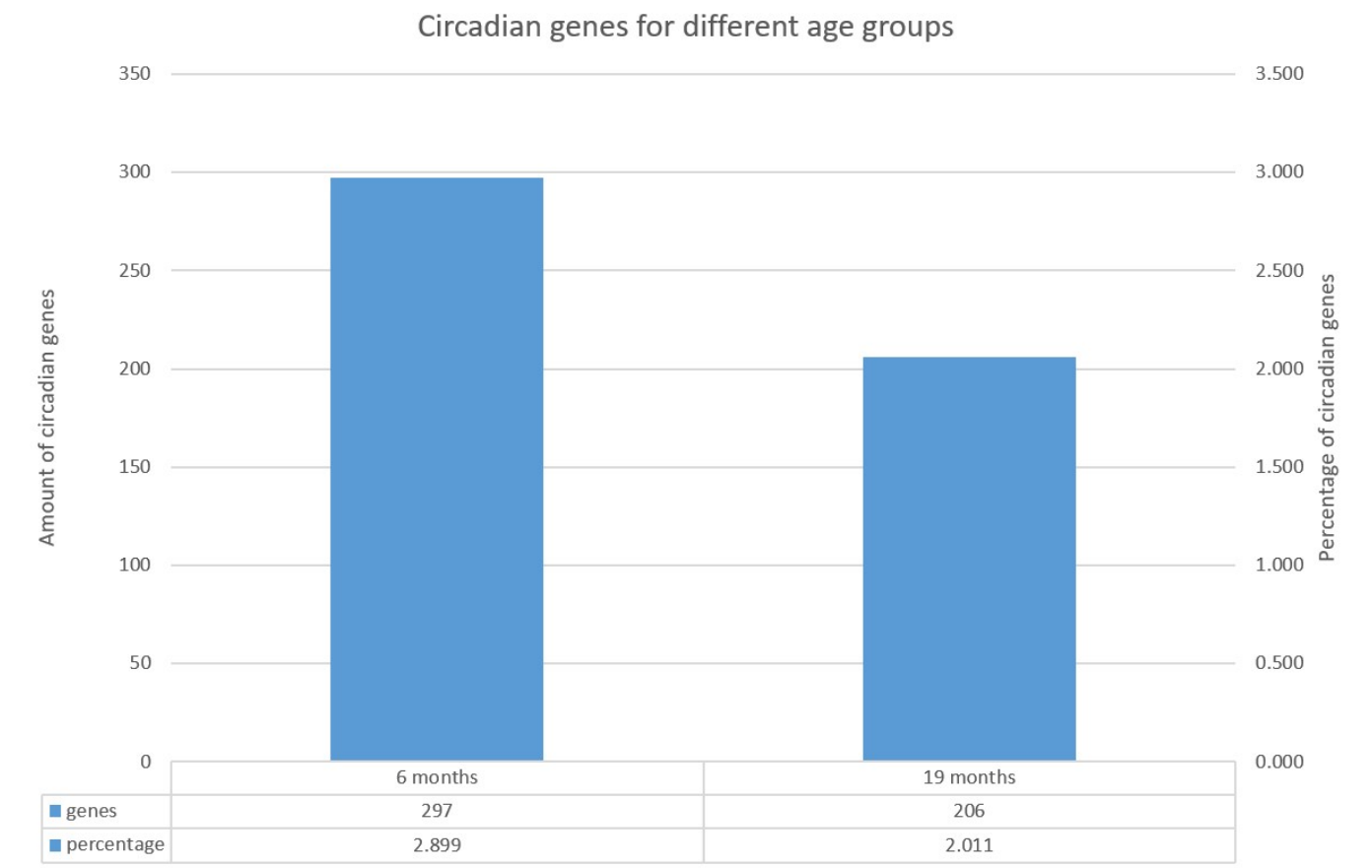


Figure 6: Amount of found genes for each age group

The results demonstrate a significant difference in the expression of circadian genes between the two age groups. The older group seems to express only 2/3 the number of circadian genes compared to the younger group. While the 1/3 decrease may not be entirely consistent with other evidence, it is in line with similar findings reported in the relevant literature, suggesting that circadian expression declines as organisms age (Mezhnina et al., 2022; Froy, 2011; Acosta-Rodríguez et al., 2022).It worth noting that the number of detected circadian genes is smaller than that reported (n=1718, for the 6 months old group and n=1507 for the 19 month old group of mice) in Acosta-Rodríguez et al. (2022). This is because different software is used for circadian gene identification. As will be shown in the sequel, circadian genes identified by GP method do exhibit a periodic change. The result hints the GP method uses a much more strict criterion than the method used in Acosta-Rodríguez et al. (2022).

