A Lab Module Report for M. Sc. in Computational Biology

Transcriptomic Analysis of Circadian Regulation in Podocytes

Submitted by,

Santhosh Gojjam Kantharaju

Under the Guidance of,

Dr. Martin Kann

Prof. Dr. Andreas Beyer

Department of Biology

University of Cologne, Germany

Contents

1 Abstract

The kidney plays a vital role in filtering blood and maintaining fluid and ion homeostasis. Approximately 90% of chronic kidney disease cases are linked to glomerular diseases, with the epithelial podocyte cell at the forefront of these pathological changes. Podocytes are specialized cells in the kidney that are terminally differentiated and essential for glomerular filtration. Recent research into circadian rhythms has highlighted the significance of the biological clock in regulating a wide array of podocytic genes. Furthermore, studies have demonstrated that disturbances in circadian rhythm within podocytes can lead to alterations in the glomerular filtration rate and podocyte injury. Disruption of circadian clocks has detrimental effects on both the structure and function of podocytes. Changes in circadian rhythm can be quantified using the Circadian Rhythm Disruption (CRD) score, which may impact podocyte function.

In our study, we identified circadian genes from mouse RNA sequencing data utilizing the cglmm() function in R. We applied these genes to calculate CRD scores in human Xenium data, also in R. Subsequently, we conducted correlation analyses to investigate the relationships between CRD scores and podocyte damage scores (PDS), as well as age, estimated Glomerular Filtration Rate (eGFR), and Albumin to Creatinine Ratio (ACR). Our findings reveal a positive correlation between CRD scores and PDS, suggesting that disruptions in circadian rhythm may contribute to podocyte loss. Additionally, we observed a positive correlation between the CRD score and age.

Keywords: Podocytes, Circadian Rhythm, Podocyte Damage Score.

2 Introduction

Transcriptomic analysis focuses on the comprehensive capture of both coding and non-coding RNA while quantifying gene expression heterogeneity across cells, tissues, organs, and even entire organisms. This analysis is crucial as it lays the groundwork for the functional characterization and annotation of genes or genomes previously uncovered through DNA sequencing. Additionally, it establishes blueprints for reconstructing genetic interaction networks, which help in understanding cellular functions, growth, development, and biological systems. Furthermore, transcriptomic analysis produces molecular fingerprints of disease processes and prognoses, enabling the identification of potential targets for drug discovery and diagnostics [8].

Standardization, portability, and reproducibility are key challenges in bioinformatics. The nf-core/rnaseq pipeline provides a reliable solution for analyzing RNA sequencing data with reference genomes. Built on Nextflow, it integrates software packages through conda and supports containerization with Docker and Singularity. nf-core delivers portable and reproducible analysis pipelines that can be executed across various computational infrastructures, as emphasized in Patel et al., [2].

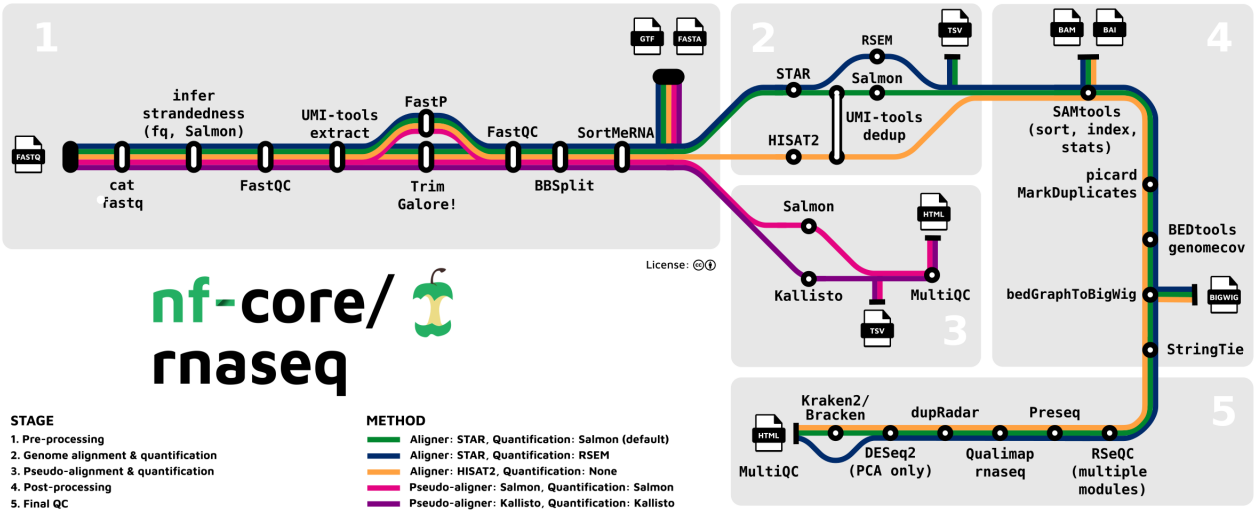

Figure 1: ***rna/seq****, Flowchart explaining the pipeline of nf-core/rnaseq [13]*

