Exploring the effects of light and sleep deprivation on the chromatin structure and transcriptional landscape of specific brain regions



Steven Paul Walsh

St Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary term 2021

Abstract

Disruption to normal circadian rhythms and frequent periods of sleep deprivation (SD), are endemic in today's 24 hour society; therefore understanding the fundamental biology underlying what happens in the suprachiasmatic nuclei (SCN) during entrainment, and other specific brain regions following SD has never been more important. Current understanding of entrainment and SD within the SCN at the molecular level is still very limited, despite SD already having been shown to have a profound effect on brain function.

The published Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq) protocol was successfully modified to allow the technique to be effective with small amounts of tissue collected from murine SCN tissue. ATAC-seq was effectively used to investigate chromatin accessibility within cells of the SCN following light pulsing at different times of the day, and following SD in the SCN, dentate gyrus, and cortex. Data from ATAC-seq light pulse experiments was compared to published gene expression data following light pulsing and determined changes in chromatin accessibility also affect the transcriptome. RNA-seq was used to investigate gene expression following SD, and gain insights into how chromatin affects the transcriptome following SD.

Results obtained indicate, chromatin accessibility is highly dynamic within all brain regions investigated. ATAC-seq revealed chromatin accessibility changes in the SCN following light pulsing, but only at certain times of the day, and discovered chromatin modification across all brain regions following SD. Several genes identified close to differential chromatin are already known to play an important role within the molecular clock, and are recognized to be important for entrainment and SD. In addition several novel genes with differential chromatin nearby have been identified that also likely have an important role during entrainment and following SD.

Several areas of differential chromatin were investigated using a luciferase reporter assay, and results obtained indicate the areas of chromatin are able to strongly drive gene expression, and respond to treatment using forskolin and horse serum, both known to have an effect on the molecular clock.

Covid-19 Statement

I returned to university mid-February following my 3-month internship straight into the coronavirus pandemic. Upon my return I proceeded as normal, planning and setting up the next set of my experiments. However, shortly after returning to university I became ill with symptoms of Covid-19, and due to the overall uncertainty, I had to cancel all my experimental plans. This meant I lost all lab time since I had returned to Oxford in February, and I did not return to the laboratory until August. In total I did not complete any successful laboratory work for 10 months.

My situation was further exacerbated as I struggled to recover from being ill, I suffered constant migraines and fatigue. Initially thought to be long covid, blood testing revealed I was suffering anaemia, but due to the coronavirus pandemic it was very difficult to ascertain further diagnosis. Despite iron supplementation my condition worsened and this consequently, limited the amount of work I was able to do remotely during lockdown, and when I returned to the laboratory I found it very difficult to complete my work. Due to restrictions on appointments with specialists caused by the coronavirus pandemic, I only received a diagnosis in December 2020, when a large potentially pre-cancerous growth was identified and removed from my large intestine.

Upon returning to the laboratory, social distancing prevented me from being able to complete several of my most important and complex experiments, with them proving to simply be impossible to carry out successfully with restrictions in place. At the same time ordering many items into the lab became difficult, and fraught with delays, this included several items needed for my experiments.

Finally, following the Christmas break 2020, several members of my household caught Covid-19, so I had to self-isolate preventing me from going into the laboratory until the beginning of February. I was then unable to carry out laboratory work for a further 4 weeks in the final months of my PhD when our laboratory moved location, and due to the coronavirus pandemic this process became very complex, taking far longer than it normally would have. The coronavirus has had a very large and significant effect on my PhD, I have tried my best to overcome many of the obstacles it posed, but its impact was often very difficult to mitigate.

Table of Contents

1.0 Main Introduction 1

1.1 The mammalian molecular clock, and circadian rhythms 2

1.2 The wider mammalian molecular clock 3

1.3 Clock controlled genes 5

1.4 Circadian entrainment in the SCN 6

1.5 Circadian entrainment in the Periphery 9

1.6 Circadian rhythms and sleep 10

1.7 Health consequences of circadian rhythm disruption 11

1.8 Chromatin structure and its role in daily circadian gene regulation 12

1.9 Daily changes in chromatin structure 14

1.10 Epigenetic modifications 16

1.11 Non-coding RNA's 17

1.12 Hypothesis 1: Light pulsing induce structural chromatin changes in the SCN 18

1.13 Hypothesis 2: Chromatin changes following sleep deprivation in the SCN, dentate gyrus and cortex 19

1.14 Methods to assay chromatin structure 20

1.15 Bioinformatic analysis 22

1.16 Outline of result chapters and key questions 23

2.0 Main Methods

2.1 Nuclei preparation 25

2.1.1 Extraction of single cells from brain tissue using enzymatic digestion and sonication 26

2.1.2 Nuclei Extraction straight from SCN tissue using Dounce homogenizer 28

Table of Contents

2.2 Preparation and sequencing of ATAC-seq DNA libraries 29

2.3 RNA extraction and RNA sequencing library preparation 30

2.4 Quantifying libraries and correct colour balance 35

2.5 Sequencing of DNA libraries 37

2.6 In-vivo animal methods 42

2.7 Cultured cell protocol 44

2.8 Transfection of cultured cells 44

2.9 Preliminary local bioinformatic analysis 46

2.9.1 Initial cleaning up of data and alignment prior to the initial analysis 47

2.9.2 Read coverage analysis and visualisation of data in genome browsers 48

2.9.3 Running the programs locally 49

2.9.4 Linux shell script 50

2.10 In-depth bioinformatic analysis for deeper analysis of data 53

2.10.1 Initial quality control of data 56

2.10.2 Alignment, filtering and visualisation of coverage tracks 59

2.10.3 Peak calling 61

2.10.4 Read counting 62

2.10.5 Differential analysis 66

2.10.6 Principal component analysis 71

2.10.7 Batch correction with RUVseq 72

2.10.8 Motif searching 75

2.10.9 Gene set enrichment analysis 80

Table of Contents

3.0 Refinement of nuclei extraction for ATAC-seq and subsequent pilot ATAC-seq experiments 82

3.1 SCN dissociation using enzymatic digestion and sonication 83

3.2 Pilot experiment 1: ATAC sequencing using nuclei extracted with papain 86

3.3 Pilot experiment 2 using nuclei isolated with a Dounce homogenizer 89

3.4 Chapter 3 Discussion 93

4.0 The effect of light on the chromatin structure within the SCN

4.1 Introduction 94

4.2 ATAC-seq experiment 1 results 99

4.2.1 Quality control 100

4.2.2 Differential analysis of aligned reads 105

4.2.3 Gene and pathway enrichment analysis 113

4.2.4 Motif enrichment analysis 115

4.2.5 Batch correction with RUVseq 124

4.2.6 Comparison of genes with differential chromatin to published micro-array data 128

4.2.7 Comparison of chromatin accessibility with publicly available data for epigenetic marks 130

4.3 ATAC-seq light pulse experiment 2 132

4.3.1 Quality control 133

4.3.2 Differential analysis of aligned reads 137

4.4 Transfection of genomic sequences of interest into cultured cells 141

4.5 Chapter 4 Discussion 151

Table of Contents

5.0 The effect of sleep deprivation on the chromatin structure and transcriptional landscape of specific brain regions

5.1 Introduction 155

5.2 Chromatin accessibility dynamics in the mouse brain following acute sleep deprivation 159

5.2.1 Quality control 160

5.2.2 Differential analysis of ATAC-seq aligned reads 165

5.2.3 Gene set enrichment analysis 170

5.2.4 Motif analysis 172

5.3 Re-analysis of published cortex ATAC-seq data 178

5.4 The effect of acute sleep deprivation on the transcriptome of the SCN, DG and cortex 183

5.4.1 RNA-seq quality control 184

5.4.2 Differential analysis of aligned reads for RNA-seq 188

5.4.3 Gene set enrichment analysis for RNA-seq 192

5.4.4 RNA-seq motif analysis 194

5.5 Comparison of chromatin accessibility with the transcriptome following sleep deprivation 198

5.6 Transfection of sleep deprivation genomic sequences of interest into cultured cells 200

5.7 Chapter 5 Discussion 205

6.0 Role of circadian regulation in the immune response to Covid-19

6.1 Introduction 211

6.2 Methods 212

6.3 Results 215

6.4 Chapter 6 Discussion 217

Table of Contents

7.0 Final project discussion 219

References 222

List of Figures

Chapter 1: Introduction

1.1 The core molecular clock 2
1.2 The wider molecular clock 4
1.3 A typical clock controlled gene 5
1.4 Light induction of PER/FOS, in the SCN 7
1.5 The Phase response curve for a nocturnal animal such as a mouse 8
1.6 Multiple inputs integrate at the promoter of *Per1/2* 9
1.7 The two process model of sleep 10
1.8 The different levels of chromatin organisation 13
1.9 Cycling genes in the liver 15
1.10 ATAC-seq reaction schematic 22

Chapter 2: Main Methods

2.1 Illumina sequencing 37
2.2 ATAC-seq library preparation workflow 40
2.3 RNA-seq library preparation workflow 41
2.4 Zeitgeber time vs circadian time 42
2.5 Light pulse timing 43
2.6 Tissue collection times for the sleep deprivation 43
2.7 A typical .fastq file 49
2.8 Code for Linux shell script 52
2.9 Further code for Linux shell script 51
2.10 Further code for Linux shell script 52
2.11 The full bioinformatic workflow for ATAC-seq 54
2.12 The full bioinformatic workflow for RNA-seq 55
2.13 Typical ATAC-seq fragment size distribution 58
2.14 Python code to automate converting .bam files .bw files 59
2.15 Code for the track hub for the pilot experiment 60
2.16 R code written to carry out read counting 62
2.17 Code written to covert .bed peak files into .GTF peak files 63
2.18 Code written to automate the read counting process 64
2.19 A typical counts table for sequencing data 65
2.20 RMarkdown for differential analysis with DESeq2 67
2.21 A typical Bland–Altman MA plot 68
2.22 RMarkdown code to contrast the different experimental conditions 69

List of Figures

2.23 A typical PCA plot 71

Chapter 3: Refinement of nuclei extraction for ATAC-seq and subsequent pilot ATAC-seq experiments

3.1 Nuclei fluorescing under ultraviolet light 85
3.2 Read distribution across the different samples for both pilot experiments 88
3.3 Read coverage for Pilot 1 near to *Per1* 88
3.4 Nuclei fluorescing under ultraviolet light following extraction using a Dounce homogenizer 91
3.5 Coverage tracks for *Fos* from pilot experiment 2 92
3.6 Coverage tracks for *Csnk1e* from pilot experiment 2 93

Chapter 4: The effect of light on the chromatin structure within the SCN

4.1 In vivo knockdown of *Sik1* results in enhanced phase shifting 95
4.2 Read distribution for light pulse experiment 1 100
4.3 Quality control for the light pulse experiment 1 sequencing run 101
4.4 Further quality control for light pulse experiment 1 103
4.5 PCA plot for read counts from the first ATAC-seq light pulse experiment 104
4.6 Bland–Altman plots for ATAC-seq light pulse experiment 1 107
4.7 Volcano plots for ATAC-seq light pulse experiment 1 108
4.8 Read coverage tracks for *Fos* and *Csnk1e* from light pulse ATAC-seq experiment 1 109
4.9 Read coverage tracks for *MafG* and *Nfil3* from light pulse ATAC-seq experiment 1 110
4.10 Read coverage tracks for *Rora* from ATAC-seq light pulse experiment 1 111
4.11 Venn plot for the numbers of statistical significant areas chromatin 112
4.12 Clustering of GO BO/MF terms for light pulse ATAC-seq experiment 1 114
4.13 Read coverage tracks for *Dtnbp1* from light pulse ATAC-seq experiment 122
4.14 Read coverage tracks for *Schip1* from light pulse ATAC-seq experiment 123
4.15 RUV-seq normalization using non-normalized data as input 125
4.16 RUV-seq normalization using DESeq2 normalized data as input 126
4.17 Venn diagram for comparison of microarray and light pulse experiment 1 data 128
4.18 Comparison of ATAC-seq read coverage with epigenetic CHIP-seq read coverage 131
4.19 Read distribution for light pulse experiment 2 133

List of Figures

4.20 Quality control for the light pulse experiment 2 sequencing run 134

4.21 Further quality control light pulse experiment 2 135

4.22 PCA plot for read counts from the second ATAC-seq light pulse experiment 136

4.23 Bland–Altman plots for ATAC-seq light pulse experiment 2 138

4.24 Read coverage tracks for *Fos* and *Csnk1e* from ATAC-seq light pulse experiment 2 139

4.25 Read coverage tracks for *MafG* and *Nfil3* from ATAC-seq light pulse experiment 2 140

4.26 Read coverage tracks for *Csnk1e* experiment 1 to show motif locations 143

4.27 Read coverage tracks for *MafG* experiment 1 to show motif locations 144

4.28 NIH3T3 cells transfected with the sequences of interest treated with forskolin 147

4.29 NIH3T3 cells transfected with the sequences of interest treated with horse serum 148

4.30 U2OS cells transfected with the sequences of interest treated with forskolin 149

4.31 U2OS cells transfected with the sequences of interest treated with horse serum 150

Chapter 5: The effect of sleep deprivation on the chromatin structure and transcriptional landscape of specific brain regions

5.1 Read distribution across the different sleep deprived ATAC-seq samples 160

5.2 FastQc results for sleep deprivation ATAC-seq 161

5.3 Further quality control for sleep deprived ATAC-seq experiment 163

5.4 PCA plot for read counts from sleep deprived ATAC-seq experiment 164

5.5 Bland–Altman plots across brain regions for sleep deprived ATAC-seq experiment 166

5.6 Bland–Altman plots within brain regions for sleep deprived ATAC-seq experiment 167

5.7 Volcano plots for areas of differential chromatin following sleep deprivation 168

5.8 Read coverage tracks for sleep deprived ATAC-seq experiment 169

5.9 Clustering of GO BO/MF terms for sleep deprived ATAC-seq experiment 171

5.10 Location of differential chromatin within the genome following SD 172

5.11 Read coverage tracks for the second ATAC-seq following SD 177

5.12 PCA plot for ATAC-seq data from Hor, Yeung et al. (2019) 179

5.13 Bland–Altman plots for differential genes comparing published data with my own 180

List of Figures

5.14 Read coverage tracks near to *MiR17hg* 180
5.15 Clustering of GO terms for Hor, Yeung et al. (2019) 182
5.16 Read distribution across the different sleep deprived RNA-seq samples 184
5.17 FastQc results for sleep deprivation RNA-seq 185
5.18 Further quality control for sleep deprived ATAC-seq experiment 186
5.19 PCA plot for read counts from RNA-seq SD experiment 187
5.20 Bland–Altman plots for RNA-seq sleep differential analysis 189
5.21 Volcano plots showing differential genes from RNA-seq following SD 190
5.22 RNA-seq read coverage tracks for Arc and Homer1a 191
5.23 GO biological process enrichment for DE genes following sleep deprivation 193
5.24 Read coverage from ATAC-seq and RNA-seq for Cirbp 199
5.25 Sleep deprived ATAC-seq read coverage tracks near to *MiR17hg* 201
5.26 Results for luciferase reporter assay using NIH3T3 cultured cells 203
5.27 Results for luciferase reporter assay using U2OS cultured cells 204
5.28 DNA Double Stranded breaks 209

Chapter 6: Role of circadian regulation in the immune response to Covid-19

6.1 R-code to determine which genes have a *Bmal1* motif in the SARS-Cov-2 gene list 214
6.2 Chip-seq binding tracks for *Bmal1* from heart, kidney and liver 216
6.3 Silencing *Bmal1* reduces SARS-CoV-2 replication 218

List of Tables

Chapter 2: Main Methods

2.1 The barcode I.D. sequences used for the pilot experiments 36

2.2 The different light/dark cycles used 43

Chapter 3: Refinement of nuclei extraction for ATAC-seq and subsequent pilot ATAC-seq experiments