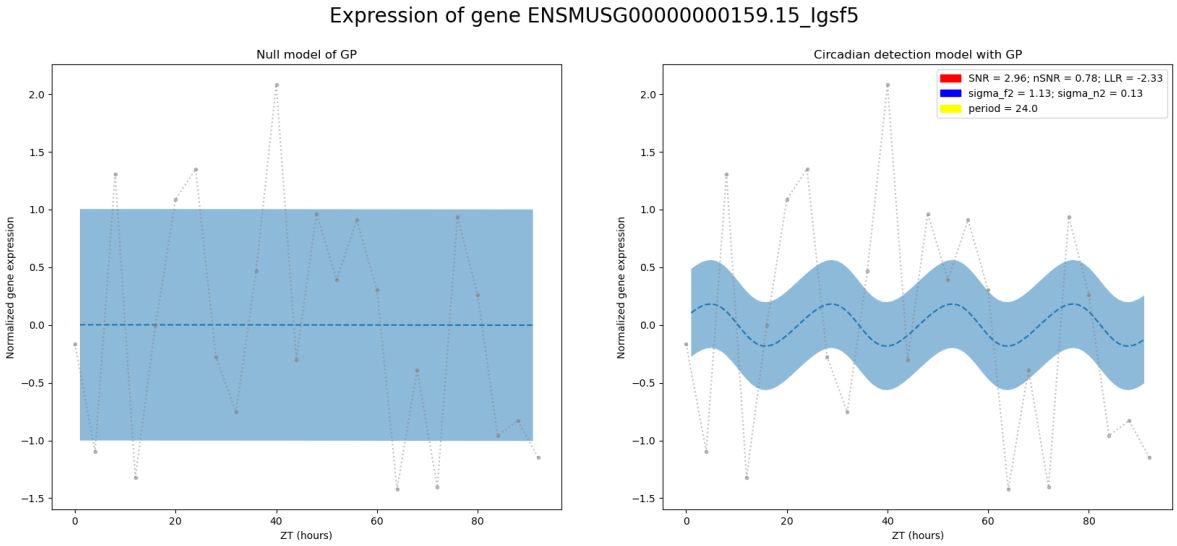


Figure 7: Expression of LGSF5 gene in 6-month-old mice, LLR -2.33

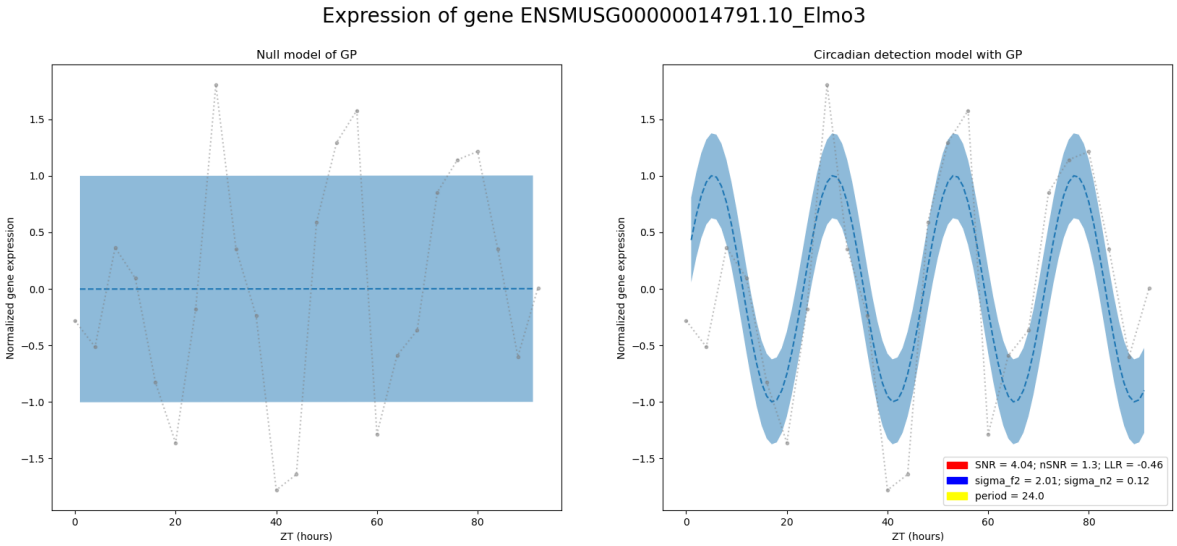


Figure 8: Expression of ELMO3 gene in 6month old mice, LLR -0.46

In mammals, a large proportion of genes do not display any detectable periodicity. As illustrated in Fig. 7, the gene presents a notably negative LLR, which substantiates the absence of periodic expression. The lack of periodic patterns is evident in the graphical representation, as the data points are dispersed throughout the visual field in a seemingly random manner. Such an example epitomizes a non-circadian gene. The findings of the conducted analysis indicate that the majority of genes classified as non-circadian share this characteristic.

In some instances, a gene might display periodic behavior, but its LLR value remains negative. These borderline cases pose interpretive challenges, as they seem to fit well

with the function but cannot be conclusively classified as circadian genes. For instance, gene illustrated Fig. 8, ELMO3, might represent a circadian gene; however, with the available data, it is not possible to determine its status with certainty. Such borderline cases may arise due to poor data quality, limitations in the experimental design, or the inherently erratic behavior of gene expression. A more comprehensive experimental approach, involving additional time points or higher-resolution measurements, could help clarify these ambiguous cases and contribute to a more accurate assessment of a gene's circadian nature. Furthermore, exploring the potential impact of environmental factors or interactions with other genes could offer valuable insights into the underlying mechanisms that influence circadian gene expression, ultimately enriching our understanding of this complex biological process.