The nf-core/rnaseq pipeline uses STAR for RNA-seq alignment with a sequential search

in suffix arrays and a clustering procedure [5]. It also employs Salmon for quantifying transcript abundance, which corrects for fragment GC-content bias and is known for its speed in read mapping [14].

Transcriptomic analysis is essential for understanding cellular functions, particularly within the spatial context of tissues. While traditional bulk and single-cell transcriptomics quantify mRNA transcripts, they often lose the natural spatial information of cells. New spatial transcriptomic methods, like the Xenium platform from 10x Genomics, allow for in situ detection of mRNA. This platform uses fluorescent in situ hybridization, offering high sensitivity and spatial resolution to localize RNA molecules within tissue sections effectively [9].

The kidneys play a vital role in filtering blood and maintaining fluid and ionic balance. This process is essential and tightly regulated, as even minor alterations in filtrate reabsorption can lead to significant changes in blood chemistry and blood pressure [16]. Glomeruli, the blood-filtering units of the kidneys, are composed of specialized cells called podocytes, mesangial cells, endothelial cells, and parietal epithelial cells. Among these, podocytes are highly specialized, terminally differentiated cells that are crucial for the glomerular filtration process. The loss of podocytes is a defining characteristic of many forms of glomerular diseases [4].

Eukaryotes regulate their 24-hour rhythms through molecular circadian clocks. In mammals, the circadian system includes a central pacemaker located in the hypothalamus and oscillators found in peripheral tissues. Both central and peripheral clocks create 24-hour rhythms of gene expression through autoregulatory feedback loops [15].

The circadian clock regulates various biological processes and adjusts physiological responses to daily environmental changes resulting from the Earth's rotation [17]. The circadian clock consists of numerous genes, but the core molecular clock mechanism is primarily made up of four key circadian genes: *Clock, Bmal1, Period* (with homologs 1, 2, and 3), and

Cryptochrome (with homologs 1 and 2). These genes encode proteins that form feedback loops to regulate the transcription of clock-controlled genes. The involvement of circadian clock proteins in regulating a variety of renal transport genes indicates that the molecular clock in the kidney plays a crucial role in controlling circadian variations in renal function. The circadian clock emerges as a significant regulator of renal function, with important implications for the treatment of renal pathologies, including chronic kidney disease [16].

Transcriptomic analyses have confirmed that 13% of kidney genes exhibit a circadian transcription pattern. Recent studies revealed that the circadian clock seems to be essential in maintaining podocyte function and structure [17][1]. A similar study by Preston et al., [15] focused on the glomerular circadian clock to identify cell-autonomous circadian oscillators present in both mouse and human glomeruli, as well as a circadian glomerular transcriptome characterized by distinct gene clusters. They employed Metacycle to pinpoint rhythmic genes, applying a Benjamini-Hochberg adjusted p-value of ¡ 0.1 to identify those with a 24-hour periodicity. Additionally, recent large-scale alterations in glomerular transcriptional programs, while highlighting individual aspects of chronic kidney disease secondary changes, do not fully clarify the gradual process of podocyte degeneration [10]. To clarify that Padvitski et al., [10] developed a scalable tool to decode disease progression, which included the Podocyte Damage Score (PDS) to monitor cellular deterioration in kidney glomerulosclerosis. PDS is a single-cell damage score used to monitor cellular damage throughout the progression of chronic kidney disease. This score reflects the *in vivo* state of cells in chronic conditions over prolonged periods, allowing for the identification of molecular pathways linked to progressive damage.

Chronobiologists analyze circadian rhythms using three key metrics: amplitude, acrophase, and the midline estimating statistic of rhythms (MESOR). Amplitude represents half the difference between the peak and trough of a given response variable. Acrophase indicates the time at which the response variable reaches its peak, while mesor denotes the rhythmic equilibrium point [12].