3.1 Pulse frequencies and durations used to dissociate SCN tissue with the Bertin Precellys24 Instrument 84

3.2 Cellular dissociation of neural tissue using a variety of enzymes 85

3.3 Cellular homogenization using a Precelleys24 86

3.4 The concentrations of the first ATAC-seq libraries generated 87

3.5 The four experimental groups for the second pilot experiment 90

3.6 The concentration of DNA libraries from the second pilot experiment prior to sequencing 91

Chapter 4: The effect of light on the chromatin structure within the SCN

4.1: The experimental groups for the sleep deprived experiment 99

4.2: Known motifs identified within differential chromatin following a 30 minute light pulse at CT16 116

4.3: De-novo motifs identified within differential chromatin following a 30 minute light pulse at CT16 116

4.4: Known motifs identified within differential chromatin following a 60 minute light pulse at CT16 117

4.5 De-novo motifs identified within differential chromatin following a 60 minute light pulse at CT16 118

4.6 Known motifs identified within differential chromatin changing across the day 119

4.7 De-Novo motifs identified within differential chromatin changing across the day 120

4.8 Contingency table for fisher exact test to compare the transcriptome with chromatin accessibility after light pulsing 129

List of Tables

Chapter 5: The effect of sleep deprivation on the chromatin structure and transcriptional landscape of specific brain regions

5.1 The experimental groups for the ATAC-seq SD experiment 159

5.2 Motifs enriched within differential chromatin identified in the SCN following SD 174

5.3 Motifs enriched within differential chromatin identified in the DG following SD 175

5.4 Motifs enriched within differential chromatin identified in the cortex following SD 176

5.5 The experimental groups for the RNA-seq SD experiment 183

5.6 Motifs located near to DE genes following SD within the SCN 195

5.7 Motifs located near to DE genes following SD within the DG 196

5.8 Motifs located near to DE genes following SD within the cortex 197

5.9 Contingency table for fisher exact test for RNA-seq cortex DE genes after SD and published data 198

Chapter 6: Role of circadian regulation in the immune response to Covid-19

6.1 Results for motif specific searching against BMAL1 core, Mouse BMAL1 and Human CLOCK 215

6.2 Contingency table for fisher exact test for SARS-Cov-2 interacting genes lists with a BMAL motif nearby, and gene lists for experimentally verified chip-seq binding sites 215

List of Commonly used Abbreviations

ATAC-seq	Assay for transposase accessible chromatin sequencing
ACE	Angiotensin converting enzyme
CCG	Clock controlled gene
CHIP-seq	Chromatin immunoprecipitation sequencing
DE	Differential expression
DG	Dentate gyrus
GSEA	Gene set enrichment analysis
IGB	Integrated genome browser
LTC	Light tight chamber
RNA-seq	RNA sequencing
SCN	Suprachiasmatic nuclei
SD	Sleep deprivation
TTFL	Transcriptional-translational feedback loop
TF	Transcription factor

Chapter 1

Introduction

Almost all life on earth has evolved in an environment of a continuous and predictable light/dark cycle, leading to many cellular processes oscillating over an approximate 24 hour period (Roenneberg and Merrow, 2002). This gives an individual their circadian rhythm, which enables them to predict light/dark cycles, preparing them for environmental factors occurring at different times of the day. Circadian rhythms are controlled by the molecular clock (Hardin et al., 1990), a complex transcriptional-translational feedback loop (TTFL), which in addition to following the passage of time at the molecular level, governs the expressions of countless clock controlled genes (CCGs).

The intricate functioning of the molecular clock is now well understood, however, there still remain many gaps in our understanding of how it is entrained to the external environment, and how an individual's recent sleep history influences gene expression. This project aims to explore the underlying molecular mechanisms that directly control the molecular clocks of cells within the suprachiasmatic nuclei (SCN), the central pacemaker of the circadian system. Taking account of exactly how light input via the optic nerve and recent sleep history, have a direct effect on molecular clock gene expression within the SCN.

Gene expression has been shown to change within the SCN as a result of light input, and in many brain regions following recent changes to sleep history, such as sleep deprivation (SD). Changes to molecular clock output are no doubt a product of these changes and their subsequent interaction, but the mechanism of action is still poorly understood. This project investigates not only changes to gene expression, and how they influence the molecular clock, but also primarily focusses on regulation at the chromatin level, which provides the principle point for any change in gene expression. Undeniably, if areas of DNA are closed and inaccessible for transcription, genes simply cannot be transcribed, and will therefore be downregulated. Currently very little is understood about how chromatin accessibility changes in the SCN following light input and recent changes to sleep history, and specifically how these changes will further influence the molecular clock.

1.1: The mammalian molecular clock, and circadian rhythms

The circadian molecular clock is the product of a TTFL, with a period of around 24 hours, consisting of four clock proteins; Circadian locomotor output cycles kaput (CLOCK), Brain and muscle ARNT like protein (BMAL), Period (PER) and Cryptochrome (CRY) (Buhr and Takahashi, 2013). BMAL1 and CLOCK heterodimerize (Huang et al., 2012), and act as the positive arm by binding to E-box promoter elements (CACGTG), in CCGs including *Per1* and *Cry1* (Figure 1.1). *Per* and *Cry* protein products in turn act as the negative arm by attenuating the action of CLOCK/BMAL1 through repression of their gene expression, therefore completing the loop. It is the rate of accumulation and subsequent degradation of these proteins that determines the period of circadian oscillation.

The TTFL loops are further complemented by an ancillary loop, consisting of two nuclear receptors REV-ERB (α/β) and retinoic acid related orphan receptor (ROR-α/β/γ) (Akashi and Takumi, 2005); these are activated by CLOCK/BMAL1. REV-ERBα represses Bmal1 transcription, however, in this loop it is counteracted by continuous activation of Bmal1 by RORα25. This alternate action of ROR and REV-ERB allows rhythmic expression of Bmal1 (Figure 1.1).

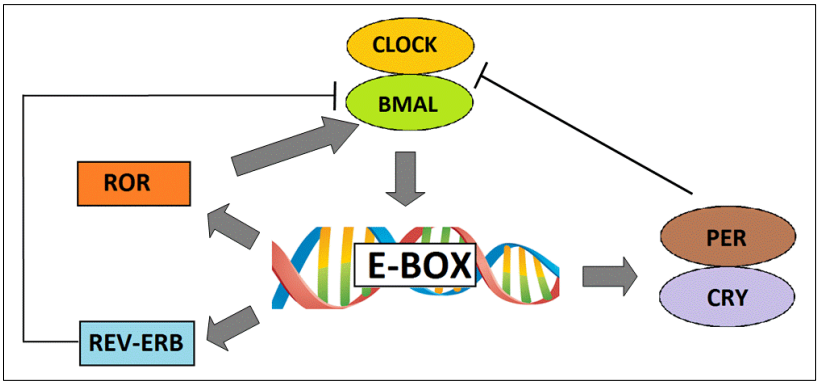


Figure 1.1: The core molecular clock. PER and CRY are transcribed when the CLOCK/BMAL1 heterodimer binds to the E-box of their gene promoter regions. PER and CRY then act to repress Clock and Bmal1 expression. Nuclear receptors REV-ERB and ROR are also activated by CLOCK/BMAL1; REV-ERB then suppresses further expression of Bmal1, while ROR continually activates BMAL.

The molecular clock is present in all cells of an individual, regardless of whether it's a neuron, hepatocyte, epithelial cell or cardiomyocyte. It is the correct alignment of the circadian rhythms of all these individual molecular clocks that ensures optimum health for an individual. The SCN, located in the hypothalamus, is responsible for ensuring that an individual's circadian rhythm is aligned to the external environment, and that the individual molecular clocks of all cells are correctly synchronised. The SCN's location within the hypothalamus enables it to synchronise cells through a combination of direct synaptic signalling on nearby neurons, or via the endocrine system to more distant cells.

1.2: The wider mammalian molecular clock

Figure 1.1, illustrates the core molecular clock, however the full molecular clock shown in Figure 1.2 is far more complex. The different elements of the molecular clock, including PER1, CRY, CLOCK, BMAL1, ROR and REV-ERB, are surrounded by a number of additional cis and trans regulatory elements within each loop. Due to the complexity of the molecular clock, the effects of the up or down regulation of any genes that directly encode for its components will likely have a significant effect across the whole clock. And, although many of the clocks components have been extensively studied, due to its intricacy there are still many of its mechanisms that are poorly understood.

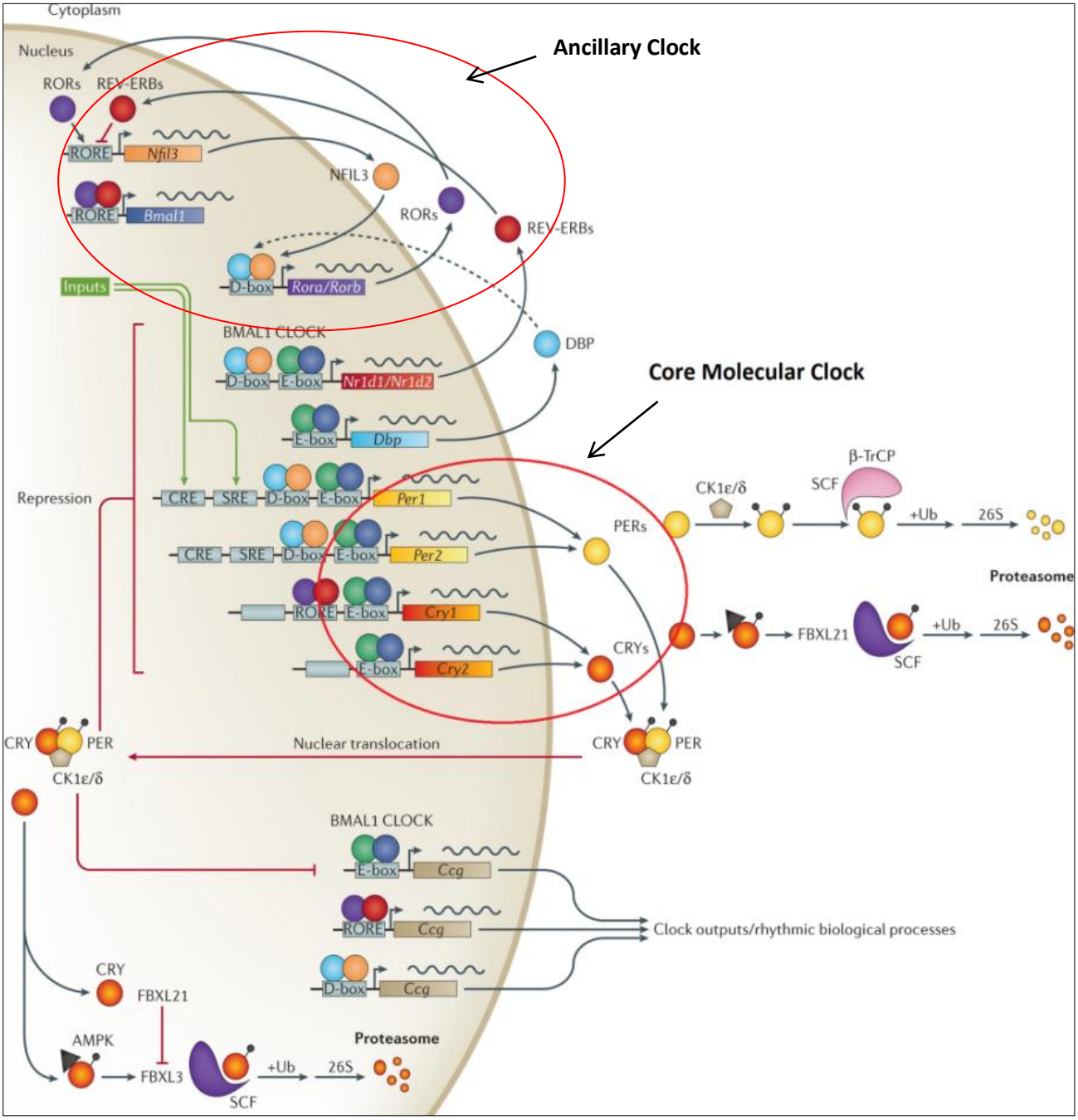


Figure 1.2: The wider molecular clock. The core and ancillary clock are surrounded by a number of cis and trans regulatory elements. Adapted from Takahashi (2017).

1.3: Clock controlled genes

The molecular clock not only simply keeps track of time at the molecular level, it also controls the activation and subsequent repression of a large number of genes; in mice almost 50% of the transcriptome (Zhang et al., 2014) is thought to be under the control of the circadian clock. This helps to ensure that the correct proteins are transcribed, at the appropriate time of day. For example the digestive tract must transcribe many digestive enzymes throughout the day, but switches this production off at night. The high level of circadian control attained by the molecular clock is achieved through the integration of binding sites for CLOCK and BMAL1 into the regulatory elements of CCG's. These sites are referred to as an enhancer box (E-box), and contain the binding sequence CANNTG (where N can be any nucleotide), with a palindromic canonical sequence of CACGTG. Figure 1.3 shows a typical CCG, with its E-box.

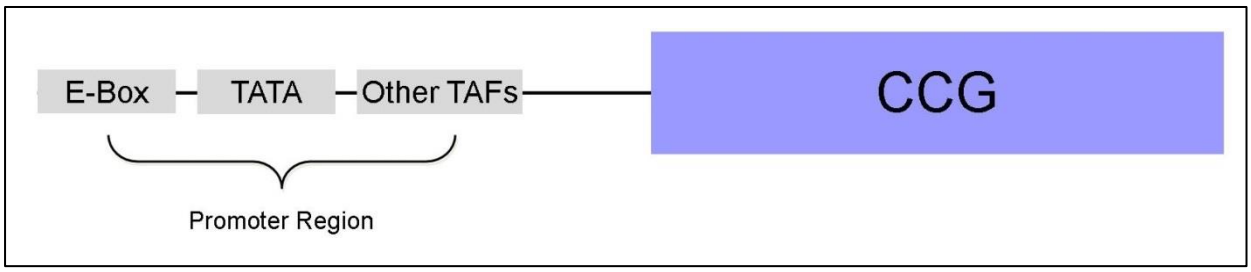


Figure 1.3: A typical CCG. The E-box can be seen within the promoter for the gene along with the TATA box and binding sites for other TATA-box-binding protein (TBP) associated factors (TAFs). CLOCK and BMAL1 will bind along within other TAFs at the promoter to form the Transcription factor II D (TFIID) protein complex, initiating transcription of the CCG.

Further circadian regulatory control is accomplished by second order CCGs through the integration of an E-box into genes that function as trans-regulatory elements such as *Myc* (Desbarats et al., 1996). These genes frequently transcribe proteins that have a direct effect on the expression of other genes enabling the molecular clock to indirectly control a wide variety of transcription factors, these often have distant effects on gene expression, and are not necessarily transcribed in the same circadian phase as first order CCGs.

As a result of the extensive control the molecular clock has on gene expression, daily circadian rhythms can have a wide ranging effect on many biological processes. These include, but are not limited to; blood pressure, glucose tolerance, body temperature, alertness, bowel movements, and sleep itself (Millar-Craig et al., 1978, Refinetti and Menaker, 1992, Voigt et al., 2019). This therefore demonstrates the potential health consequences of circadian disruption, and as a result the likely benefits of any new drugs or therapies that can help to restore co-ordinated circadian rhythms, correctly aligned to the exterior environment.

1.4: Circadian Entrainment in the SCN

Cells of the SCN like all cells contain their own molecular clock. However, SCN neurons have the distinctive ability to quickly align their own molecular clocks to the outside environment by synaptic input from the eyes via the retinohypothalamic tract (RHT). As previously described, *Per1/2* is a core element of the molecular 24h clock, but remarkably in the SCN, *Per1*/2 are also activated via light input (Schwartz et al., 2011). Photic signalling via the RHT releases neurotransmitters, which activate several G-protein coupled receptor (GPCR) signalling pathways and results in an influx of Ca2+ and cAMP within SCN neurons. This leads to subsequent activation of two mitogen activated protein kinase (MAPK) pathways; extracellular-signal regulated kinase (ERK) and c-JUN NH2-terminal kinase (JNK) (Obrietan et al., 1998). Phosphorylated ERK activates protein kinase A (PKA) and calmodulin dependent kinase (CamK) which in turn phosphorylates CREB (Lonze and Ginty, 2002), allowing it to bind to CRE elements in promoters of *Per1* and *Fos* to induce their expression (Figure 1.4). In parallel, ERK also activates Elk-1, which binds to serum response elements (SRE) in promoters of *Per1* and *Fos* to induce their expression independent of CREB.