4.2 Example regression results

This section examines example results obtained from the analysis. Fig. 9, panel A, presents example regressions of known core clock genes BMAL1, RORC, and CRY1. These genes were accurately identified as circadian, corroborating findings in other literature. Moreover, these genes are shared circadian genes between the 6-month and 19-month age groups.

Fig. 9, panel B, displays three examples of non-circadian genes with varying log-likelihood ratios. The RNPC3 gene exhibits no periodicity in its expression, resulting in a significantly negative LLR. The SSH3 gene demonstrates some periodicity in its regression but also deviates considerably from the calculated periodic function. The CAPRIN1 gene exhibits apparent periodicity; however, imperfections or errors in sample analysis might have affected the test results, leading to skewed data. As mentioned in Konopka and Rooman (2010), CAPRIN1 and SSH3 could also be circadian genes, but their expression data might be influenced by the periodicity of other genes, causing disruptions in the expression patterns.

Differences in results for the same gene based on the sample age group also exist, as illustrated in Fig. 9, panel C. In this case, NR1D1 is a core clock gene considered circadian for both groups. Using the chosen Gaussian analysis method, the gene expression values are normalized for analysis, so it is not possible to compare the amplitude or value of the expression in the genes. Analyzing the differences in amplitude would have allowed us to determine the extent to which the expression of the gene decreases or increases as the subjects grow, providing a deeper understanding of the relationship between aging and circadian gene expression.