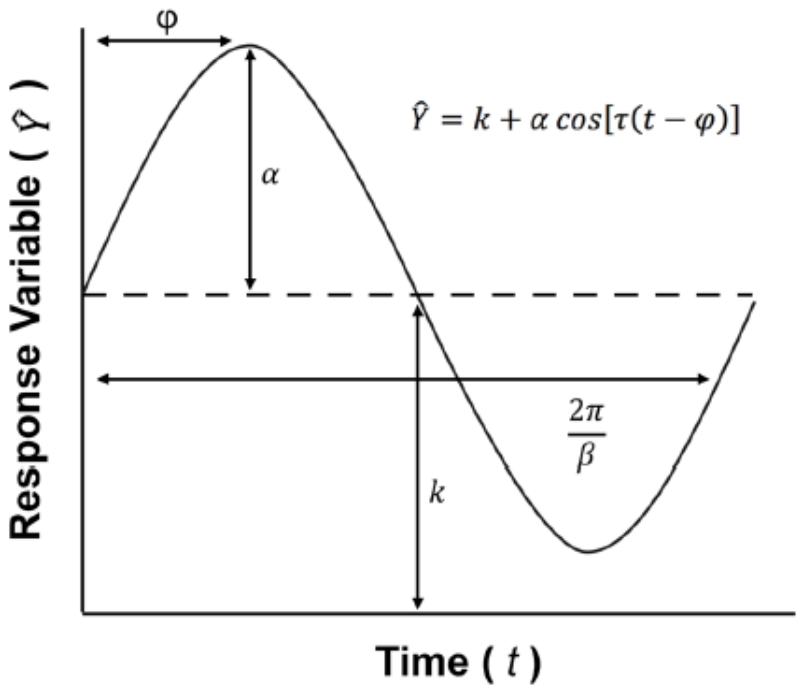

Figure 2: ***Rhythmic Curve****, The cosine curve and its metrics fitted to a hypothetical dataset of rhythmic genes [12].*

GLMM cosinor enables users to fit generalized linear models based on rhythmic data using a cosinor model. The linear coefficients of these models are estimated using the maximum likelihood method. To retrieve the parameters for acrophase and amplitude, the original parameters are transformed and returned along with the mesor as output [11].

Disruption of the circadian rhythm results in abnormalities in blood pressure and disrupts the circadian regulation of water and sodium homeostasis, ultimately contributing to the onset and progression of kidney diseases [6]. To assess this disruption of circadian rhythm, a computational algorithm was developed to evaluate the status of circadian disruption at the scRNA-seq level [7].

In this study, we aim to construct bulk RNA sequencing data derived from isolated mouse glomeruli collected over a 48-hour period [15]. We will convert the FASTQ files from the database into bulk RNA-seq format using the nf-core/rnaseq pipeline. Next, we

will utilize this bulk RNA-seq data to identify significantly cycling genes with the cglmm() package. Using these identified genes, we plan to calculate the CRD score for podocyte cells based on human kidney Xenium data. Furthermore, we will conduct a correlation analysis between the CRD scores and the previously calculated PDS scores. Additionally, we will explore other correlations, including the relationship between CRD scores and age, CRD scores and glomerular filtration rate (eGFR), the rate of primary urine formation as a key indicator of renal function [1], and CRD scores with the albumin-to-creatinine ratio (ACR), which serves as an early marker for diabetic nephropathy [3].

3 Methods

3.1 Acquisition of Datasets from the Server

The datasets referenced in Preston et al., [12] were obtained from the Array Express server utilizing the nf-core/fetchngs pipeline. To facilitate this process, a sample sheet was prepared that included all the accession numbers for the FASTQ files, which were then uploaded to the Array Express server.

3.2 RNAseq of fastq files

After downloading the FASTQ files, FastQC was performed to assess the quality of the read files. Files with low-quality reads were excluded, and the remaining files were utilized for subsequent analyses.

For RNA sequencing, the nf-core/rnaseq (v.3.18.0) pipeline from Nextflow was employed, utilizing the STAR aligner and the Salmon quantifier. To facilitate this process, a sample sheet was created that included the sample names, paths to read 1 and read 2, and strandness as columns.

In a Docker environment, the input files were provided, which included the reference genome FASTQ file and the genome annotation GTF file for the mouse. A sh script was generated and executed on a cluster to obtain the output files.

The raw counts for each sample were merged into a single matrix to generate the bulk RNA-seq data, which was then used for further analysis. Additionally, a metatable was created utilizing information available from the Array Express server.

3.3 Significantly Cycling Genes Detection

To identify the cycling genes within the RNAseq data, the GLMM cosinor package (v.0.2.1.9000) in R was utilized. Additionally, other libraries such as ggplot2 for plotting

and visualization, and tidyverse for data frame management, were loaded. The RNAseq data, along with its corresponding metadata, were incorporated into the analysis.