On the other hand, phosphorylation via the JNK pathway leads to activation of JUN proteins, which heterodimerize with cFOS to form the activating protein-1 (AP-1) transcription factor complex (Halazonetis et al., 1988). This AP1 factor induces *Per1* expression by binding to the AP-1 binding site within the *Per1/2* promoters (Kornhauser et al., 1990).

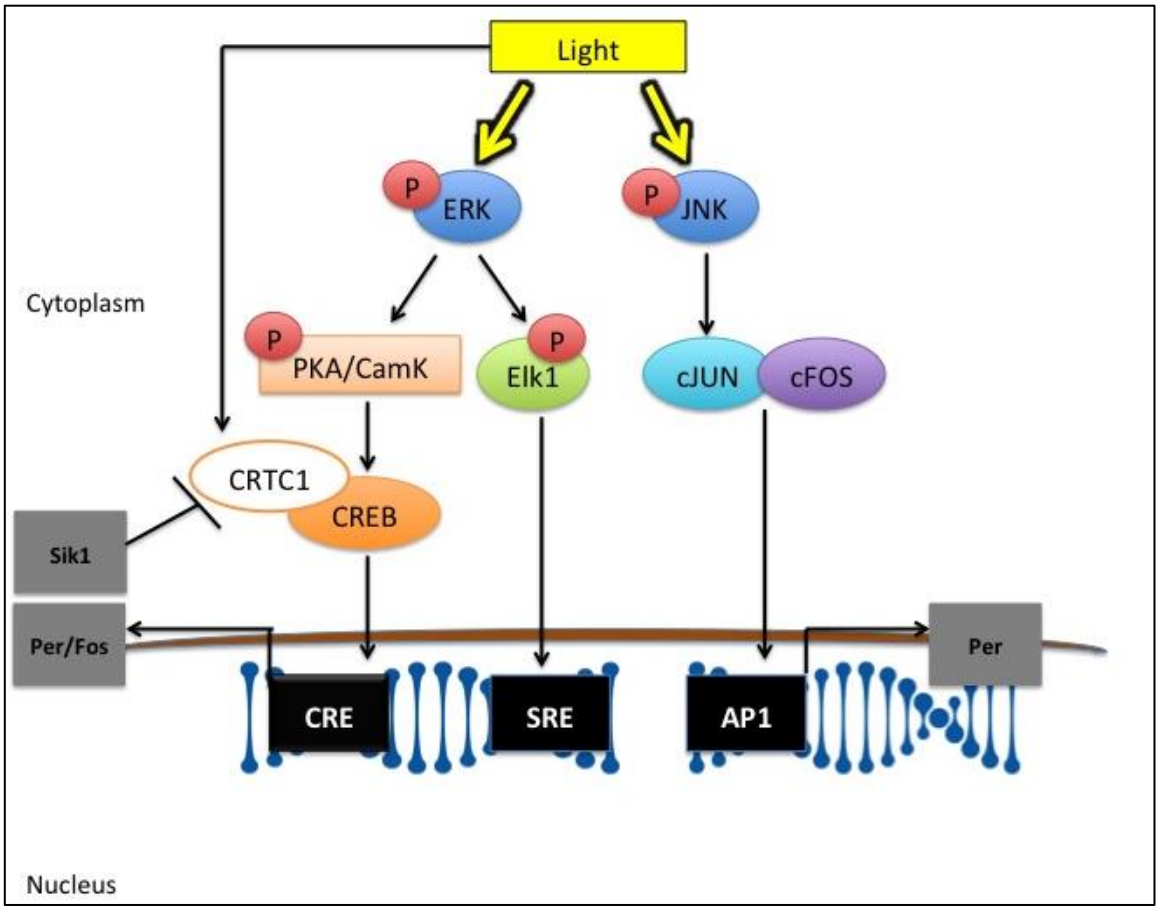


Figure 1.4: Light induction of PER/FOS, in the SCN. This occurs via two pathways, either through CREB/Elk1 via ERK, or by the AP1 complex via JNK.

However, circadian entrainment is far more complex; indeed the above model suggests the potential for integration of multiple inputs at the level of clock gene promoters. Entrainment mechanisms also allow for the modulation of pathways that provide input. As a key example, light does not phase shift the clock equally at all times (Ralph and Menaker, 1988); for a diurnal human the largest shifts occur during the morning at dawn, while for a nocturnal mouse it is mainly at dusk. Furthermore, depending on the time of day, either a phase advance or a phase delay will occur, this is recognised as the phase response curve (PRC), and is shown in Figure 1.5. In addition to the PRC, other physiological factors such as sleep history also modulate the impact of light on the clock (Franken, 2013), allowing circadian timing to be fine-tuned to the multiple demands of the environment and internal physiology.

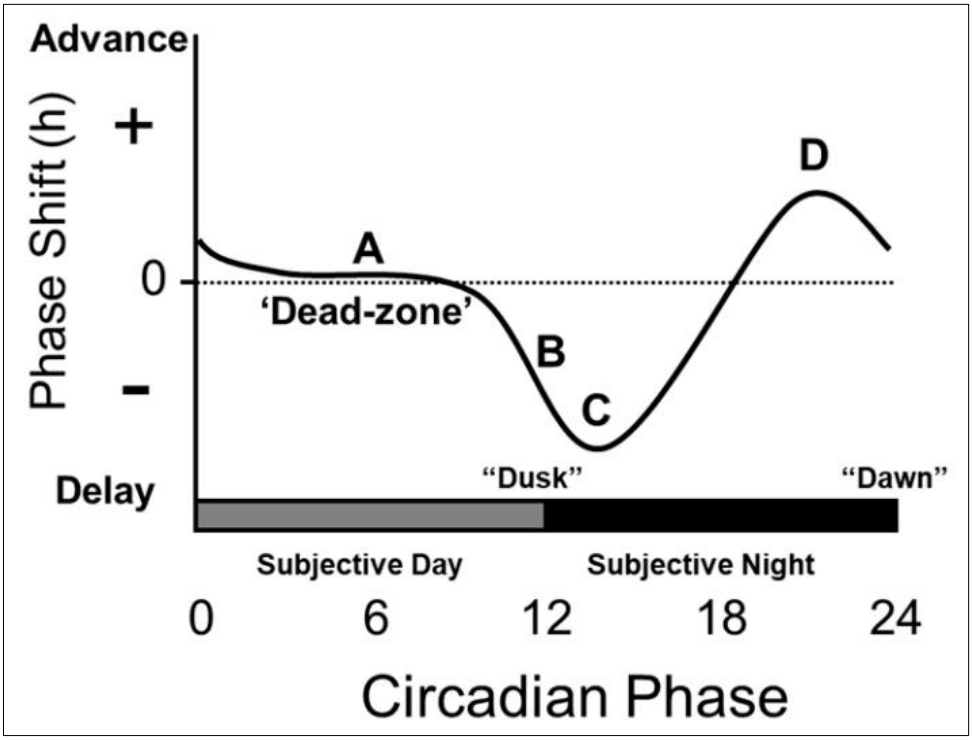


Figure 1.5: The Phase response curve for a nocturnal animal such as a mouse. Phase shifts to the molecular clock occur primarily during the subjective night, taken from Foster et al. (2020).

1.5: Circadian Entrainment in the Periphery

Within the SCN, the foremost Zeitgeber is light, but within peripheral tissue the process of entrainment becomes more complex, and results from a combination of many different factors that all integrate at the promoter of clock genes through multiple regulatory enhancers (Figure 1.6). The SCN has a diminished effect on peripheral clocks, and must depend on in-direct signalling to tissues throughout the body. By interacting with other brain regions, it can modulate signalling molecules including hormones such as cortisol (Quigley and Yen, 1979), which will then influence peripheral clocks (Yamazaki et al., 2000). These cues, along with others in the periphery such as insulin and glucocorticoid signalling, and localised inflammation will all combine to define the circadian rhythms in other tissues (Leone et al., 2012, Balsalobre et al., 2000).

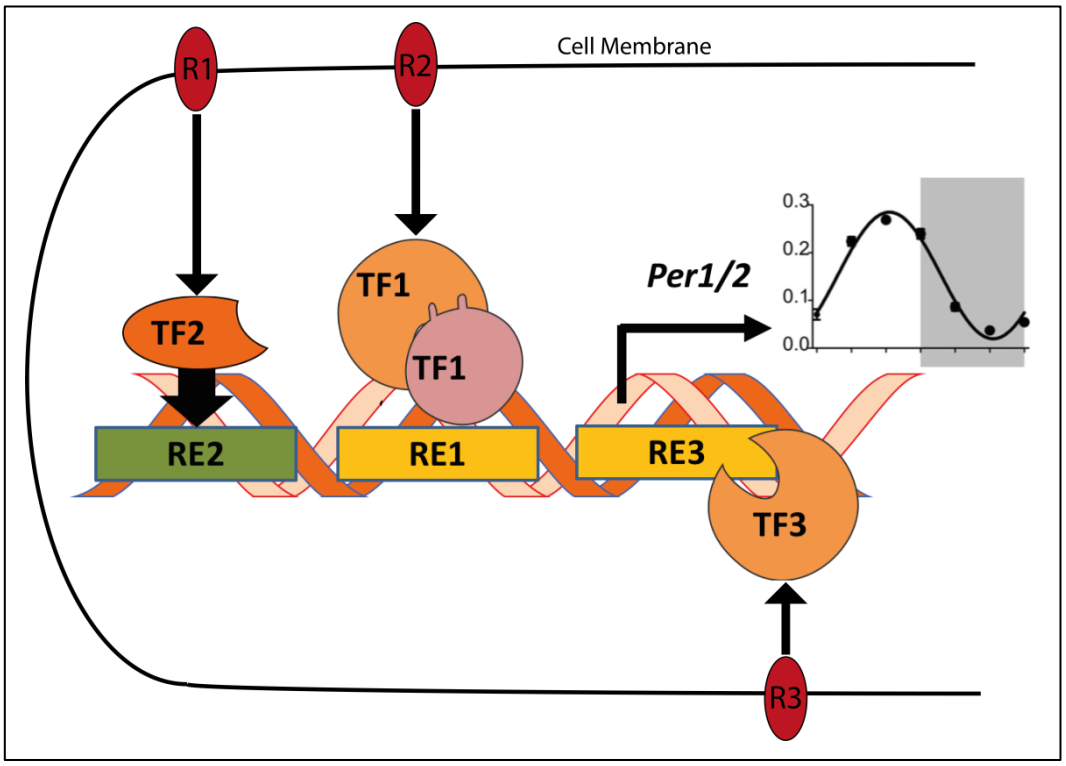


Figure 1.6: Multiple inputs integrate at the Promoter of *Per1/2*. In this example; input from three different membrane receptors (R) activate three different regulatory enhancers (RE) through three transcription factors (TF), which all ultimately converge at the promoter of *Per1/2*. This enables entrainment of the molecular clock through multiple factors, by varying *Per1/2* expression.

Due to this additional level of complexity, circadian rhythms within the periphery are not always exactly matched to those in the SCN, and frequently modern society works to exacerbate this discrepancy (Kervezee et al., 2018a). Desynchronization between the SCN and the periphery often strongly contributes to many of the negative health implications resulting from today's society (Nedeltcheva et al., 2009, Kervezee et al., 2018b) . To help elevate this desynchronization, it is more important than ever to understand the complex molecular pathways of the SCN. By ensuring that the SCN and periphery are quickly realigned following disruption from activities such as shift work, we can help to mitigate the negative health implications of modern life (Hower et al., 2018).

1.6: Circadian Rhythms and sleep

An individual's daily sleep/wake routine is regulated by a combination of both their daily circadian rhythm, and homeostatic sleep pressure resulting from their recent sleep history, this is known as the two process model of sleep (Figure 1.7; Borbély, 1982). Homeostatic sleep pressure and circadian rhythms must be in perfect synchronization for an individual to attain the optimum quality and length of sleep. Frequently, it is when these two processes fall out of synchronisation that an individual begins to develop sleep problems, and subsequent health problems commonly associated with an insufficient amount of poor quality sleep (Franken, 2013).

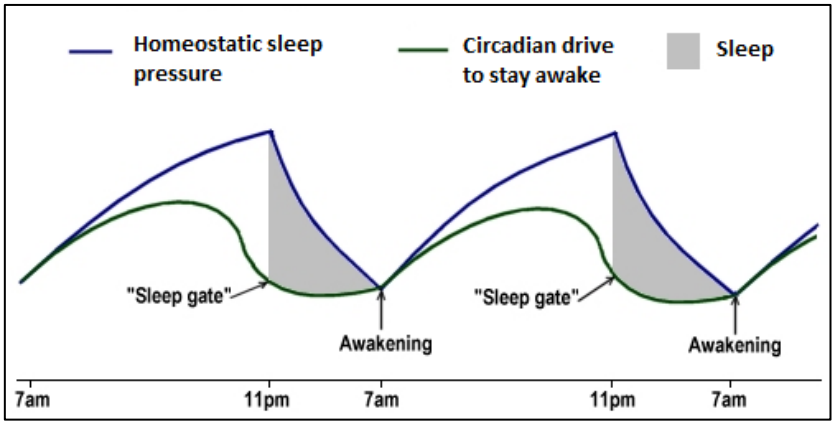


Figure 1.7: **The two process model of sleep.** As sleep pressure begins to rise, which causes an individual to feel tired, the circadian drive to stay awake also rises preventing an individual from falling asleep at the wrong time of the day. In the evening when sleep pressure is at its highest, the circadian drive begins to drop, enabling an individual to fall asleep in the late evening, 11pm in this example. Once an individual falls asleep, sleep pressure quickly falls, until around 7am in this example, at this point the circadian drive is once again high enough to cause arousal from sleep.

At the molecular level, it is still disputed what exactly drives homeostatic sleep pressure and therefore causes an individual to feel tired. However, in addition to light input, other stimuli may also play a role in circadian entrainment within the SCN, and an individual's sleep history has been shown to have an effect on the transcription of *Per1/2 (*Wisor et al., 2002*)*. Indeed when an individual is more tired, for example in the late evening, light input via the RHT will have a diminished effect on increasing *Per1/2* in the SCN (Khalsa et al., 2003). Light input to the SCN for a diurnal individual will have the greatest effect on *Per1/*2 induction first thing in the morning when levels of homeostatic sleep pressure are at their lowest. This therefore indicates that on a molecular level homeostatic sleep pressure and the circadian molecular clock are intertwined. The Jagannath research group has recently made associations between sleep history and the circadian clock at the molecular level, demonstrating that greater levels of adenosine within the brain attenuate how robustly the SCN will react to light input (Jagannath et al., 2021). During the active/wake period when sleep pressure increases, there is an accumulation of adenosine in the brain as a result of wakeful metabolism; these levels are then restored to normal when an individual sleeps. Adenosine therefore gives a good molecular indicator of sleep pressure; experiments have shown the action of increased adenosine on $A1/A2_A$ receptors of the SCN can inhibit the molecular mechanisms of entrainment already described here. This would indicate at the molecular level when sleep pressure is high, entrainment by the SCN following light input via the RHT will be diminished, and this is exactly what occurs.

1.7: Health consequences of circadian rhythm disruption

An individual's circadian rhythm helps ensure many biological processes are fully synchronised with their external environment, making certain these processes are optimised to occur at precisely the right time of the day. For example the digestive system will be fully activate during the day, at a time when an individual is most likely to eat food, with all genes required for expression of digestive enzymes being fully switched on. Disruption to the normal light/dark cycle has been shown to lead to a variety of health problems including, obesity, diabetes and heart disease (Shi et al., 2013). This has particular significance for individuals whose daily lives cause unavoidable disruption to their normal circadian rhythm. Such examples include, shift workers, workers who are 'on call' at night,

individuals who continually travel to different time zones, and those who have to take medication that causes circadian disruption.