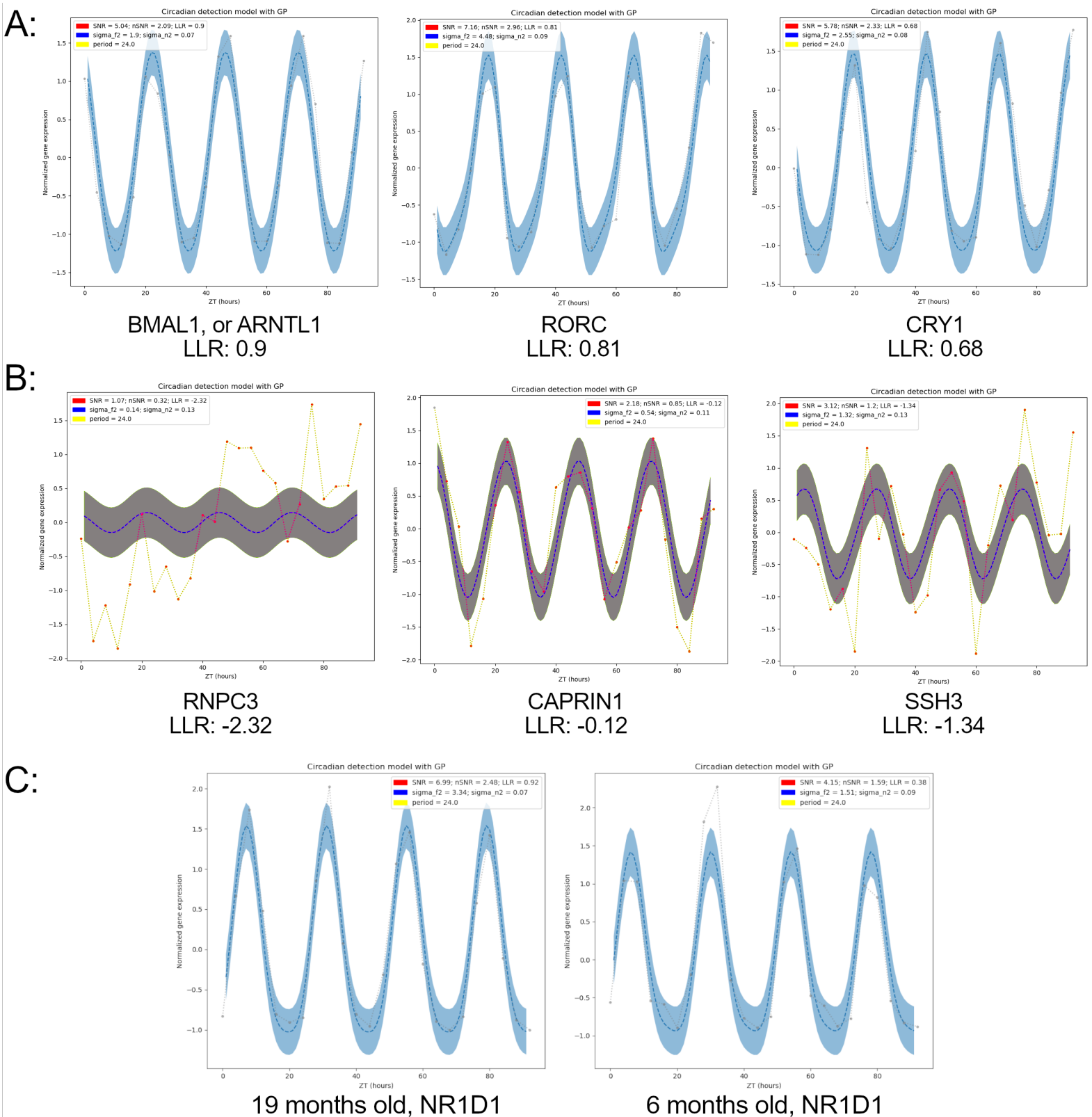


Figure 9: Expression of various different genes: A (circadian genes) and B (non-circadian genes) are for 6 months old mice, C compares a NR1D1 for both age groups.