Initially, feature selection was conducted to reduce the dimensionality of the gene data. This involved retaining genes that have an expression value of more than 10 raw counts in at least 10 samples. Subsequently, a variance filter was applied, leading to the exclusion of genes with a variance below 20 raw counts. Finally, any remaining genes with a mean expression value under 10 raw counts were also removed.

The resulting data underwent reshaping and transformation, converting a wide-format table into a long-format suitable for input into the generalized linear model (GLM). This long table contains features and their expressions, along with the meta information for each sample.

This formatted table was utilized in the `cglmm()` function to identify cycling genes. The function was executed under a Gaussian family with the model `Expression ~ Gene`. The results were presented in a table that included the amplitude, acrophase, and p-value for each gene. Genes with a p-value of less than 0.01 were selected as significantly cycling genes, known as glomerular cycling genes. Additionally, the autoplot function from cglmm was employed to visualize the cosine wave nature of gene expression.

3.4 CRD Score Calculation

CRD scores were calculated for the podocyte cells within the Human Xenium data using the podocyte-specific cycling genes derived from the study by Padvitski et al., [10] To accomplish this, the RDS files of the Xenium data were loaded, and the expression matrix was extracted through the use of the `GetAssayData()` and `WhichCells()` functions, focusing exclusively on podocyte cells. For this expression matrix, transcript per million (TPM) values and log transformations were computed using a custom function. Initially, gene lengths were obtained from the Ensembl database by employing the EDASeq package's `getGeneLength&GCContent()` function, with org.db and hg38 also provided. After acquiring

the gene lengths, TPM was calculated, incorporating the log transformation within the process. The final output from this custom function is a gene matrix that features both log transformation and TPM normalization.

The gene matrix utilized in the `cal_CRD_score()` function of the CRDscore package in R, along with circadian genes sourced from the `cglmm()` function, was employed to calculate the CRD score for podocyte cells. A similar investigation was conducted for the glomerular-specific cycling genes, which were obtained from the previous study.

3.5 Correlation Studies

This study utilized the CRD scores of both podocyte cycling genes and glomerular cycling genes, as well as the PDS score obtained from the research by Padvitski et al., [10] to investigate the correlation between CRD and PDS scores. The correlation was assessed using the Spearman test, and the results were visualized in a scatter plot created with ggplot2.

In addition to this, other correlation studies were conducted. The average CRD score was calculated, and correlations were examined with various patient factors such as age, eGFR rate, and ACR rates.

4 Results

4.1 RNAseq of Mouse Samples

The fastq files for the samples were downloaded from Array Express. The dataset included 48 fastq files representing 24 paired RNA-seq samples. Among these, 3 samples were found to be corrupted due to low data size and were excluded from the RNA-seq analysis. The database also contained metadata, which was organized into a metatable with samples in the rows and age and circadian time in the columns. All mice were 8 weeks old, and the collection of kidneys was carried out every three hours, over a period of 48 hours.

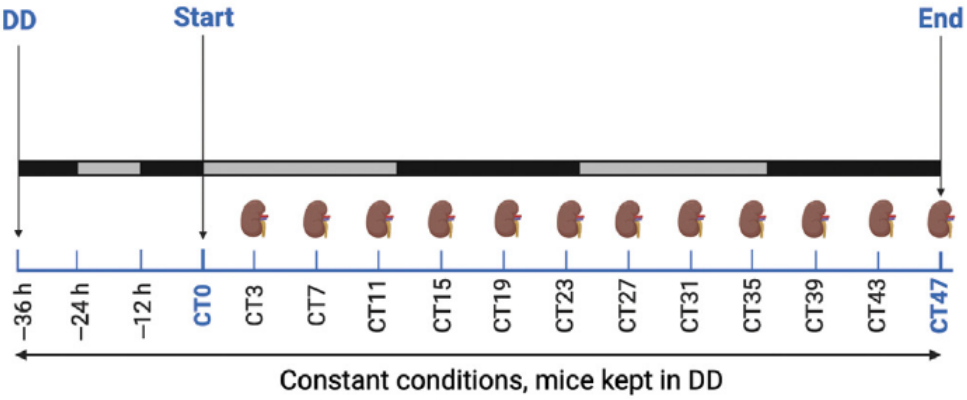

Figure 3: ***Graphical information of Mouse Samples****, Circadian time series tissue collection of kidneys for bulk RNA sequencing obtained from Preston et al., [15]*

RNA-seq was performed using the Nextflow nf-core/rnaseq pipeline. Results were generated upon the successful completion of the job. The output directory contained QC metrics, bigwig files, raw gene counts, and both scaled and TPM-normalized gene counts, along with gene lengths. Additionally, simpler count tables were available in both .rds and .tsv formats. The .tsv files for all samples were merged to create a combined gene count matrix for the 21 samples, with samples organized in columns and genes in rows. This matrix includes a total of 77981 genes in the rows.