There are currently very few treatment options available to those individuals whose daily lives inevitably lead to circadian disruption. This is despite the fact that the rectification of circadian disruption in numerous cases, will greatly improve the symptoms of many mental illnesses (Freeman et al., 2017), and more generally significantly improve the health of any individual.

This project aims to explore the underlying molecular mechanisms that directly control the molecular clock of cells within the SCN. By gaining a greater understanding of the molecular mechanisms underpinning entrainment of circadian rhythms, it is hoped that new drug targets can be identified. This in-turn will lead to the development of new treatments, to correct circadian rhythm mis-alignment, ensuring correct synchronisation to the external environment.

1.8: Chromatin structure and its role in daily circadian gene regulation

In eukaryotic cells, chromatin is a complex of DNA and proteins, with its primary function to package long strands of DNA into small compact, structures (Kornberg, 1974). The total length of human DNA in each cell is approximately 2 metres long; therefore this must be condensed to fit into a nucleus, with a diameter of only 6 μm. The major protein components of chromatin are histones (Kelley, 1973); these bind to DNA and act as anchors around which DNA strands are wound. DNA wraps around histones like cotton around a bead, forming nucleosomes, giving chromatin its most fundamental structure euchromatin. Chromatin is then further structured into small nucleosome arrays of multiple histones wrapping into a 30 nm fibre called heterochromatin. Further higher-level supercoiling of this fibre produces the chromosomes, often the most familiar DNA structure, Figure 1.8 details the different levels of chromatin organisation.

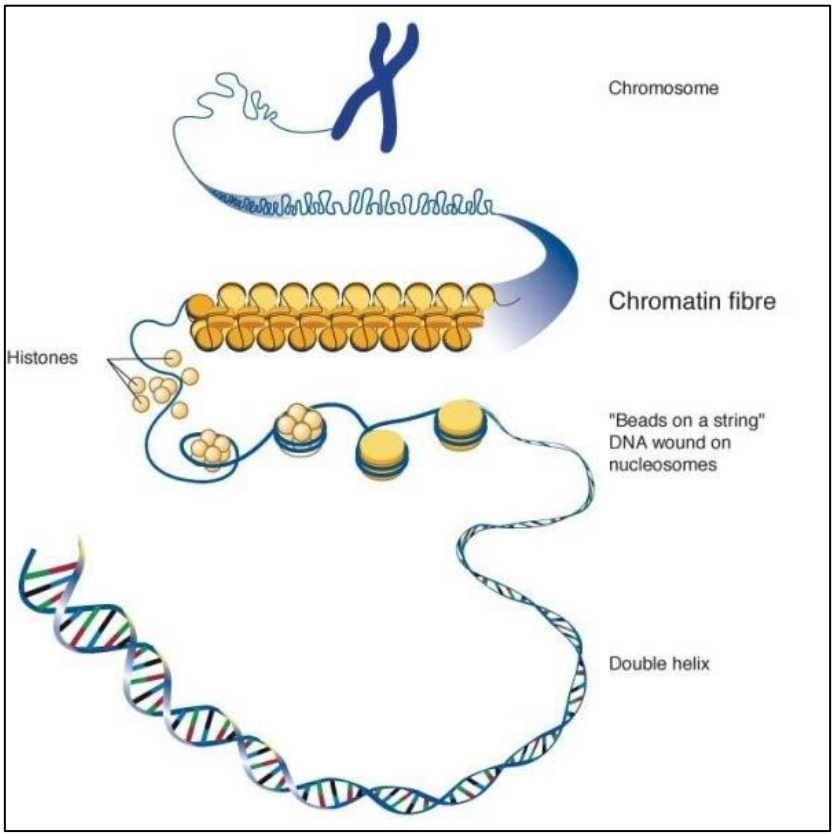


Figure 1.8: The different levels of chromatin organisation. The DNA double helix is wrapped around histone proteins into nucleosomes; these are then ordered into chromatin fibre and ultimately the common metaphase chromosome itself. (Figure from The National Human Genome Research Institute)

Chromatin not only helps to fit DNA into the nucleus, it also prevents DNA becoming tangled, has an important role in reinforcing DNA during cell division, aids in preventing DNA damage, and plays a role in regulating gene expression and DNA replication (Chung et al., 2016). Chromatin structure helps ensure the correct areas of the genome are physically open, so transcription factors can bind, and makes certain DNA is orientated is the correct direction, so the sense strand is arranged correctly for transcription to begin. Every cell in the human body has exactly the same DNA, despite carrying out very divergent roles, clearly a hepatocyte functions very differently from a cardiomyocyte, which is equally as distinct from a neuron. It is the structure of the chromatin that in a large part allows certain genes to be expressed in one cell but not another, areas of DNA coding for genes essential for correct function of the liver will be open and easily accessible within hepatocytes, but inaccessible where it is not required, such as within a neuron.

The structure of chromatin within the nucleus changes significantly during mitoses/meiosis, but even within a cell that is not actively replicating, chromatin structure does not remain static, it is instead highly dynamic. Chromatin structure not only varies between different cell types, allowing cells in different tissues to express different genes (Kamakaka and Thomas, 1990), but can also vary within cells of the same tissue. Fine tuning of gene regulation can be achieved through continuous chromatin re-arrangement; areas of the genome usually open are sometimes wrapped up into histone proteins and closed off, so therefore become inaccessible for transcription (Bodnar and Bradley, 1996). This frequently occurs at the promoters and other regulatory elements of genes, therefore preventing transcription factors and other trans-regulatory factors from directly binding DNA. This continuous opening and closing of chromatin throughout the day is the first point of transcriptional regulation, if a promoter or enhancer is physically closed off, it will likely have a significant effect on the transcription of the gene it is associated with (Heintzman et al., 2007). Regardless of whether transcription factors are present to switch on a gene, if they cannot bind their binding site due to inaccessibility, they cannot initiate transcription.

1.9: Daily changes in chromatin structure

Given that the molecular clock is conserved and functional in all cells, how it achieves temporal and tissue-specific regulation is an important question. It is hypothesised that one way this may be accomplished is by differential regulation of chromatin accessibility of clock protein binding motifs. Recent experiments using chromatin immunoprecipitation sequencing (ChIP-seq) in pancreas and liver have shown tissue-specific regulation can be achieved by distinct DNA binding patterns for the clock proteins in different tissues. Perelis et al. (2015) showed that in pancreatic tissue CLOCK/BMAL1 bind to specific sites at three distinct regulatory regions at the Nr1d1 locus, including promoter and enhancer regions. However, a previous study showed that CLOCK/BMAL1 did not bind to the same regions in hepatocytes (Fang et al., 2014), therefore demonstrating tissue specific regulation.

In fact in the liver only 22% of cycling mRNAs are driven by transcription (Koike et al., 2012). In addition to cycling genes, epigenetic marks also show persistent patterns of cycling and include H3K4me3, H3K9ac and H3K27ac. H3K4me3 indicates methylation of histones, and this usually leads to the closing up of chromatin and therefore the repression of gene expression. Both H3K9ac and H3K27ac indicate acetylation of histones, which usually has the opposite effect, the opening up of chromatin, enabling gene transcription to take place. Figure 1.9 below shows the most common cycling genes in the liver.

Furthermore, although CLOCK/BMAL1 binding to promoters has been shown to contribute to cycling chromatin reorganization, in combination with chromatin re-modellers (Koike et al., 2012), this does not provide sufficient explanation for such diversity of circadian outputs among different tissues. Recent findings have revealed the presence of circadian interactomes; these appear as organised hubs within the genome where circadian genes are expressed in a co-ordinated manner on a daily basis (Aguilar-Arnal et al., 2013). Within these hubs, clock controlled transitions within the structure of chromatin contribute to the synchronization of circadian gene expression through the physical association and accessibility of genes (Xu et al., 2016).

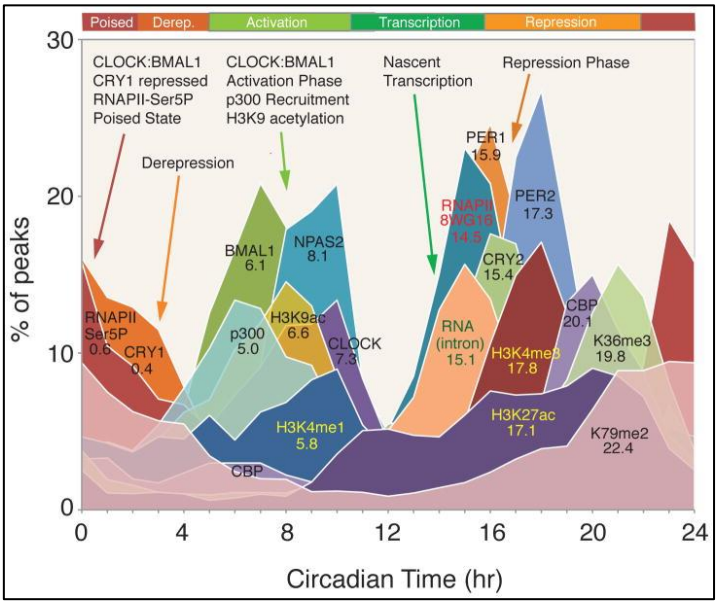


Figure 1.9: Cycling genes in the liver. The core clock genes, *Clock, Per1/2, Bmal1 and Cry1* can all be seen to be cycling. In addition several histone epigenetic marks including H3K4me3, K36me3 H3K4me1, H3K9ac and H3K27ac also appear to be highly rhythmic. Figure from Koike et al. (2012).

It is therefore likely that chromatin structure is playing a greater role in co-ordinated rhythmic output than initially anticipated through differential accessibility. The extent to which chromatin organization can influence the circadian transcriptome is still scarcely understood; therefore this project hopes to further elude how changes in chromatin accessibility may account for such a diversity of circadian outputs. Currently little is known about histone modification within the SCN, this project hypothesises that as H3K4me3, H3K9ac and H3K27ac all cycle within the liver, they will likely cycle within cells of the SCN too. With the SCN's role as the central pacemaker, any resulting changes to circadian rhythms from chromatin modification may impact how it entrains to differing light/dark cycles. It is therefore easy to foresee how changes in chromatin structure that leads to the up or down regulation of genes, will in turn have a substantial effect on such a complex molecular network as the whole molecular clock.

1.10: Epigenetic modifications

Epigenetic modifications are stable and heritable modifications in gene expression, attained by reversible changes affecting chromatin structure, usually without changes to the original DNA sequence. The two most common epigenetic modifications are DNA methylation and histone modification; both of these have a direct effect on DNA chromatin structure, leading to differential access of DNA, often closing or opening up particular areas of the genome to increase or decrease transcription factor binding.

Methylation of DNA occurs when a methyl group is added to a cytosine-guanine dinucleotide (CpG), and is considered the most frequent DNA epigenetic modification. It is accomplished by a family of enzymes called DNA methyltransferases (DNMTs), these include: DNMT1, DNMT2, DNMT3A, DNMT3B and DNMT3-like (DNMT3L). DNMT3A and DNMT3B are required for the creation of de novo DNA methylation marks (Okano et al., 1999), while DNMT3L has no enzyme activity of its own, but instead acts as a partner of DNMT3A and DNMT3B (Jia et al., 2007). Methylation occurs in both a dynamic manner changing on a day to day basis (Xia et al., 2015), and has been observed to accumulate to a greater extent as we age (Horvath and Raj, 2018). DNA methylation has already been

shown to change in key clock genes following SD (Cedernaes et al., 2015); this will therefore likely have a direct effect on the expression of these genes. Using ATAC-sequencing areas of significant chromatin change can be compared to areas of frequent DNA methylation, if changes are occurring in similar sites, this will confirm that changes in DNA methylation are having a direct effect on accessibility, and therefore likely to alter gene expression. The accessibility of DNMTs can also be examined, to determine if this is changing, and may indicate an increase or decrease in the activity of DNMTs, and therefore the level of any changes to DNA methylation.

Histone proteins pack and order DNA into organisational units called nucleosomes; therefore any modifications to histones will have a direct effect on both the structure and organisation of DNA. Histone modification has been well studied in the nervous system (Pattaroni and Jacob, 2013); acetylation of histones weakens their interaction with DNA, and therefore leads to a more relaxed chromatin conformation, which facilitates recruitment of transcriptional machinery. Histones can be acetylated by histone acetyltransferases (HATs) (Allfrey et al., 1964), phosphorylated/ dephosphorylated by kinases and phosphatases (Oki et al., 2007), and methylated by methyltransferase (HKMT) (Ng et al., 2008). Both acetylation and phosphorylation will change the charge of the histone protein, while methylation of histone proteins will not, but amino acids can be methylated multiple times. Histone modification has even been revealed to play an important role in long term memory formation in the hippocampus (Levenson et al., 2004), and has already by shown to influence the molecular clock (DiTacchio et al., 2011). The chromatin structure of enzymes involved in histone modification will be investigated, and the level of chromatin accessibility will be assayed in areas of the genome where histone modification has been shown to frequently occur.

1.11: Non-coding RNA's

Further to epigenetic modification, additional transcriptional control is accomplished by non-coding RNA's, these are encoded from DNA, but are never translated, yet despite this they can still have a significant effect on the expression of other genes. A class of non-coding RNA, long non-coding RNAs (lncRNAs) are greater than 200bp long, and their transcription has been shown to be

significantly affected by SD (Soshnev et al., 2011, Davis et al., 2007). The majority of lncRNAs have been found in the brain (Derrien et al., 2012). Although non coding RNA's are not transcribed, they are still translated from DNA in the same manner as other genes; therefore the accessibility of DNA determined by chromatin structure may well influence their levels of translation.

1.12: Hypothesis 1: Light pulsing induces structural chromatin changes in the SCN

The SCN is the primary conduit for environmental time signals and time-keeping is its chief role. Therefore it is vital to understand how chromatin and promoter accessibility of key molecular clock components and CCGs within the SCN change across the day, and in response to input signals such as light. However, in contrast to peripheral tissues, relatively little is known about such changes within the SCN. This is partly due to the small size of the structure, consisting of only around 20,000 neurons, making techniques such as ChIP-Seq or DNAse1 hypersensitivity very challenging. Despite this, Crosio et al. have shown that light causes rapid phosphorylation of S10 on histone 3, indicating light can remodel SCN chromatin (Crosio et al., 2000). Similarly, Azzi et al. have shown that non-24h light dark cycles can alter promoter DNA methylation within the SCN, leading to lasting changes in fundamental properties of the circadian rhythm, such as period length (Azzi et al., 2014). These studies suggest that the SCN chromatin landscape is dynamic and can influence the SCN's timekeeping properties.

Within the SCN genes associated with light input and transmission of information, are up or down regulated throughout the day. As previously described, the induction of PER and FOS, via light input can lead to shifts in the molecular clock. However, these shifts can only occur in the most light-responsive times of the day (Ralph and Menaker, 1988); for a diurnal human this occurs largely in the morning at dawn, while for a nocturnal mouse it is mainly at dusk. Very little is known about how these mechanisms are switched on or off at different times of the day; it is hypothesised that daily chromatin re-modelling may play a role. Daily chromatin re-modelling has already been shown in both the liver and pancreas, and it is therefore not unreasonable to propose that it may also be

occurring in the SCN. However, in this case instead of simply preventing genes being switched on at the wrong time of the day, it could potentially prevent the genes being switched on that are actually able to entrain the molecular clock itself. This would therefore enable the molecular clock to dynamically respond to light input only at the correct time of day.

This project hypothesises that promoter and enhancer sequences of CLOCK regulated proteins in the SCN, where binding motifs such as the CRE or AP1 motif are located, may be opened up or closed within chromatin, depending on the time of the day. This would enable genes required for entrainment to be in an open state ready for transcription initiation at the most light responsive time of day, whilst at other times of the day to be closed preventing transcription from taking place.

1.13: Hypothesis 2: Chromatin changes follow sleep deprivation in the SCN, dentate gyrus and cortex.