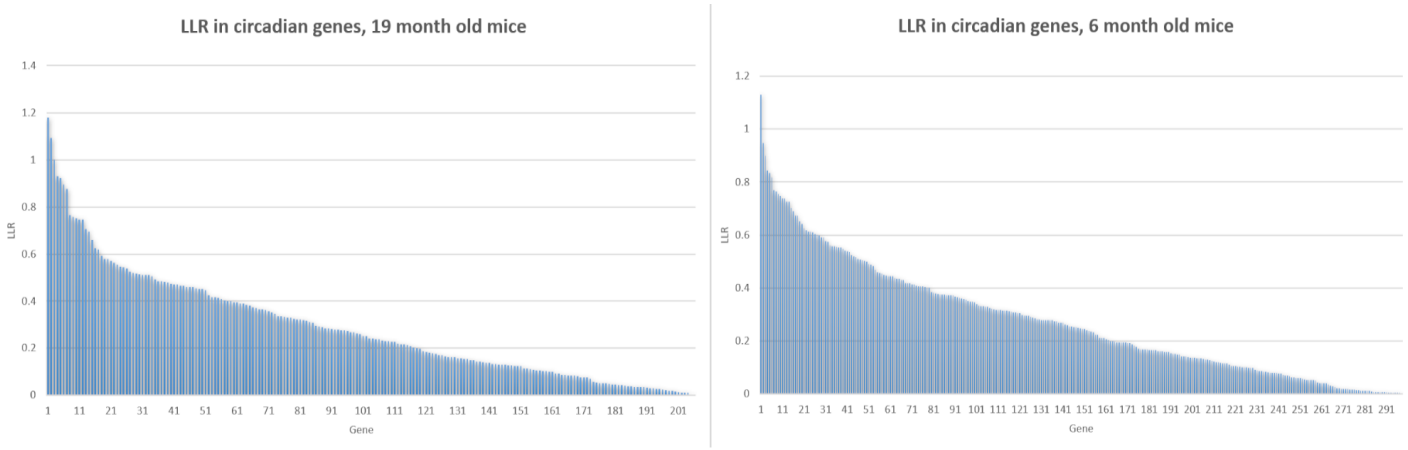
4.3 Comparison to literature

The results obtained from the performed analysis demonstrate that the model performs reasonably well. The most well-known core clock genes are all clearly identifiable as circadian genes, supporting the validity of the approach taken in this thesis.

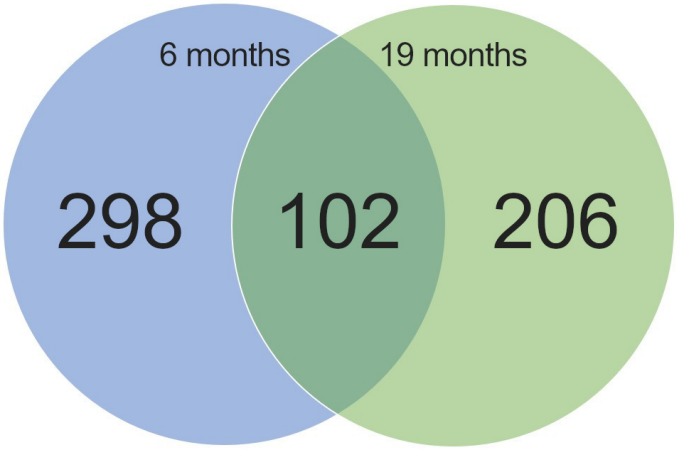
The results obtained from the analysis align with some key findings of the source study Acosta-Rodríguez et al. (2022), although the number of circadian genes identified is significantly lower. As shown in Fig. 10, the younger group exhibits circadian expression in 298 genes, while the older group does so in 206 genes. In contrast, Acosta-Rodríguez et al. (2022) identified 1,718 circadian genes for the younger group and 1,507 for the older group. This discrepancy may result from several factors. The analysis employed in this thesis exclusively attempts to identify circadian genes with a 24-hour phase. In the Acosta-Rodríguez et al. (2022) study, the phase was allowed to vary, resulting in a higher number of identified genes. Furthermore, they utilized periodic analysis algorithms, such as ARSER, JTK_CYCLE, and RAIN, which analyze and identify circadian genes using different methods than a Gaussian model.