4.2 Cycling Genes Detection

The prepared combined gene count matrix is loaded and first undergoes feature selection. Three filters are applied to exclude genes with low expression values and minimal variability. After feature selection, 15551 genes are retained, which are then used to identify cycling genes.

The count data for each gene across all samples is analyzed using the `cglmm()` function with the model specified as `Expression ~ amp_acro(Time, period=24, family=gaussian())`. The resulting table includes the amplitude values, acrophase values, and p-values for all genes. Genes with a p-value of less than 0.05 are filtered out, and those genes are considered significantly rhythmic. Ultimately, a total of 256 genes are identified, and the top five genes are visualized using the autoplot function of `cglmm()`.

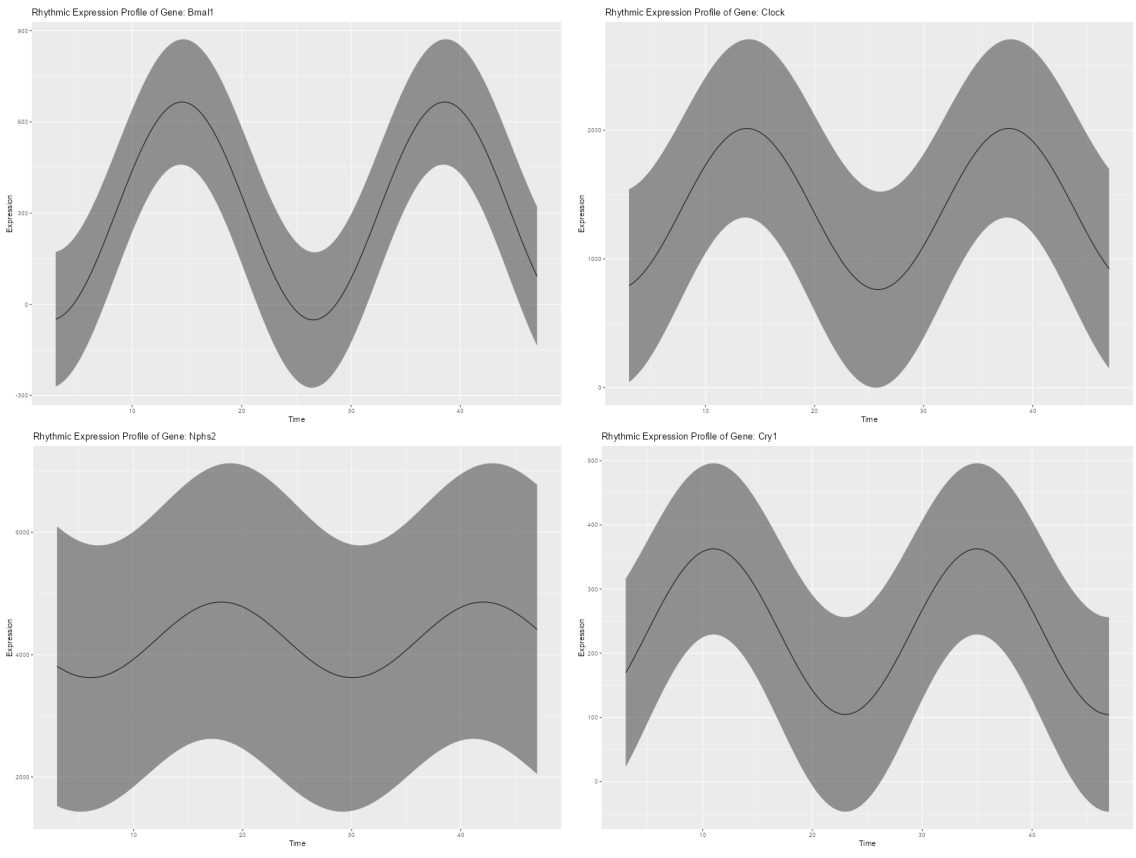

Figure 4: ***Circadian Expression****, Expression Profiles for Rhythmic Genes: Bmal1 (Top Left), Clock (Top Right), Nphs2 (Bottom Left), Cry1 (Bottom Right)*

4.3 CRDscore Calculation

The CRD scores were calculated using the glomerular and podocyte cycling genes of mice applied to the human Xenium data. Scores were obtained for each cell within the podocyte class. For the 16 samples of Xenium data, the CRD score was calculated and stored in an RDS file. The available PDS scores for these 16 samples were utilized for correlation studies alongside the CRD scores.