Due to the fact entrainment within the SCN occurs primarily in the morning for diurnal animals, this time of the day is also when sleep pressure is at the lowest. Therefore it is reasonable to hypothesize that increased sleep pressure later in the day may inhibit increased *Per1/2* transcription that is usually initiated by light input to the SCN. The mechanism by which this process occurs is still largely unknown; this project hypothesises an increase in sleep pressure during the course of the day may itself initiate further changes to chromatin structure. Recent discoveries by the Jagannath research group describing a link between adenosine and sleep pressure, provide an ideal candidate for further consideration when investigating sleep pressure. It is further hypothesised that changes in chromatin structure result in regulatory elements of genes associated with entrainment in the SCN becoming more responsive to transcription factors downstream of the $A1/A2_A$ receptors. This may therefore have a negative effect on gene transcription, and result in an attenuation of the increased *Per1/2* induction usually associated with light input to the SCN.

SD has already been shown to initiate chromatin remodelling in the murine cortex (Hor et al., 2019), so it is therefore highly probable SD may cause such changes to occur in other brain regions including

the SCN. Chromatin remodelling has already been demonstrated in the dentate gyrus (DG) following electrical stimulation (Su et al., 2017), and therefore this brain region therefore provides a further candidate to investigate chromatin changes following SD. Indeed many existing studies looking at SD have focussed on the hippocampus where the DG is located. At the molecular level SD has been shown to completely change epigenetic status (Trivedi et al., 2017, Duan et al., 2016), and have a significant effect on memory formation and hippocampal function (Yoo et al., 2007). Therefore any data generated about chromatin accessibility changes following SD in the DG, can be viewed in the full context of the existing literature.

It is also surmised that insights into why we actually sleep may be revealed; there are still many unanswered questions surrounding the role of sleep, with most theories centring on the idea of memory consolidation and cellular maintenance and repair. From an epigenetic perspective, many studies indicate sleep is very important for DNA repair (Zada et al., 2019), and chromatin structure must change to facilitate this process (Surrallés et al., 1998). Therefore by assaying chromatin accessibility, insights may be revealed relating to DNA repair occurring during sleep.

Following SD, sleep pressure substantially increases; it therefore provides the ideal experimental condition for investigating potential chromatin remodelling occurring during sleep, and at times of increased sleep pressure. This project will investigate chromatin changes following SD in the cortex, the SCN, and DG; this will provide insights into epigenetic modification occurring across different brain regions following SD.

1.14: Methods to assay chromatin structure

A number of methods are currently available to assay chromatin structure, and these include, CHIP-seq, DNase I hypersensitive sites Sequencing (DNase-seq), Formaldehyde-Assisted Isolation of Regulatory Elements sequencing (FAIR-seq), and Chromosome conformation capture (3C). All these techniques require far more starting material than is available from even several pooled mouse SCNs. Therefore, the number of mice required to carry out any of these techniques would be both impractical

from an experimental point of view, and likely be unethical due to the number of animals required to be sacrificed.

The project will therefore take advantage of a recent next generation sequencing technique (NGS); Assay for Transposase-Accessible Chromatin sequencing (ATAC-seq; Buenrostro et al., 2015), to investigate chromatin structure in the SCN. By using ATAC-seq it is possible to assay chromatin structure of a far smaller number of cells, small enough that a pool of only 4 mouse SCN should be sufficient. In fact the technique has been successfully used to assay chromatin from only 500 cells, although to attain high quality data, it is likely many times this number of cells will be required.

ATAC-seq makes use of transposase enzymes; these occur naturally within nuclei of eukaryote cells, but usually only move transposable elements of DNA within the genome. However, for ATAC-seq a hyper active transposase enzyme Tn5 is used. This enzyme will effectively cut out any open areas of DNA within chromatin, rather than just the small number of transposable elements (Goryshin and Reznikoff, 1998). These fragments of DNA are tagged with sequencing adapters as they are cleaved (Adey et al., 2010), and barcode sequences are then added when the fragments are amplified using Polymerase Chain Reaction (PCR). The fragments can then be sequenced, and aligned to the mouse reference genome; the relative number of sequencing reads that align to particular areas of the genome can then be used to determine levels of chromatin openness (Buenrostro et al., 2013). It is expected that areas of the genome where multiple DNA fragments align in the same location are likely to be the most open, as a larger number of fragments would have been originally cut from these areas by the Tn5 enzyme. The ATAC-seq reaction is shown below in Figure 1.10.

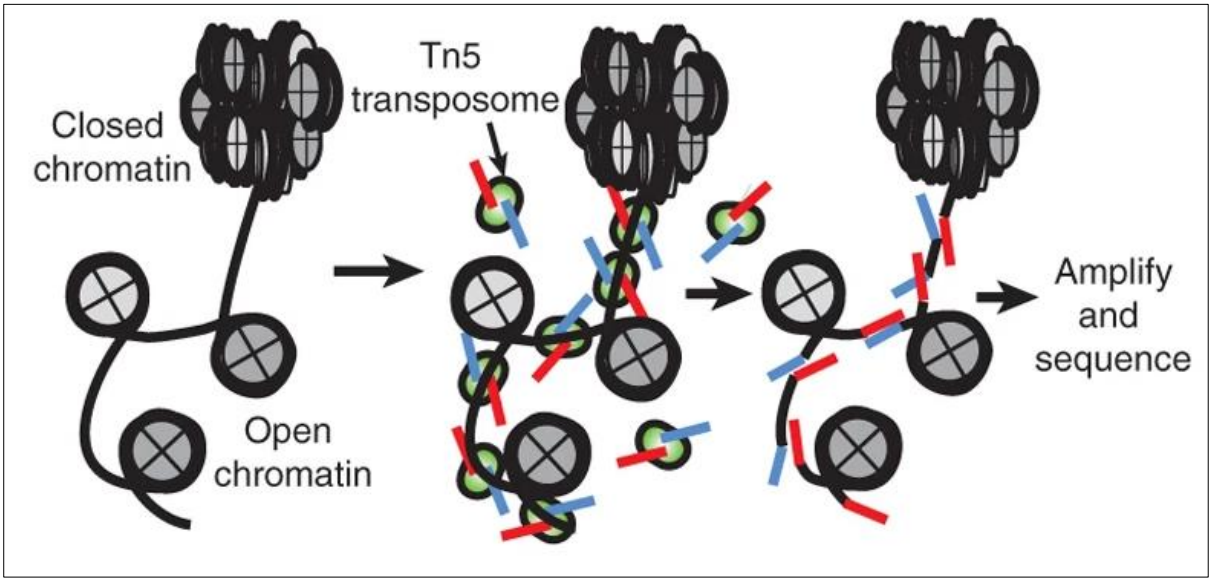


Figure 1.10: ATAC-seq reaction schematic. Tn5 enzyme (green), which has been loaded with sequencing adaptors (red and blue), inserts in regions of open chromatin between nucleosomes (grey) and generates DNA fragments with adapter sequences which can be PCR-amplified. Taken from Buenrostro et al. (2013).

1.15: Bioinformatic Analysis

ATAC-seq will generate a large amount of data for analysis; currently there are many programs and pipelines available to analyse data from RNA-sequencing (RNA-seq), but for ATAC-seq these are more limited. It is likely for this project that a combination of stand-alone programs, pipelines, and scripts written de-novo will be required. Existing programs for RNA-seq will have to be adapted for ATAC-seq, and any results derived from a program originally written to analyse RNA-seq data will have to be carefully considered, to determine their likely accuracy. The platform upon which the data is analysed will have to be chosen carefully, laptops and desktop computers are likely to have insufficient processing power to analyse all the data in a timely manner. Therefore a high performance computer cluster will have to be used to analyse the data generated during this project.

1.16: Outline of result chapters and key questions

The first results chapter, chapter three investigates the feasibility of the project; through pilot experiments the technical approach of nuclei isolation was optimised for SCN tissue punches coupled with ATAC-seq. This technique had never been carried out before with using SCN tissue, and the process was new to our research group. This chapter answers one key question:7

- Can ATAC-seq by used to assay the chromatin structure of mouse SCN, given its small size?

Chapter four implements the techniques refined in previous chapter to investigate changes in chromatin structure following light exposure at specific times of the day, and the potential implications this may have on gene expression within the molecular clock, exploring the following questions:

- Does daily variation in chromatin structure within the SCN, predispose the clock to be light responsive at certain times of day, and is there differential chromatin remodelling that occurs in response to light input at different times of day?

- Are circadian entrainment pathways down regulated by chromatin structural changes at certain times of the day?

- Can new genes and pathways that regulate circadian entrainment, be identified by daily variations in chromatin structure?

- Do changes in chromatin structure result in changes to the transcriptome?

The fifth chapter aims to understand whether SD has an effect on chromatin structure, and does this by asking the following fundamental questions:

- Does SD lead to changes in chromatin structure, and are these changes responsible for preventing shifts to the molecular clock at certain times of the day?

- Are changes in chromatin structure as a result of SD, leading to changes in gene expression?

The final chapter presents preliminary data generated to assist with a Covid-19 research project. Due to the pandemic, all lab work was halted, and results from a COVID bioinformatic analysis to contribute to the pandemic response are presented in this chapter. The analysis aimed to access the feasibility of a project investigating the importance of circadian regulation of Angiotensin converting enzymes in lung epithelial cells.

Chapter 2

Main Methods

This project takes advantage of a new NGS technique ATAC-seq, and the more common technique RNA-seq, these are two of a number of recently developed methodologies, which take advantage of the latest advances in DNA sequencing. NGS experiments are essentially broken down into two sections; firstly sequencing libraries must be generated using a variety of molecular biology techniques, often requiring extensive optimisation of existing experimental techniques. For ATAC-seq; nuclei are first isolated from dissociated cells, and hyper active transposase enzyme Tn5 is then used to cut out all open DNA from the nuclei, whilst simultaneously tagging DNA with adapter sequences. These tagged DNA sequences are then amplified using the polymerase chain reaction (PCR), and barcode sequences are added, giving the libraries, which can then be sequenced. The full experimental workflow for ATAC-seq is shown in Figure 2.1. For RNA-seq, the process is simpler; cDNA is generated from RNA, and then subsequently used to generate NGS libraries as shown in Figure 2.2.

These libraries are sequenced; which generates a large quantity of data, the second part of a NGS experiment is the subsequent extensive analysis of this data, using a variety of bioinformatic techniques. For this project both parts of this process were carried out in-house; this chapter describes the methods required for the laboratory techniques, before moving onto the bioinformatic techniques.

2.1: Nuclei Preparation

The first aim of the project was to refine existing ATAC-seq protocols to work with SCN tissue; several protocols were followed in an attempt to achieve this goal. These included enzymatic digestion, sonication, and douncing using a pestle and mortar.

2.1.1: Extraction of single cells from brain tissue using enzymatic digestion and sonication

The first method tested to obtain nuclei, attempted to use enzymatic digestion to dissociate SCN tissue to individual cells, before lysing the cells using the ATAC lysis buffer from the published protocol to attain nuclei. Cells were examined immediately following digestion to ascertain viability, and then lysed using the published ATAC seq lysis buffer if viable cells were present. The following protocol was followed for each enzyme tested:

i) **Brain removal and Cortex punch:** Brains were removed and placed into a brain matrix (Kent Scientific, Torrington CT, USA). A skin graft blade (Swann-Morton, Sheffield, UK) was positioned at Bregma −0.10 mm (Figure S1). A second blade was placed 1 mm (Bregma −1.10) caudal from the first, and a 1 mm thick brain slice was dissected. Cortex punches were taken using a sample corer (1 mm internal diameter, Fine Science Tools GmbH, Heidelberg, Germany) from the brain slice (n = 4).

ii) **Enzymatic incubation:** 4 brain punches were enzymatically digested at 37°C using 200 µl of Trypsin, Collagenase, Dispase or papain (0.2 mg/ml), with/without gentle agitation by pipetting up and down with a p1000 pipette, for 5, 10, 15 or 20 min.

iii) **Centrifugation:** 500 µl PBS was added to the sample, and the sample was spun at 500g for 10 min, cells were collected as a pellet.

- Either proceed to step 4 to immediately examine the cells following digestion, or straight to step 5 to carry out cellular lysis.

iv) **Cell viability:** Cells were checked under a microscope using a microscope slide with coverslip at 10x and 20x zoom**.**

v) **Centrifugation and nuclei extraction:** 500 µl ATAC lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630) was added to the sample, and the sample was spun at 500g for 10 min, the nuclei were collected as a pellet.

vi) **Nuclei viability:** Pellet was re-suspended in 100µl PBS with 1 µl ThermoFisher NucBlue™ Live ReadyProbes™ Reagent added. Nuclei were viewed fluorescing under ultra violet light, using a microscope slide with coverslip at 10x and 20x zoom.

vii) Viable cells were subsequently counted using a haemocytometer.

The second approach to dissociate tissue entailed using a tissue homogenizer to physically break up the cells into a single cell suspension, or straight to nuclei. The Bertin Precellys24 Instrument was used; this method homogenised the tissue for varying lengths of time at varying pulse frequencies, to see which combination best breaks down the tissue. Samples were examined under the microscope straight after sonication, and following lysis with the published ATAC seq lysis buffer.

The following protocol was followed:

i) **Brain removal and Cortex punch:** Brains were removed and placed into a brain matrix (Kent Scientific, Torrington CT, USA). A skin graft blade (Swann-Morton, Sheffield, UK) was positioned at Bregma −0.10 mm (Figure S1). A second blade was placed 1 mm (Bregma −1.10) caudal from the first, and a 1 mm thick brain slice was dissected. Cortex punches were taken using a sample corer (1 mm internal diameter, Fine Science Tools GmbH, Heidelberg, Germany) from the brain slice (n = 4).

ii) **Sonication:** Samples were sonicated at varying pulse frequencies using the Precelleys24 sonicator.

iii) **Centrifugation:** 500 µl PBS was added to the sample, and the sample was spun at 500 g for 10 min, nuclei/cells were collected as a pellet.

- Either proceed to step 4 to immediately examine the cells/nuclei following sonication, or straight to step 5 to carry out cellular lysis.

iv) **Cell/Nuclei viability:** Nuclei/cells were checked under a microscope using ThermoFisher NucBlue™ Live ReadyProbes™ Reagent, as above.

v) **Centrifugation and nuclei extraction:** 500 µl ATAC lysis buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% IGEPAL CA-630) was added to the sample, the pellet was re-suspended by pipetting up and down, and the sample was spun at 500 g for 10 min, the nuclei were collected as a pellet.

vi) **Nuclei viability:** Nuclei were checked under a microscope using ThermoFisher NucBlue™ Live ReadyProbes™ Reagent, as above.

vii) Viable nuclei were subsequently counted using a haemocytometer.

2.1.2: Nuclei Extraction straight from tissue using Dounce homogenizer

Initial experiments using enzymatic dissociation of tissue, showed this to be a slow and unreliable approach that resulted in poor library preparation. To overcome this, a new approach was developed; SCN were homogenized directly from tissue to nuclei, rather than individual cells first, a similar approach to the Bertin Precellys24 was applied but using a new protocol. Dissociation directly to nuclei was achieved by modifying the nuclei isolation protocol from Su et al. (2017), using a Dounce homogenizer and sucrose cushion. The Dounce homogenizer quickly broke down the SCN tissue to nuclei, and the sucrose cushion provided a concentration gradient that would allow a pellet of nuclei to form, while leaving cellular debris suspended. The viability of nuclei was again checked under a microscope using ThermoFisher NucBlue™ Live ReadyProbes™ Reagent.