Another factor to consider when comparing the results is the possible influence of noise in the gene expression data. Noise can stem from various sources, including biological variability, experimental errors, or technical limitations in the data acquisition process. This noise may lead to false positives or negatives in the identification of circadian genes, potentially contributing to the discrepancies observed between the two studies

It is essential to consider these methodological differences when comparing the results of this thesis with those reported in the literature. Future studies might benefit from exploring alternative analytical approaches and incorporating a broader range of phase values, potentially leading to a more comprehensive understanding of circadian gene expression and its relationship with aging and other biological processes. Furthermore, accounting for potential sources of noise in the data can help improve the reliability and accuracy of the analysis, enabling more meaningful comparisons with existing literature.



A: The Log-likelihood-Ratio for genes identified as circadian, 6 months and 19months old mice respectively



A: a Venn diagram displaying circadian genes found for each age group and the circadian genes shared by both

Figure 10: Differences in gene expressions for the two age groups

5 Conclusion

This thesis aimed to (1) review the literature regarding circadian oscillators in humans, (2) identify circadian oscillators from public data using a Gaussian process-based approach, and (3) compare the analysis results with conclusions reported in the literature. The study successfully met these objectives, providing insights into the genetic base of circadian rhythms, discussing the applicability of research conducted, and highlighting the potential for further investigation in this area.

By analyzing the expression data of 10,246 genes in 6-month-old and 19-month-old male mice, the study identified 298 circadian genes in the younger group and 206 circadian genes in the older group. This observation supports the literature, which suggests a

decline in circadian gene expression with age (Mezhnina et al., 2022; Froy, 2011; Acosta-Rodríguez et al., 2022). Moreover, the analysis demonstrated the effectiveness of the Gaussian process in detecting periodicity in gene expression data, despite some limitations in borderline cases.

The findings, however, showed a discrepancy in the number of circadian genes compared to the study by Acosta-Rodríguez et al. (2022). This difference can be attributed to the methodology, as the Gaussian process-based approach focused on identifying genes with a 24-hour phase, while the original study allowed for varying phase lengths. Additionally, the algorithms used in the original study (ARSER, JTK_CYCLE, and RAIN) employed different methods for identifying circadian genes.

In light of these findings, future research could focus on exploring alternative analytical approaches, incorporating a broader range of phase values, and examining the role of environmental factors and lifestyle choices in the regulation of circadian gene expression. This may contribute to a more comprehensive understanding of the relationship between circadian rhythms, aging, and overall health.

In conclusion, this thesis contributes to the understanding of circadian rhythms and their molecular mechanisms by successfully employing a Gaussian process-based approach to identify circadian genes from expression data. This thesis highlights the importance of circadian rhythms in the regulation of various biological processes and demonstrates the potential of Gaussian processes in detecting periodicity in molecular data. While some discrepancies exist between the findings and the original study, these differences emphasize the need for further research in developing analysis methods, focusing on the optimization of models and the exploration of other factors influencing biology and health.

References

Acosta-Rodríguez Victoria , Rijo-Ferreira Filipa , Izumo Mariko , Xu Pin , Wight-Carter Mary , Green Carla B. and Takahashi Joseph S. . Circadian alignment of early onset caloric restriction promotes longevity in male C57BL/6J mice. *Science (New York, N.Y.)*, 376(6598):1202, 6 2022. ISSN 10959203. doi: 10.1126/SCIENCE.ABK0297. URL /pmc/articles/PMC9262309//pmc/articles/PMC9262309/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC9262309/.

Bass Joseph and Takahashi Joseph S. . Circadian Integration of Metabolism and Energetics. *Science (New York, N.Y.)*, 330(6009):1354, 2010. ISSN 00368075. doi: 10.1126/SCIENCE.1195027. URL /pmc/articles/PMC3756146//pmc/articles/PMC3756146/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3756146/.

Bishop Christopher M. . *Pattern Recognition and Machine Learning*. 6 Springer Science+Business Media, LLC, 2006. ISBN 978-0387-31073-2.

Froy Oren . Circadian rhythms, aging, and life Span in mammals. *Physiology*, 26(4):225–235, 8 2011. ISSN 15489213. doi: 10.1152/PHYSIOL.00012.2011/ASSET/IMAGES/LARGE/PHY0041100740003.JPEG. URL https://journals.physiology.org/doi/10.1152/physiol.00012.2011.