4.4 Correlation Studies

A correlation study was conducted using cells from samples that had both CRD and PDS scores. We observed that the CRD score increases as the PDS score increases, which is evident in the scatter plot. This trend was consistent across all samples. The aggregated scatter plot produced a Spearman correlation coefficient of 0.19.

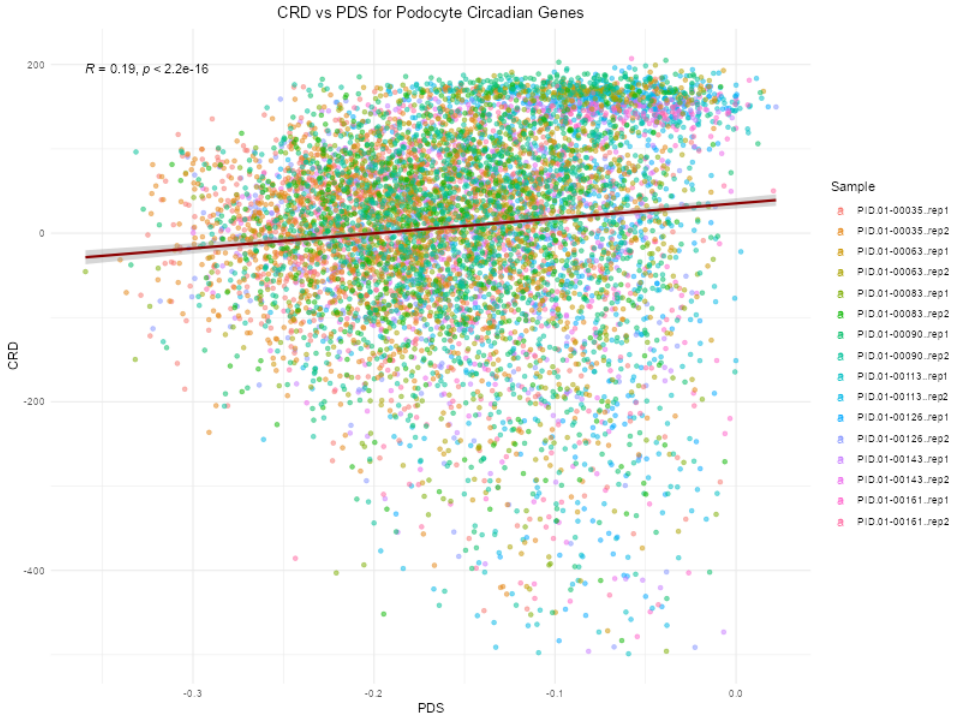

Figure 5: ***Correlation Study CRD vs PDS****, Scatter plot illustrating the correlation study between the CRD score for Podocyte Circadian Genes on the y-axis and PDS on the x-axis, for all samples combined and colored according to sample, with* $R = 0.19$*.*

We also examined the metadata in the same manner, analyzing the relationship between the age of the samples and the mean CRD scores. The scatter plot displaying mean CRD scores (on the y-axis) against age (on the x-axis) demonstrated a positive correlation, with an R value of 0.28, signifying an increase in CRD scores with age.