The following protocol was followed for nuclei isolation:

i) **Brain removal and punch:** Protocol from section 2.11 followed, with 4 SCN punches taken instead of cortex.

ii) **Tissue homogenization:** SCN punches were processed using a Dounce homogenizer (Sigma, D8938) in 1 ml lysis buffer (20 mM Tricine-KOH, 25mM $MgCl_2$, 0.25 M sucrose, 1 mM DTT, 0.15 mM spermine, 0.5 mM spermidine, 0.3% IGEPAL-630, EDTA-free protease inhibitor) for 10 strokes.

iii) **Centrifugation:** Homogenate was then filtered through a 40 µm strainer into a 50 ml falcon tube, it was then transferred to a 1.5 ml epindorf tube, under layered with a cushion buffer (0.5 mM $MgCl_2$, 0.5 mM DTT, EDTA-free protease inhibitor, 0.88 M sucrose), and centrifuged at 2800 g for 10 min at 4°C. The pellet was re-suspended and washed in 500 µl cold PBS, then re-spun at 700 g for 10 min. Nuclei were collected from the resulting pellet.

iv) **Cell viability:** Cells were checked under a microscope using ThermoFisher NucBlue™ Live ReadyProbes™ Reagent, as per section 2.11.

v) Viable nuclei were subsequently counted using a haemocytometer

2.2: Preparation and Sequencing of ATAC-seq DNA libraries

Following successful isolation of intact nuclei, ATAC-seq libraries were prepared using the Tn5 transposase with the following protocol:

i) Transposition:

1. The reaction mix was prepared as follows:

    - 25 µl 2x TD Buffer (Illumina Cat #FC-121-1030)
    - 2.5 µl Tn5 Transposes (Illumina Cat #FC-121-1030)
    - 22.5 µl Nuclease Free H_2O

2. Nuclei from 4 punches, containing approximately 40,000 nuclei were re-suspended in this mix and incubated at 37°C for 30 min for transposition to occur.

3. DNA was then purified using the Qiagen MinElute Kit and eluted with 23µL warm 10mM Tris, pH8 and stored at -20°C.

ii) PCR Amplification:

1. To amplify transposed fragments and attach the Illumina adapters, the following was combined in a PCR tube:

    - 10 µl Transposed DNA
    - 10 µl Nuclease Free H_2O
    - 2.5 µl 25 µM Customized Nextera PCR Primer 1 (Barcodes given in supplementary data)
    - 2.5 µl 25 µM Customized Nextera PCR Primer 2 (Sigma Barcode Ad1_noMX)
    - 25 µl Fast Sybr Green 2x PCR Master Mix (Qiagen Cat #4385612)

2. Cycle using the following conditions on an Applied Biosystems qPCR machine:

    (1) 72°C, 5 min
    (2) 98°C, 30 s
    (3) 98°C, 10 s
    (4) 63°C, 30 s
    (5) 72°C, 1 min
    (6) Repeat steps 3-5 for as many cycles as appropriate.

3. Hold at 4°C

2.3: RNA Extraction and RNA sequencing library preparation

RNA-seq, when required was carried in addition to ATAC-seq to investigate how the transcriptome changes, this is because changes to chromatin structure will likely have a direct effect on the transcriptome. RNA-seq therefore provides a method to investigate how changes in chromatin structure may relate to the transcriptome.

A total RNA-seq protocol was carried out to sequence all RNA with the exception of ribosomal RNA; this gives a good insight into both which genes are being expressed, and potentially the expression of non-coding RNAs, which may also be having an effect on transcription.

RNA was first extracted using the following protocol:

i) **Brain removal and SCN punch:** Protocol from section 2.11 followed, with 2 punches taken per sample.

ii) **Tissue homogenization and RNA stabilisation:** Tissue was homogenised in 100 µl Qiagen QIAzol Lysis Reagent (Cat. No. 79306) using a Microtube Homogenizer (Bel-Art F65000-0000) for 10 seconds, and incubated for 5 min at room temperature.

iii) **Phase separation using chloroform:** An additional 400 µl Qiagen QIAzol Lysis Reagent added, along with 100 µl chloroform, sample shaken vigorously, incubated for 10 min at room temperature, and spun at 12,000 g for 15 min.

iv) **RNA purification:** Clear band containing RNA carefully removed and purified using the Qiagen RNeasy Micro Kit (Cat. No. 74004); genomic DNA was removed using DNase.

v) **RNA quantity/quality check:** RNA quantity and RNA integrity number (RIN) was determined using The Agilent 4150 TapeStation, with high sensitivity RNA ScreenTape (Cat. No. 5067-5579), samples with a RIN above 8.5 were considered of high quality.

Following RNA extraction, Illumina RNA sequencing libraries were prepared using the Illumina Stranded Total RNA kit (Cat. No. 20040529). The following protocol was used:

i) **RNA concentration normalisation**: All samples normalised in 10µl to an appropriate concentration for the libraries being prepared at the same time.

ii) **Ribosomal RNA depletion:** Ribosomal RNA depleted using Illumina Ribo-Zero Plus rRNA Depletion Kit (Cat. No. 20037135):

1. Hybridization of DNA probes complementary to rRNA to target depletion enzyme; per sample 3.6 µl depletion probe buffer mixed with 1.2 µl depletion Probes. Sample incubated at 95°C, and temperature reduced by 0.1°C per second until the temperature reaches 37°C.

2. Enzymatic digestion of rRNA; per sample, 4.8 µl Ribo Depletion buffer was mixed with 1.2 µl Ribo Depletion enzyme, sample incubated at 37°C for 15 min.

3. Probe removal by DNase digestion; per sample, 7.7 µl probe removal buffer mixed with 3.3 µl DNase I probe removal enzyme, sample incubated at 37°C for 15 min and then 70°C for 15 min.

iii) **Clean up RNA using beads:**

1. 60 µl Agencourt RNAClean XP (Cat. No. A63987) added to each sample; incubate at room temperature for 5 min.

2. Place on magnetic stand for 5 min, remove 80µl supernatant, keep on the magnetic stand and wash with 175 µl fresh 80% ethanol, remove ethanol and wash a second time, remove ethanol and air-dry on the magnetic stand for 2 min.

3. Add 10.5 µl elution buffer to each well, and pipette until the beads are re-suspended.

4. Incubate at room temperature for 2 min, place on the magnetic stand and incubate for 2 min, transfer 8.5 µl supernatant to the next step.

iv) **Fragment, denature and prime RNA;** this ensures RNA fragments are small enough to be sequenced, the RNA is denatured and primed with random hexamers ready for single strand cDNA synthesis:

Add 8.5 µl Illumina Elute, Prime, and Fragment High Mix to each sample, and incubate at 94°C for 2 min.

v) Synthesize First Strand cDNA; this step reverse transcribes the hexamer-primed RNA fragments to produce the first complementary cDNA:

1. Add 9µl Illumina first strand synthesis mix and 1 µl reverse transcriptase to each sample.
2. Run the following program on the thermocycler:
    - 25°C for 10 min
    - 42°C for 15 min
    - 70°C for 15 min

vi) Synthesize second Strand cDNA; this step removes the RNA template and synthesizes a replacement strand to generate blunt-ended, double-stranded cDNA fragments.

Add 25 µl Second Strand Marking Master Mix, and incubate at 16°C for 1 hour.

vii) Clean up cDNA:

1. Add 90 µl Agencourt AMPure XP beads (Cat. No. A63880) to each sample; incubate at room temperature for 5 min.
2. Place on magnetic stand for 5 min, remove 130µl supernatant, keep on the magnetic stand and wash with 175 µl fresh 80% ethanol, remove ethanol and wash a second time, remove ethanol and air-dry on the magnetic stand for 2 min.
3. Add 19.5 µl elution buffer to each well, and pipette until the beads are re-suspended.
4. Incubate at room temperature for 2 min, place on the magnetic stand and incubate for 2 min, transfer 17.5 µl supernatant to the next step.

viii) Adenylate 3′ Ends; this step adds an adenine nucleotide to the 3′ ends of the blunt fragments to prevent them from ligating to each other during adapter ligation. A corresponding thymine nucleotide on the 3′ end of the adapter provides a complementary overhang for ligating the adapter to the fragment:

Add 12.5 µl Illumina A-Tailing mix to each well and mix well, and incubate for 37°C for 30 min and 70°C for 5 min.

ix) Ligate Anchors; this step adds pre-index anchors to the ends of the double-stranded cDNA fragments to prepare them for dual indexing:

1. Add the following in exactly this order:

    - 2.5 µl resuspension buffer to each sample
    - 2.5 µl RNA Index Anchors to each sample
    - 2.5 µl Illumina ligation mix to each sample

2. Incubate samples at 30°C for 10 min

3. Add 5 µl Illumina stop ligation mix to each sample and mix well.

x) Clean Up Fragments:

1. Add 34 µl Agencourt AMPure XP beads to each sample; incubate at room temperature for 5 min.

2. Place on magnetic stand for 5 min, remove 67 µl supernatant, keep on the magnetic stand and wash with 175 µl fresh 80% ethanol, remove ethanol and wash a second time, remove ethanol and air-dry on the magnetic stand for 2 min.

3. Add 22 µl resuspension buffer to each well, and pipette until the beads are re-suspended.

4. Incubate at room temperature for 2 min, place on the magnetic stand and incubate for 2 min, transfer 20 µl supernatant to the next step.

xi) Amplify Library

1. Add the following in exactly this order:

    - 10µl Illumina index primer mix containing both i5 and i7 indexes
    - 20µl Illumina enhanced primer mix

2. Run the following program on the thermocycler:

    i. 98°C for 30 seconds

    ii. 14 cycles of: 98°C for 10 seconds, 60°C for 30 seconds and 72°C for 30 seconds

    iii. 72°C for 5 min

xii) Clean up amplified library

1. Add 50 µl Agencourt AMPure XP beads to each sample; incubate at room temperature for 5 min.

2. Place on magnetic stand for 5 min, remove 90 µl supernatant, keep on the magnetic stand and wash with 175 µl fresh 80% ethanol, remove ethanol and wash a second time, remove ethanol and air-dry on the magnetic stand for 2 min.

3. Add 17 µl resuspension buffer to each well, and pipette until the beads are re-suspended.

4. Incubate at room temperature for 2 min, place on the magnetic stand and incubate for 2 min, transfer 15 µl supernatant to the next step.

2.4: Quantifying libraries and correct colour balance

Once both the ATAC-seq and RNA-seq libraries had been prepared and amplified, their concentration was determined by qPCR, using the KAPA Library Quantification Kit (KK4824). The following protocol was used:

i) **Library Dilution:** Libraries diluted by serial dilution to 1:10, 1:100, 1:2000 and 1:4000, using 10 mM Tris-HCl

ii) **PCR Amplification:** 1:2000 and 1:4000 diluted libraries amplified by qPCR with the following reaction mix:
   - 12 µl of pre-mixed KAPA SYBR FAST + Primer Premix
   - 4 µl of DNA Standard/library dilution
   - 4 µl PCR-grade water

iii) **qPCR Cycling:**
   - 95°C, 5 min
   - 35 cycles 95°C, 30 s
   - data acquisition at 60°C, 45 s

iv) **Data Analysis:** Absolute quantification against standard curve 20 pM – 0.2 pM, an average ATAC-seq fragment length of 180bp is assumed.

Libraries were then diluted to the concentration of the individual DNA library whose concentration was lowest. This is to ensure that all libraries are of the same concentration when pooled together, helping ensure each library is sequenced to the same depth on the sequencing machine. Table 2.1 below shows the barcodes used for each sample for pilot experiments 1 and 2 and displays a good colour balance across the index sequences.

Experiment	Sample name	Barcode I.D.	Index Barcode Sequence
Pilot 1	Sample 1	50	**TCCTGAGC**
Pilot 1	Sample 2	30	**CGTACTAG**
Pilot 2	CT8 SHAM	70	**TAGGCATG**
Pilot 2	CT8 LP30	60	**GGACTCCT**
Pilot 2	CT16 SHAM	20	**TAAGGCGA**
Pilot 2	CT16 LP30	40	**AGGCAGAA**

Table 2.1: The barcode I.D. sequences used for the pilot experiments. The barcode sequences have been coloured according to whether they luminesce either red or green when sequenced, this is to access colour balance between the barcode sequences, these sequences show good colour balance.

2.5: Sequencing of DNA libraries

Following successful generation and quantification of ATAC-seq and RNA-seq DNA library libraries, these were sequenced using an Illumina NextSeq 500 sequencing system. Sequencing using an Illumina machine is carried out using a flow cell containing four lanes; DNA fragments will pass down each lane, and as the barcoded DNA fragments flow through the flow cell, they are bound by the Illumina adapter sequence, which for ATAC-seq was added by the Tn5 enzyme, when it cut out open areas of DNA, and for RNA-seq was ligated in when the anchors were added. These bound fragments then form clusters on the flow cell, these clusters of bound fragments are then amplified from the adapter sequence, using fluorescently tagged nucleotides which will fluoresce as they are added. As the nucleotides fluoresce, the emission wavelength and intensity determines how each base is called, hundreds of millions of clusters will be sequenced simultaneously in parallel. Once the fragment has been sequenced, it is washed off, the barcode sequences that were added during amplification are then amplified in the same manner, and these sequences are used to identify which sample the fragment came from. Figure 2.1 shows the process:

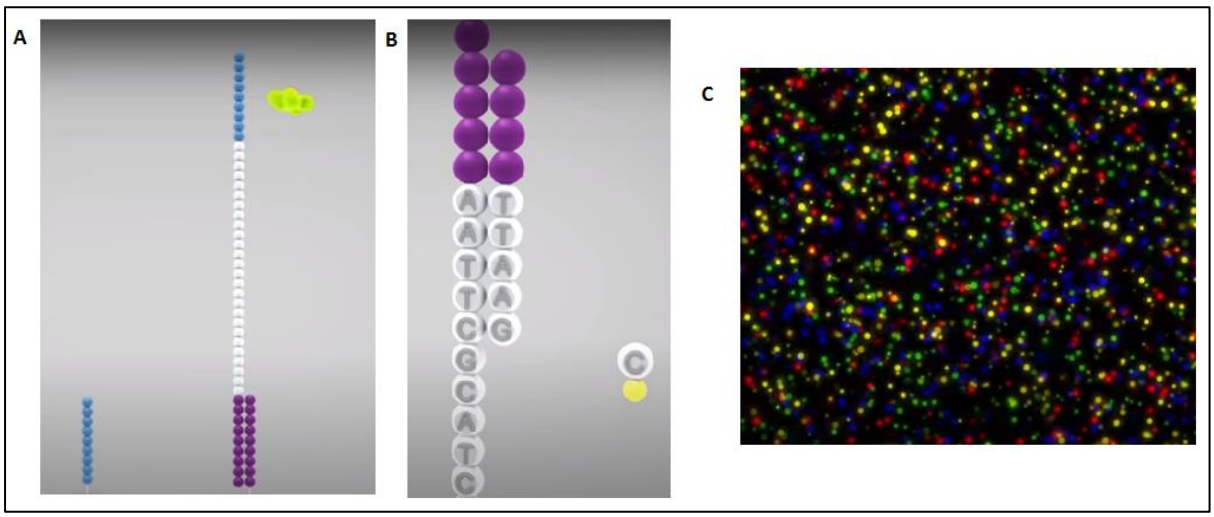


Figure 2.1: Illumina sequencing. A: Fragments first bind to the flow cell; **B:** Fragments are then amplified using fluorescently tagged nucleotides; **C:** As the nucleotides fluoresce, their base is called. Adapted from the Illumina website: (https://emea.illumina.com/science/technology/next-generation-sequencing/sequencing-technology.html).

The following Illumina protocol was used to sequence the libraries:

i) Normalization of library concentration:

1. All libraries were normalised and diluted to 0.5nM, 1nM, 2nM, or 4nM, depending on the lowest library concentration being sequenced.

ii) Denature libraries:

1. Libraries were denatured with 0.2M NaOH according the table below:

Starting Library Concentration	Library	0.2M NaOH
4nM	5µl	5µl
2nM	10µl	10µl
1nM	20µl	20µl
0.5nM	40µl	40µl

2. Vortex briefly and then centrifuge at 280 × g for 1 minute.

3. Incubate at room temperature for 5 min.

4. The following volumes of 200 mM Tris-HCl, pH 7 were added:

Starting Library Concentration	200 mM Tris-HCl, pH 7
4nM	5µl
2nM	10µl
1nM	20µl
0.5nM	40µl

5. Vortex briefly and then centrifuge at 280 × g for 1 minute.

iii) Dilute Denatured Libraries to 20 pM:

1. The following volumes of pre-chilled Hybridization Buffer (HT1) were added to the tube of denatured libraries.

Starting Library Concentration	Prechilled HT1
4nM	985 µl
2nM	970 µl
1nM	940 µl
0.5nM	880 µl

2. Vortex briefly and then centrifuge at 280 × g for 1 minute.
3. Place the 20 pM libraries on ice until you are ready to proceed to final dilution.

iv) Dilute Libraries to Loading Concentration of 1.8pM for High Output kit:

1. Add 117µl 20pM denatured library solution to 1183 µl pre-chilled HT1.
2. Invert to mix and then pulse centrifuge.

v) Sequence libraries using Illumina NextSeq with 70 cycle High Output sequencing cartridge using the following settings:

1. Indexing: Paired
2. Number of cycles: 38

The full ATAC-seq and RNA-seq library preparation methods are shown in Figures 2.2 and 2.3.