Hughes Michael E , Hogenesch John B , Abruzzi Katherine C , Allada Ravi , Anafi Ron , Arpat Alaaddin Bulak , Asher Gad , Baldi Pierre , De Bekker Charissa , Bell-Pedersen Deborah , Blau Justin , Brown Steve , Ceriani M Fernanda , Chen Zheng , Chiu Joanna C , Cox Juergen , Crowell Alexander M , Debruyne Jason P , Dijk Derk-Jan , Ditacchio Luciano , Doyle Francis J , Duffield Giles E , Dunlap Jay C , Eckel-Mahan Kristin , Esser Karyn A , Fitzgerald Garret A , Forger Daniel B , Francey Lauren J , Fu Ying-Hui , Gachon Frédéric , Gatfield David , De Goede Paul , Golden Susan S , Green Carla , Harer John , Harmer Stacey , Haspel Jeff , Hastings Michael H , Herzel Hanspeter , Herzog Erik D , Hoffmann Christy , Hong Christian , Hughey Jacob J , Hurley Jennifer M , De La Iglesia Horacio O , Johnson Carl , Kay Steve A , Koike Nobuya , Kornacker Karl , Kramer Achim , Lamia Katja , Leise Tanya , Lewis Scott A , Li Jiajia , Li Xiaodong , Liu Andrew C , Loros Jennifer J , Martino Tami A , Menet Jerome S , Merrow Martha , Millar Andrew J , Mockler Todd , Naef Felix , Nagoshi Emi , Nitabach Michael N , Olmedo Maria , Nusinow Dmitri A , Ptáček Louis J , Rand David , Reddy Akhilesh B , Robles Maria S , Roenneberg Till , Rosbash Michael , Ruben Marc D , Rund Samuel S C , Sancar Aziz , Sassone-Corsi Paolo , Sehgal Amita , Sherrill-Mix Scott , Skene Debra J , Storch Kai-Florian , Takahashi Joseph S , Ueda Hiroki R , Wang Han , Weitz Charles , Westermark Pål O , Wijnen Herman , Xu

Ying , Wu Gang , Yoo Seung-Hee , Young Michael , Zhang Eric Erquan and Zielinski Tomasz . Guidelines for Genome-Scale Analysis of Biological Rhythms. *JOURNAL OF BIOLOGICAL RHYTHMS*, 32(5):380–393, 2017. doi: 10.1177/0748730417728663. URL https://doi.org/10.1177/0748730417728663.

Husse Jana , Eichele Gregor and Oster Henrik . Synchronization of the mammalian circadian timing system: Light can control peripheral clocks independently of the SCN clock. *BioEssays*, 37(10):1119–1128, 10 2015. ISSN 1521-1878. doi: 10.1002/BIES.201500026. URL https://onlinelibrary.wiley.com/doi/full/10.1002/bies.201500026https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201500026https://onlinelibrary.wiley.com/doi/10.1002/bies.201500026.

Jouffe Céline , Weger Benjamin D. , Martin Eva , Atger Florian , Weger Meltem , Gobet Cédric , Ramnath Divya , Charpagne Aline , Morin-Rivron Delphine , Powell Elizabeth E. , Sweet Matthew J. , Masoodi Mojgan , Uhlenhaut N. Henriette and Gachon Frédéric . Disruption of the circadian clock component BMAL1 elicits an endocrine adaption impacting on insulin sensitivity and liver disease. *Proceedings of the National Academy of Sciences of the United States of America*, 119(10), 3 2022. ISSN 10916490. doi: 10.1073/PNAS.2200083119/SUPPL{_}FILE/PNAS.2200083119.SD06.XLSX. URL https://www.pnas.org/doi/abs/10.1073/pnas.2200083119.

Konopka Tomasz and Rooman Marianne . Gene expression model (in)validation by Fourier analysis. *BMC systems biology*, 4(1):123, 9 2010. ISSN 17520509. doi: 10.1186/1752-0509-4-123/FIGURES/4. URL https://bmcsystbiol.biomedcentral.com/articles/10.1186/1752-0509-4-123.