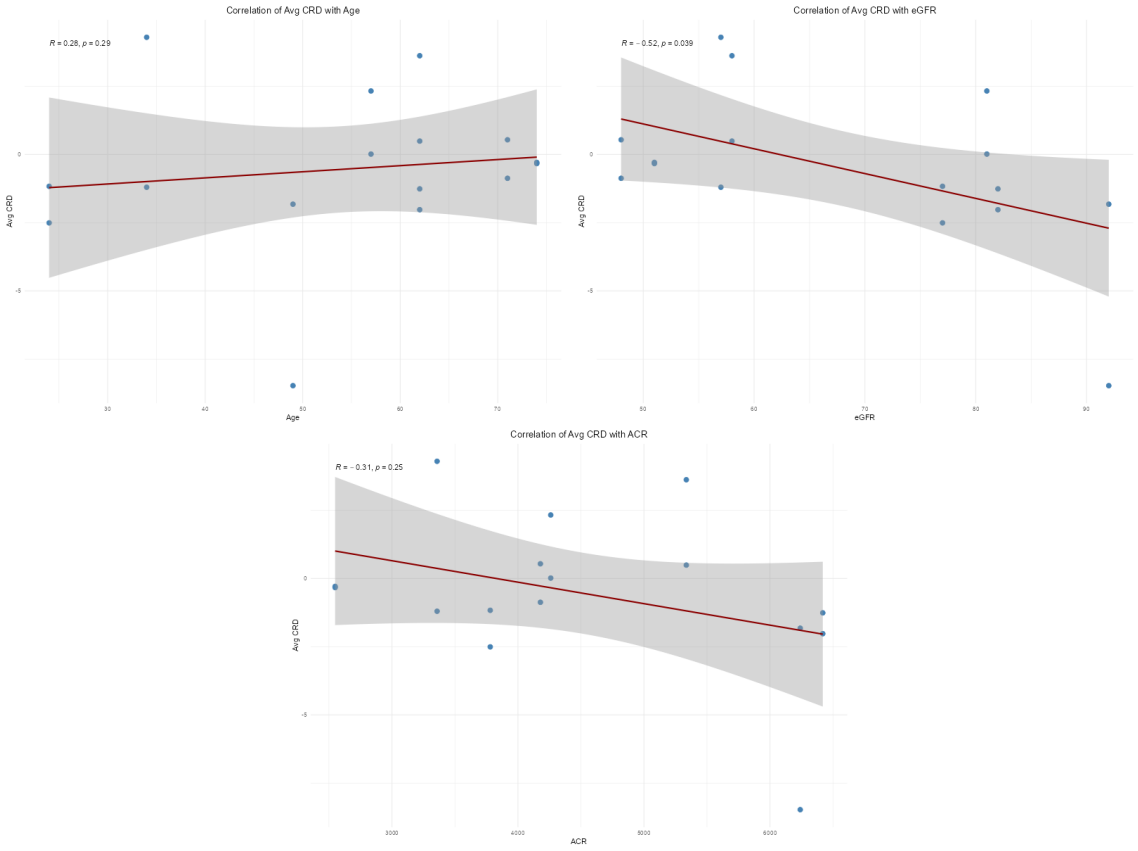

Figure 6: ***Correlation Studies****, Correlation studies conducted for Avg CRD scores calculated for Podocyte Circadian Genes with Age [R = 0.28] (Top Left), eGFR [R = -0.52] (Top Right), and ACR [R = -0.31] (Bottom).*

Similar investigations were performed on the eGFR and ACR values, where the scatter plot showed a negative correlation with both eGFR and ACR, with R values of -0.52 and -0.31, respectively. Correlation studies using CRD scores from Glomerular-specific cycling genes yielded similar results.

5 Discussion

Current research in renal circadian physiology is focused on identifying the molecular mechanisms underlying the circadian rhythmicity of renal functions, determining whether renal functional rhythms are driven by systemic circadian time cues (such as activity, food components, hormones, and body temperature) or by local mechanisms, and determining whether disruption of circadian rhythms causes human kidney diseases. [6]. The study by Preston et al., [15] provides evidence that both glomerular and podocyte-specific circadian clock mechanisms exist, likely influencing rhythmic fluctuations in glomerular physiology. Podocytes are significantly affected by changes in circadian rhythms [17][1]. However, Padvitski et al., [10] revealed a positive correlation between the degree of damage in podocytes (amongst 7 different podocyte disease mouse models) and a disruption in circadian rhythm, pointing towards a bidirectional interplay between circadian control and podocyte health.

We also sought to examine the relationship between CRD and PDS and found evidence indicating a positive correlation, specifically, that CRD increases as PDS increases. This suggests that disruptions in the circadian rhythm hint towards an impact on the podocyte health. This reasoning is supported by findings from Wang et al., [17] who investigated whether podocytes exhibit circadian oscillations, given that podocytes are subjected to glomerular hydrostatic pressure and various hematological substances, including electrolytes, amino acids, and glucose. Their study concluded that disturbances in circadian rhythm play a critical role in podocyte loss.

In our studies, we calculated the CRD score for human samples using the cycling genes identified from mouse samples. This suggests that cycling genes are conserved between animal models and humans, as we used the cycling genes obtained from the mouse model and used these to calculate the CRD score in Human data.

As we age, the CRD score increases, indicating greater circadian rhythm disruption. This disruption is linked to the loss of podocytes, which affects the circadian rhythm [17]. In their review article, Welz et al., [18] highlighted that deficiencies in core circadian clock genes

are associated with several aging-related hallmarks. These include abnormal activation of nutrient-sensing pathways, mitochondrial dysfunction, and stem cell exhaustion, all of which contribute to a shorter lifespan.

Based on our study, we propose that there may be a positive correlation between circadian rhythm disruption and the decline in podocyte health. Additionally, this relationship appears to be associated with the aging process, requiring further investigation into the implications of circadian rhythm disturbances.

References

[1] Camille Ansermet, Gabriel Centeno, Svetlana Nikolaeva, Marc P. Maillard, Sylvain Pradervand, and Dmitri Firsov. The intrinsic circadian clock in podocytes controls glomerular filtration rate. *Scientific Reports*, 9(1):16089, November 2019.