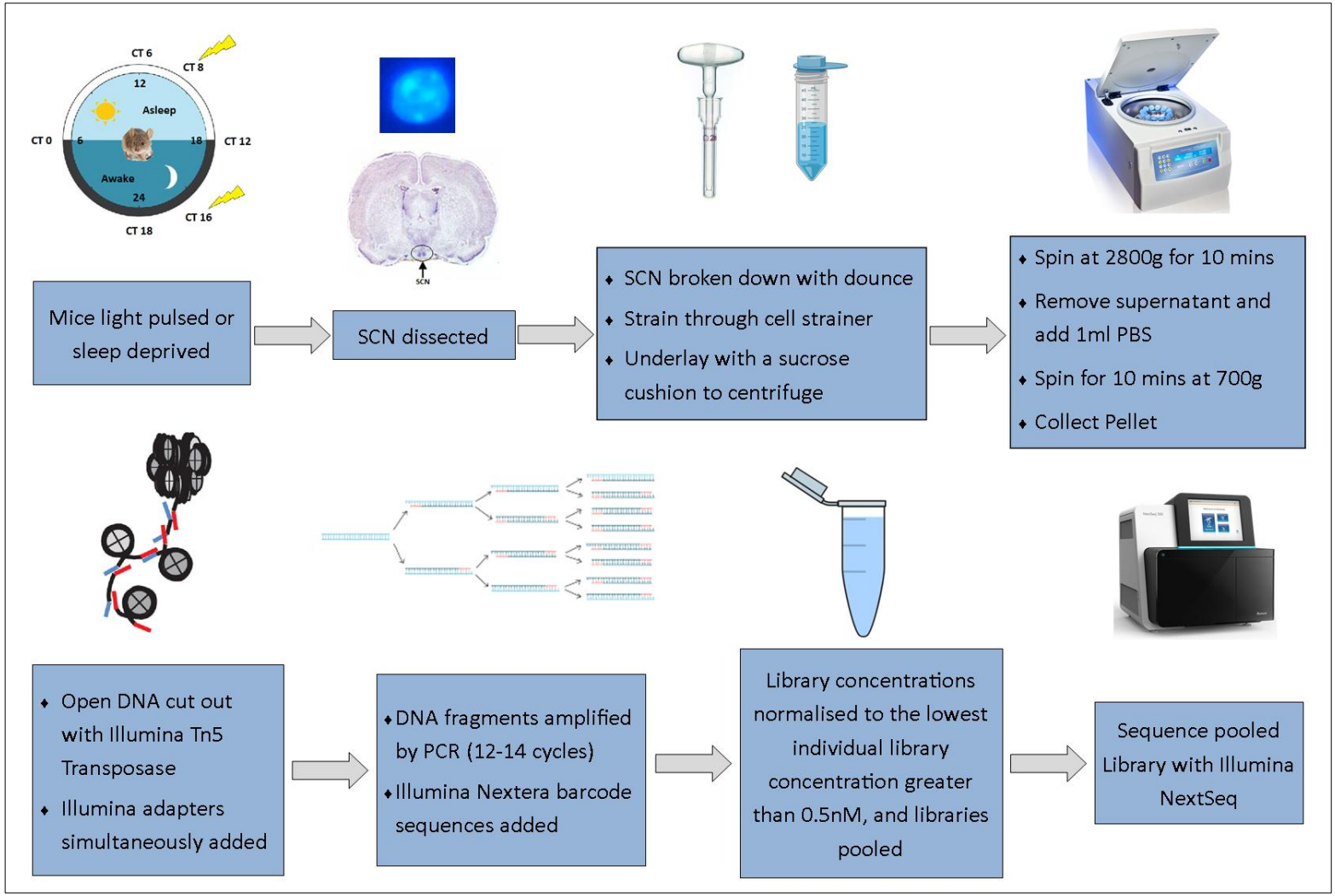


Figure: 2.2: ATAC-seq library preparation workflow.

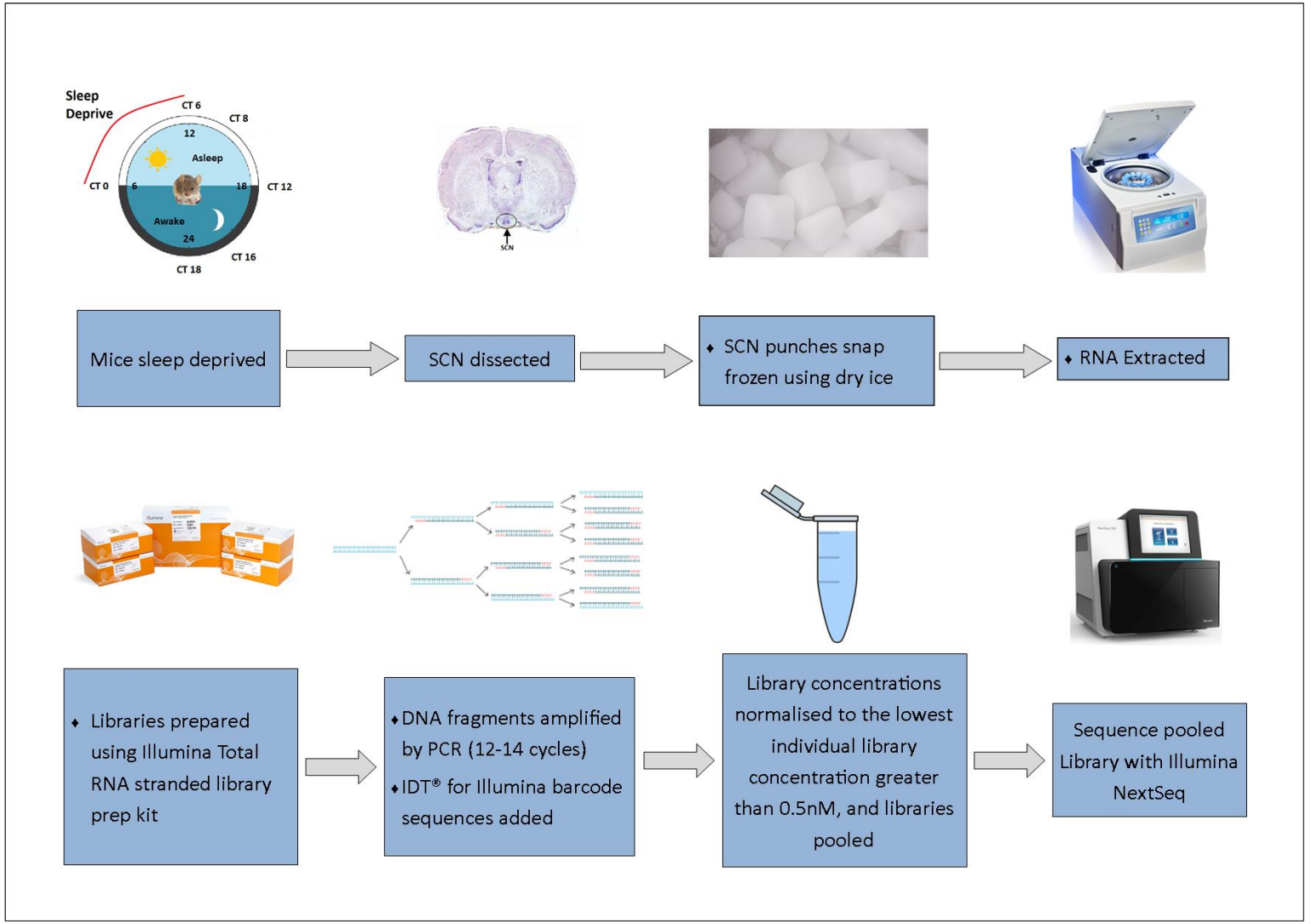


Figure: 2.3: RNA-seq library preparation workflow.

2.6: In-vivo animal methods

Unless otherwise stated, C57BL/6 black 6 mice were used for all in-vivo experiments; animals were housed in light tight chambers (LTC), with light/dark cycles being strictly controlled, were fed a standard diet, and the 3Rs methodology of live animal research was followed at all times (Russell and Burch, 1959). Mice began artificial light entrainment at 4 weeks old, whenever possible were housed in cages of 6, and were subsequently culled at approximately 12 weeks old. All experiments were carried out in the dark; ensuring mice were culled in circadian time (CT), rather than Zeitgeber time (ZT), Figure 2.4 illustrates the difference between ZT and CT. Prior to all experiments, mice were subject to their current light dark cycle for a minimum of 4 weeks, to make certain all animals were fully entrained. Mice were culled as quickly as possible using cervical dislocation to prevent further chromatin changes due to stress.

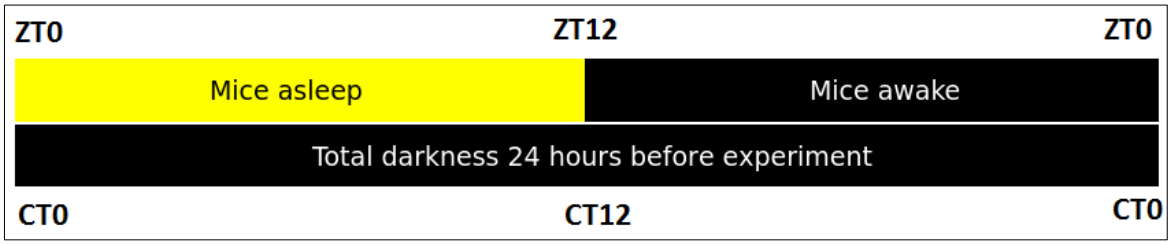


Figure 2.4: Zeitgeber time vs circadian time. Lights are turned off prior to all experiments guaranteeing mice oscillate by their internal molecular clocks, rather than by any external Zeitgeber such as light.

Light Pulse in-vivo Procedures

Mice were kept on a 12 hour light/dark cycle; the timing of this cycle was varied depending on the experimental plan to ensure mice were culled at the correct CT, Table 2.2 summarises the different light cycles that were used. All light pulsing was carried out at 400 lux for a period of either 30 or 60 mins, if multiple mice were culled, light timings were staggered to allow time for dissection. Mice were culled in groups according to their cage as this was unavoidable, but mice from the same experimental condition were not culled sequentially in an attempt to mitigate batch effects. Mice were light pulsed and culled at CT8 and CT16 (Figure 2.5), lights within LTCs were adjusted to ensure CT8 and CT16 occurred during working hours (Table 2.2)

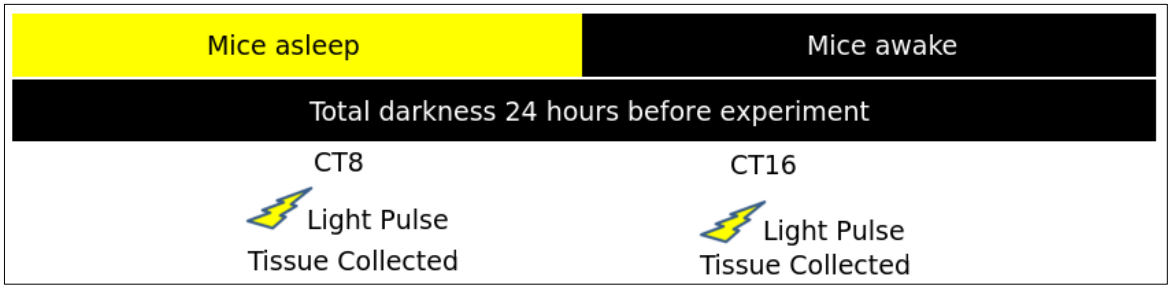


Figure 2.5: Light pulse timing. Mice were put into total darkness 24 hours before the experiment, to ensure mice were light pulsed in CT time at CT8 and CT16.

Experiment Time	lights on	lights off
CT8 at 10am	2am	2pm
CT16 at 10am	6pm	6am
CT8 at 2pm	6am	6pm
CT16 at 2pm	10pm	10am

Table 2.2: The different light/dark cycles used. This time used depended on experimental timing.

Sleep Deprivation in-vivo Procedures

Mice were kept on a 12 hour light/dark cycle, and were sleep deprived for a period of 6 hours (CT0-CT6) when they would usually be asleep (Figure 2.6). Mice were kept awake by stimulation using novel objects and gentle handling; all mice were kept awake until culled.



Figure 2.6: Tissue collection times for the SD. Mice were sleep deprived for 6 hours at the start of the light phase when they would normally be asleep.

2.7: Cultured cell protocol

Cells were incubated in a T25 Corning™ Cell Culture Flask (product no. 430825) with 12ml Dulbecco's Modified Eagle's Medium (DMEM; product no. BE12-604Q), containing 10% bovine calf serum (product no. 16170078) at 37°c/15% CO_2. Cells were split once confluent using the following protocol:

i) **Remove all culture medium**

ii) **Wash cells with 12ml PBS**

iii) **Lift cells using Trypsin**: Cells incubated for 5 min with 5ml Trypsin at 37°c.

iv) **Deactivate Trypsin and centrifuge:** 7ml DMEM added and cells spun at 500g for 5 min.

v) **Re-suspend and split cells:** Supernatant removed and cells re-suspended in 2ml DMEM; 200µl (10%) of cell suspension returned to flask containing 11.8ml DMEM.

Cells were used for transfection once confluent, remaining cells were split as outlined above and returned to the incubator.

2.8: Transfection of cultured cells

A transfection was used to artificially incorporate a number of nucleic sequences of interest into mouse NIH3T3 and human U2OS cells. By using lipid reagents, a chemical-mediated transfection was utilized to transfer plasmids containing the sequence of interest along with a luciferase reporter into the cells. When activated, the sequence of interest drives transcription of the luciferase gene, and the subsequence luminescence was used to measure the activity of the sequence of interest.

The sequences of interest were first incorporated into a plasmid using Invitrogen GeneArt (https://www.thermofisher.com/uk/en/home/life-science/cloning/gene-synthesis/geneart-gene-synthesis.html). The sequence of interest was then cloned from the GenArt plasmid into the Promega GL4.27 [luc2P/minP/Hygro] plasmid (https://www.promega.co.uk/) containing the luciferase gene (Cloning carried out by Norbert Vaga). This plasmid was then transfected into cultured cells to complex with lipid reagents, the Invitrogen Lipofectamine® 3000 reagent was used, this process enables negatively charged DNA molecules to pass through the negatively charged cell membrane.

The cells were incubated for 24 hours following transfection before being stimulated, and the subsequent luminescence recorded using a BMG plate reader.

The following protocol was followed for transfection:

i) **Cloning sequences of interest into plasmid using Invitrogen GeneART:** Sequences were determined from the Integrated Genome Browser (IGB), a restriction enzyme cutting site was added to either end of the sequence before submission to company for cloning.

ii) **Cloning into plasmid with luciferase gene:** Sequences of interest were cloned from GeneArt plasmid into the Promega GL4.27 [luc2P/minP/Hygro] plasmid. (Carried out by Norbert Vaga)

iii) **Plasmid DNA/lipid complexes:**

1. Per sample: 9.4µl Invitrogen Lipofectamine™ 3000 LP3000 Transfection Reagent (ThermoFisher: L3000015) diluted 1:10 with 0.6µl ibc Opti-MEM Reduced Serum Medium (ThermoFisher: 31985070).