Kristjanson Duvenaud David . *Automatic Model Construction with Gaussian Processes Declaration*. Ph.D. thesis, 2014.

Mezhnina Volha , Ebeigbe Oghogho P. , Velingkaar Nikkhil , Poe Allan , Sandlers Yana and Kondratov Roman V. . Circadian clock controls rhythms in ketogenesis by interfering with PPARα transcriptional network. *Proceedings of the National Academy of Sciences of the United States of America*, 119(40), 10 2022. ISSN 10916490. doi: 10.1073/PNAS.2205755119/SUPPL{_}FILE/PNAS.2205755119.SAPP.PDF. URL https://www.pnas.org/doi/abs/10.1073/pnas.2205755119.

Moreira Ayrton Custodio , Antonini Sonir Rauber and Castro de Margaret . A sense of time of the glucocorticoid circadian clock: from the ontogeny to the diagnosis of Cushing's syndrome. , 4 2018. URL https://www.researchgate.net/profile/Margaret-Castro-2/publication/324549940_MECHANISMS_IN_ENDOCRINOLOGY_A_sense_of_time_of_the_glucocorticoid_circadian_clock_from_the_ontogeny_

to_the_diagnosis_of_Cushing's_syndrome/links/5adb6f57a6fdcc29358a2cb7/MECHANISMS-IN-ENDOCRINOLOGY-A-sense-of-time-of-the-glucocorticoid-circadian-clock-i pdf.

Mukherji Atish , Bailey Shannon M , Staels Bart and Baumert Thomas F . The circadian clock and liver function in health and disease. *Journal of Hepatology*, 71:200–211, 2019. doi: 10.1016/j.jhep.2019.03.020.

O-Neill John S. , Van Ooijen Gerben , Dixon Laura E. , Troein Carl , Corellou Florence , Bouget Franois Yves , Reddy Akhilesh B. and Millar Andrew J. . Circadian rhythms persist without transcription in a eukaryote. *Nature*, 469(7331):554, 1 2011. ISSN 00280836. doi: 10.1038/NATURE09654. URL /pmc/articles/PMC3040569//pmc/articles/PMC3040569/?report=abstracthttps://www.ncbi.nlm.nih.gov/pmc/articles/PMC3040569/.

Saini Reena , Jaskolski Mariusz and Davis Seth J. . Circadian oscillator proteins across the kingdoms of life: structural aspects. *BMC Biology 2019 17:1*, 17(1):1–39, 2 2019. ISSN 1741-7007. doi: 10.1186/S12915-018-0623-3. URL https://bmcbiol.biomedcentral.com/articles/10.1186/s12915-018-0623-3.

Trott Alexandra J. and Menet Jerome S. . Regulation of circadian clock transcriptional output by CLOCK:BMAL1. *PLOS Genetics*, 14(1), 1 2018. ISSN 1553-7404. doi: 10.1371/JOURNAL.PGEN.1007156. URL https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1007156.

Wilkinson Maxim , Maidstone Robert , Loudon Andrew , Blaikley John , White Iain R. , Singh Dave , Ray David W. , Goodacre Royston , Fowler Stephen J. and Durrington Hannah J. . Circadian rhythm of exhaled biomarkers in health and asthma. *European Respiratory Journal*, 54(4), 10 2019. ISSN 0903-1936. doi: 10.1183/13993003.01068-2019. URL https://erj.ersjournals.com/content/54/4/1901068https://erj.ersjournals.com/content/54/4/1901068.abstract.

Zhang Ray , Lahens Nicholas F. , Ballance Heather I. , Hughes Michael E. and Hogenesch John B. . 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*, 111(45):16219–16224, 10 2014. ISSN 10916490. doi: 10.1073/PNAS.1408886111/SUPPL{_}FILE/PNAS.1408886111.SD05.XLS. URL https://www.pnas.org/doi/abs/10.1073/pnas.1408886111.

List of Figures