[2] Françoise Baylis. To publish or not to publish. *Nature Biotechnology*, 38(3):271–271, March 2020.

[3] Axel C Carlsson, Johan Sundström, Juan Jesus Carrero, Stefan Gustafsson, Markus Stenemo, Anders Larsson, Lars Lind, and Johan Ärnlöv. Use of a proximity extension assay proteomics chip to discover new biomarkers associated with albuminuria. *European Journal of Preventive Cardiology*, 24(4):340–348, March 2017.

[4] Joanna Cunanan, Daniel Zhang, Anna Julie Peired, and Moumita Barua. Podocytes in health and glomerular disease. *Frontiers in Cell and Developmental Biology*, 13:1564847, April 2025.

[5] Alexander Dobin, Carrie A. Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R. Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, January 2013.

[6] Dmitri Firsov and Olivier Bonny. Circadian rhythms and the kidney. *Nature Reviews Nephrology*, 14(10):626–635, October 2018.

[7] Lei He. Single-cell transcriptomic analysis reveals circadian rhythm disruption associated with poor prognosis and drug-resistance in lung adenocarcinoma. *John Wiley & Sons Ltd.*

[8] Zhihua Jiang, Xiang Zhou, Rui Li, Jennifer J. Michal, Shuwen Zhang, Michael V. Dodson, Zhiwu Zhang, and Richard M. Harland. Whole transcriptome analysis with se-

quencing: methods, challenges and potential solutions. *Cellular and Molecular Life Sciences*, 72(18):3425–3439, September 2015.

[9] Xiaokuang Ma, Peng Chen, Jing Wei, John Zhang, Chang Chen, Hanqiu Zhao, Deveroux Ferguson, Aaron W. McGee, Zhiyu Dai, and Shenfeng Qiu. Protocol for Xenium spatial transcriptomics studies using fixed frozen mouse brain sections. *STAR Protocols*, 5(4):103420, December 2024.

[10] Tsimafei Padvitski, Paula Unger Avila, He Chen, Cem Özel, Fabian Braun, Roman-Ulrich Müller, Pål O. Westermark, Paul T. Brinkkötter, Bernhard Schermer, Thomas Benzing, F. Thomas Wunderlich, Martin Kann, and Andreas Beyer. Single-Cell Resolution of Cellular Damage Illuminates Disease Progression, February 2025.

[11] Rex Parsons, Oliver Jayasinghe, Nicole White, and Oliver Rawashdeh. GLMMcosinor: Fit a Cosinor Model Using a Generalized Mixed Modeling Framework, January 2024. Institution: Comprehensive R Archive Network Pages: 0.2.1.

[12] Rex Parsons, Richard Parsons, Nicholas Garner, Henrik Oster, and Oliver Rawashdeh. CircaCompare: a method to estimate and statistically support differences in mesor, amplitude and phase, between circadian rhythms. *Bioinformatics*, 36(4):1208–1212, February 2020.

[13] Harshil Patel. nf-core/rnaseq: nf-core/rnaseq v3.19.0.

[14] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. Salmon provides fast and bias-aware quantification of transcript expression. *Nature Methods*, 14(4):417–419, April 2017.

[15] Rebecca Preston, Ruby Chrisp, Michal Dudek, Mychel R.P.T. Morais, Pinyuan Tian, Emily Williams, Richard W. Naylor, Bernard Davenport, Dharshika R.J. Pathiranage,

Emma Benson, David G. Spiller, James Bagnall, Leo Zeef, Craig Lawless, Syed Murtuza Baker, Qing-Jun Meng, and Rachel Lennon. The glomerular circadian clock temporally regulates basement membrane dynamics and the podocyte glucocorticoid response. *Kidney International*, 107(1):99–115, January 2025.

[16] Kristen Solocinski and Michelle L. Gumz. The Circadian Clock in the Regulation of Renal Rhythms. *Journal of Biological Rhythms*, 30(6):470–486, December 2015.

[17] Lulu Wang, Han Tian, Haiyan Wang, Xiaoming Mao, Jing Luo, Qingyun He, Ping Wen, Hongdi Cao, Li Fang, Yang Zhou, Junwei Yang, and Lei Jiang. Disrupting circadian control of autophagy induces podocyte injury and proteinuria. *Kidney International*, 105(5):1020–1034, May 2024.

[18] Patrick-Simon Welz and S.A. Benitah. Molecular Connections Between Circadian Clocks and Aging. *Journal of Molecular Biology*, 432(12):3661–3679, May 2020.