2. Plasmid DNA concentration determined using NanoDrop 2000/2000c Spectrophotometers (ThermoFisher: ND-2000).

3. Plasmid DNA diluted with Opti-Mem and Invitrogen Lipofectamine™ 3000 P3000 transfection reagent (ThermoFisher: L3000015) solution to 20ng/µl.

4. 10µl of (1) mixed with 10µl (3) to give final plasmid DNA/lipid complex containing 200ng DNA.

iv) **Cultured cell counting:** Cells spun at 500g for 5 min and pellet re-suspended in 3ml of growth medium. 10 uL of cell suspension added to 50 uL Trypan blue (ThermoFisher: 15250061) and 40 uL of PSB, cells were counted using a haemocytometer.

v) **Preparing 96-well plate:** 10µl DNA/lipid complex and 20,000 cells in 90 µl of growth medium added to each well of the plate.

vi) **Cells and DNA plasmid/lipid complex left to incubate for 24 hours**

The following protocol was followed for treatment of cells and recording of bioluminescence:

i) **Mix medium for recording:** Per sample; 5µl B-27 Supplement (ThermoFisher: 17504044) and 0.1µl luciferin added to 94.9 µl recording medium.

ii) **Prepare cells for plate reader:** Growth medium removed from all cells and 100µl (1) added to each well

iii) **Cells placed in BMG plate reader:** The cells were left for 5 cycles to record a baseline level of luminescence

iv) **Treatment added to cells**: 10µl recording medium removed, and the same volume of relevant treatment added to each well.

v) **Cells returned to plate reader:**

2.9: Preliminary Local Bioinformatic Analysis

NGS techniques generate a large quantity of data, as a result a wide variety of bioinformatic techniques have been used to analyse the data. All analyses were carried out in-house, with many scripts being written de-novo to specifically analyse ATAC-seq data; many existing programs available to analyse NGS data are tailored for RNA-seq data, rather than ATAC-seq data. The following preliminary local analysis was carried out for pilot ATAC-seq data, code snippets have been included throughout the bioinformatic methods to facilitate an easier understanding of methods employed.

Preliminary QC of Data

Quality control was first carried out within the Illumina BaseSpace platform; this was primarily to assess the read number distribution across the different samples. Ideally if samples were correctly diluted correctly and tagmentation was successfully carried out, the number of reads within the sequencing run would be divided equally across the samples.

2.9.1: Initial Cleaning up of data and alignment prior to the initial analysis

Reads were initially trimmed using Trimmomatic (Bolger et al., 2014), to ensure that all adapter and barcode sequences were removed prior to alignment. Reads were then aligned to the mouse reference genome mm10, using Bowtie2 (Langmead et al., 2009). Following alignment, the aligned reads were then filtered using Samtools (Li et al., 2009) and Bamtools ((Barnett et al., 2011) to remove the following:

- **Duplicates:** Sequence duplicates occur during the amplification step, only unique reads are kept, these are reads starting and terminating at unique points within the genome. It is highly unlikely that two reads will both start and terminate at the same point, so any reads where this is the case, are considered amplification duplicates.

- **Mitochondrial aligned reads:** Any reads that align to the mitochondrial genome will give no useful information, and may affect later analysis steps, such as differential analysis.

- **Reads considered of low quality:** Reads of low quality may be sequenced incorrectly, and could therefore give inaccurate information.

- **Unpaired reads:** These are reads whose paired mate is not identified.

- **Reads aligning to ATAC-seq black regions:** There are areas of the genome which are always very accessible, and as a result will always give a very large number of reads, these regions are called ‘black regions’ (Amemiya et al., 2019). Any reads from these regions are removed, as they will likely skew later analysis steps, in particular peak calling.

2.9.2: Read coverage analysis and visualisation of data in genome browsers

The coverage of reads across the genome was determined using the package bamCoverage from deepTools (Ramírez et al., 2016). This determines where within the genome, the greatest number of reads are found, and therefore the areas of the genome that would be considered the most open. The program will bin the sequencing reads to a user defined bin width, usually of around 100bp, at this point normalisation is applied to the data in an attempt to account for differences in sequencing depth amongst the different samples. For this experiment the bin width was set to 100bp, and the Bins Per Million (BPM) normalisation method was applied, which carried out the following normalization calculation:

$$BPM\ (per\ bin) = \frac{number\ of\ reads\ per\ bin}{sum\ of\ all\ reads\ per\ bin\ (in\ millions)}$$

The output from bamCoverage is a .bigwig (.bw) file; this can be viewed using many genome browsers, and gives the tracks that have become characteristic for the analysis of both ATAC-seq and RNA-seq data, amongst many other sequencing methods. Throughout, all sequencing tacks within a single figure are scaled the same, unless otherwise stated.

Although this is really only the beginning of the analysis process; the tracks can be used to confirm if the sequencing method has worked correctly, and allows for a quick glance at how open any candidate genes may be. However, this method does not actually allow for any statistical testing to be carried out, and scrolling through the genome can be slow and cumbersome. The aligned read coverage tracks were viewed locally using The Integrated Genome Browser (IGB) (Freese et al., 2016), and online using the WashU browser (Zhou et al., 2011).

2.9.3: Running the programs locally

Initially, before access had been gained to the CGAT computer cluster, analysis to determine the success of pilot experiments was carried out locally using a laptop computer. Due to the large amounts of data being analysed, this process had to be both automated to allow the laptop to be left running for continuous periods of time, and also had to be tailored to the specifications of the laptop being used. To achieve this, a Linux shell script was written that would automatically start the various programs needed to complete the analysis, carrying out several tasks in tandem, making maximum use of computing resources available without over-taxing and crashing the laptop.

The Illumina NextSeq machine will output sequencing files in .fastq format, Figure 2.7 below shows a typical .fastq file.

```
@SEQ_ID
GATTTGGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTTGTTCAACTCACAGTTT
+
!"*((((***+))%%%++)(%%%%).1***-+*"))**55CCF>>>>>>CCCCCCC65
```

Figure 2.7: A typical .fastq file. The first line is the sequence identifier, the second line gives the sequence itself, the third line is the strand, and the last line gives data in relation to the quality score for each based that has been sequenced.

Eight .fastq files are output for each sample, these consist of a file for reads sequenced from each one of four lanes of an Illumina NextSeq DNA sequencer in the forward and reverse direction. The files at this point are either merged into two larger files, one for reads in the forward direction, and one for reads in the reverse. Or are analysed individually; as files were being analysed locally, due to laptop resources, the files were not merged for this preliminary analysis.

2.9.4: Linux shell Script

This script first takes the input files; it will take the forward and reverse .fastq file for each lane from the output of the Illumina NextSeq sequencer, and save them as a variable, the code is shown in Figure 2.8 below:

```
echo -n "Enter the path where the compressed .fastq files are located>"
read path

echo -n "Enter lane 1 forward alignment file name>"
read forward1
echo -n "Enter lane 1 reverse alignment file name>"
read reverse1
echo -n "Enter lane 2 forward alignment file name>"
read forward2
echo -n "Enter lane 2 reverse alignment file name>"
read reverse2
echo -n "Enter lane 3 forward alignment file name>"
read forward3
echo -n "Enter lane 3 reverse alignment file name>"
read reverse3
echo -n "Enter lane 4 forward alignment file name>"
read forward4
echo -n "Enter lane 4 reverse alignment file name>"
read reverse4
echo -n "Enter file path for bowtie2 genome >"
read genome
echo -n "Enter output file name >"
read output
workingdir=$(pwd)
```

Figure 2.8: Code for Linux shell script. This section reads in the input files, and assigns them to different variables.

Following the input of data, the first lane is aligned with Bowtie2 against mm10, the output of this first alignment is then filtered using Bamtools and Samtools, whist the alignment of the second lane begins. The alignments are run initially on 6 processor threads and then only 5 whilst the filtering steps use 1. This leaves 2 remaining processor threads so the laptop can still be used. The script gives updates to the user using echo commands, and pauses with the 'wait' command to allow all this code to finish, ensuring the laptop does not get overloaded, the code is shown below in Figure 2.9.

```
bowtie2 -p 6 -x $genome -1 $path/$forward1 -2 $path/$reverse1 | samtools view -bS - > alignment_temp.bam

echo "Alignment 1 completed, will now filter alignment 1 and start alignment 2....."
bowtie2 -p 5 -x $genome -1 $path/$forward2 -2 $path/$reverse2 | samtools view -bS - > alignment2_temp.bam &

bamtools index -in alignment_temp.bam
bamtools sort -in alignment_temp.bam -out sorted_temp.bam
bamtools index -in sorted_temp.bam
mkdir ATAC_raw_alignment
mv alignment_temp.bam $workingdir/ATAC_raw_alignment
samtools view -b sorted_temp.bam chr1 chr2 chr3 chr4 chr5 chr6 chr7 chr8 chr9 chr10 chr11 chr12 chr13 chr14 chr15
chr16 chr17 chr18 chr19 chX chY> filtered_temp.bam
rm sorted_temp.bam
echo "Mitochondrial DNA alignment 1 filtered"
bamtools filter -isDuplicate false -in filtered_temp.bam -out filtered2_temp.bam
echo "Duplicates Mitochondrial DNA alignment 1 filtered"
bamtools filter -isMapped true -in filtered2_temp.bam -out filtered_temp.bam
echo "non mapped alignment 1 filtered"
bamtools filter -isPaired true -in filtered_temp.bam -out filtered2_temp.bam
echo "non paired alignment 1 filtered"
bamtools filter -isPrimaryAlignment true -in filtered2_temp.bam -out filteredcomplete1.bam

rm filtered2_temp.bam
rm filtered_temp.bam
wait
echo "Alignment 2 Complete and alignment 1 has been filtered, will now filter alignment 2 and start alignment 3"
```

Figure 2.9: Further code for Linux shell script. This section carries out the Bowtie2 alignment, and filters the data output.

Once the first two alignments are completed, and the first one filtered, the script continues in the same fashion, aligning one lane whilst filtering the output from the previous alignment. Once all four lanes have been aligned, they are merged, and sorted, as shown in Figure 2.10 below:

```
echo "Alignment 4 complete, will now Merge bam files...."

bamtools merge -in filteredcomplete1.bam -in filteredcomplete2.bam -in filteredcomplete3.bam -in
filteredcomplete4.bam -out $output
bamtools index -in $output
rm filteredcomplete1.bam
rm filteredcomplete2.bam
rm filteredcomplete3.bam
rm filteredcomplete4.bam
```

Figure 2.10: Further code for Linux shell script. This section merges all the filtered .bam files from the different lanes, and then removes any remaining temporary files.

All temporary .bam files are deleted, leaving just a merged, indexed, and sorted .bam file for this sample.

2.10: In-depth Bioinformatic Analysis for deeper analysis of data

Bioinformatic analysis carried out for the second pilot experiment was sufficient to determine a viable method to extract nuclei and successfully generate ATAC-seq libraries had been achieved. However, a far more in-depth analysis was required to analyse the subsequent data generated from both larger ATAC-seq and RNA-seq experiments. Figure 2.11 gives the full workflow for the more advanced bioinformatic analysis carried out for ATAC-seq and Figure 2.12 for RNA-seq.

Running programs on the Computational Genomics: Analysis and Training (CGAT) cluster

Following completion of the first large ATAC-seq experiment, due to the amount of data, it was no longer viable to carry out the analysis using a laptop computer; instead the analysis was carried out using the CGAT computer cluster. The CGAT cluster has many bioinformatic pipelines already installed for the analysis of NGS data, including ATAC-seq. This allowed the initial analysis of the data to be quickly carried out, without the necessity of writing further Linux shell scripts such as the one written to analyse data from the pilot experiment.

The CGAT pipelines are all written in Python using a library called Ruffus (Goodstadt, 2010), this library has been specifically designed to quickly permit pipelines to be written, stringing together many different programs, exactly what is required to analyse ATAC-seq and RNA-seq data. The Ruffus library enables a user to write code that will automatically read in all files of a specific file type from a directory, for example all .fastq sequencing files can be read in at the beginning of the analysis. The pipeline will then output to the same, or a different directory. This is particularly useful when analysing large amounts of data on the cluster, as the cluster can be left running over a prolonged period of time, using Ruffus pipelines to automatically read in any input files, and output to whichever directory is appropriate.

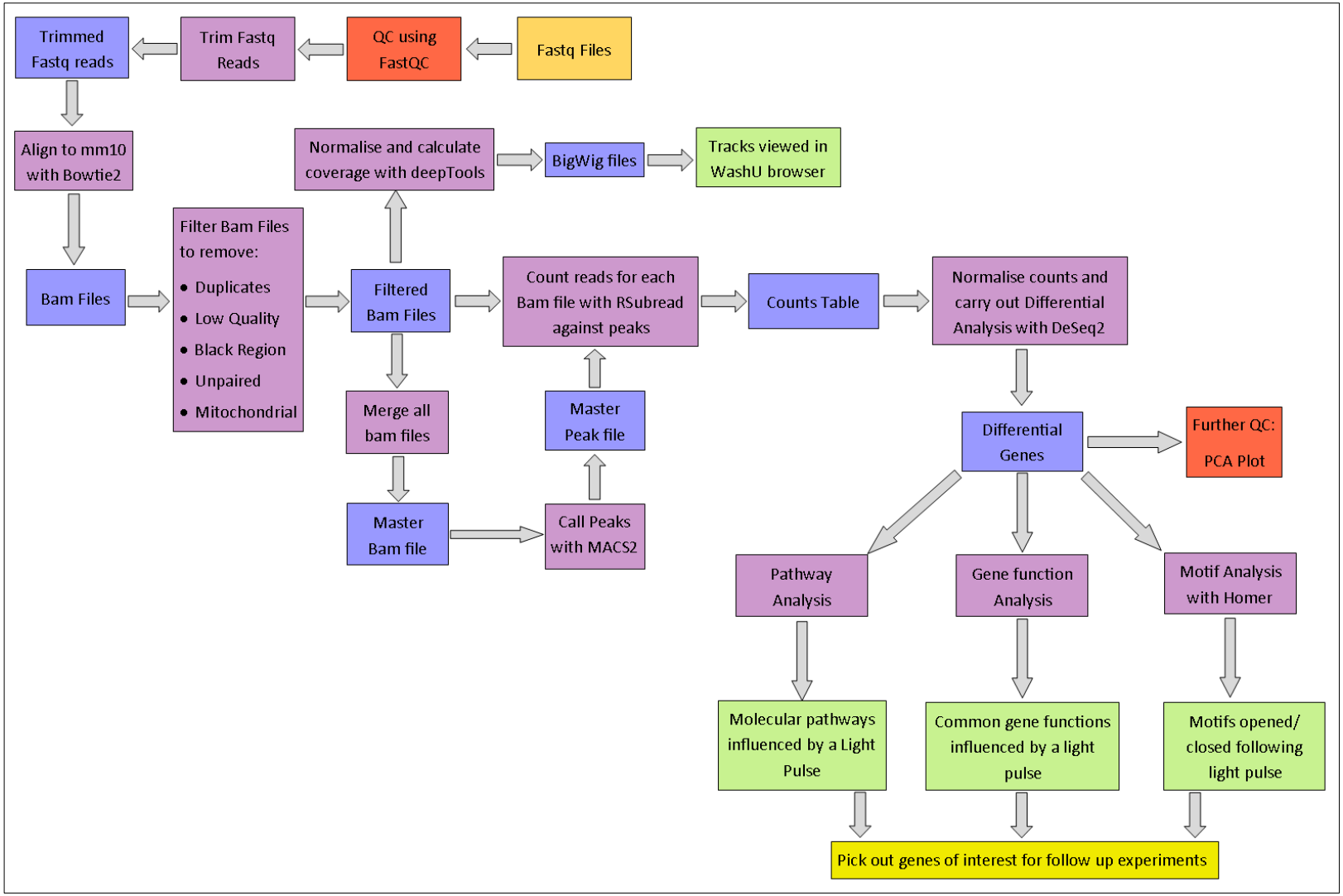


Figure 2.11: The full bioinformatic workflow for ATAC-seq.

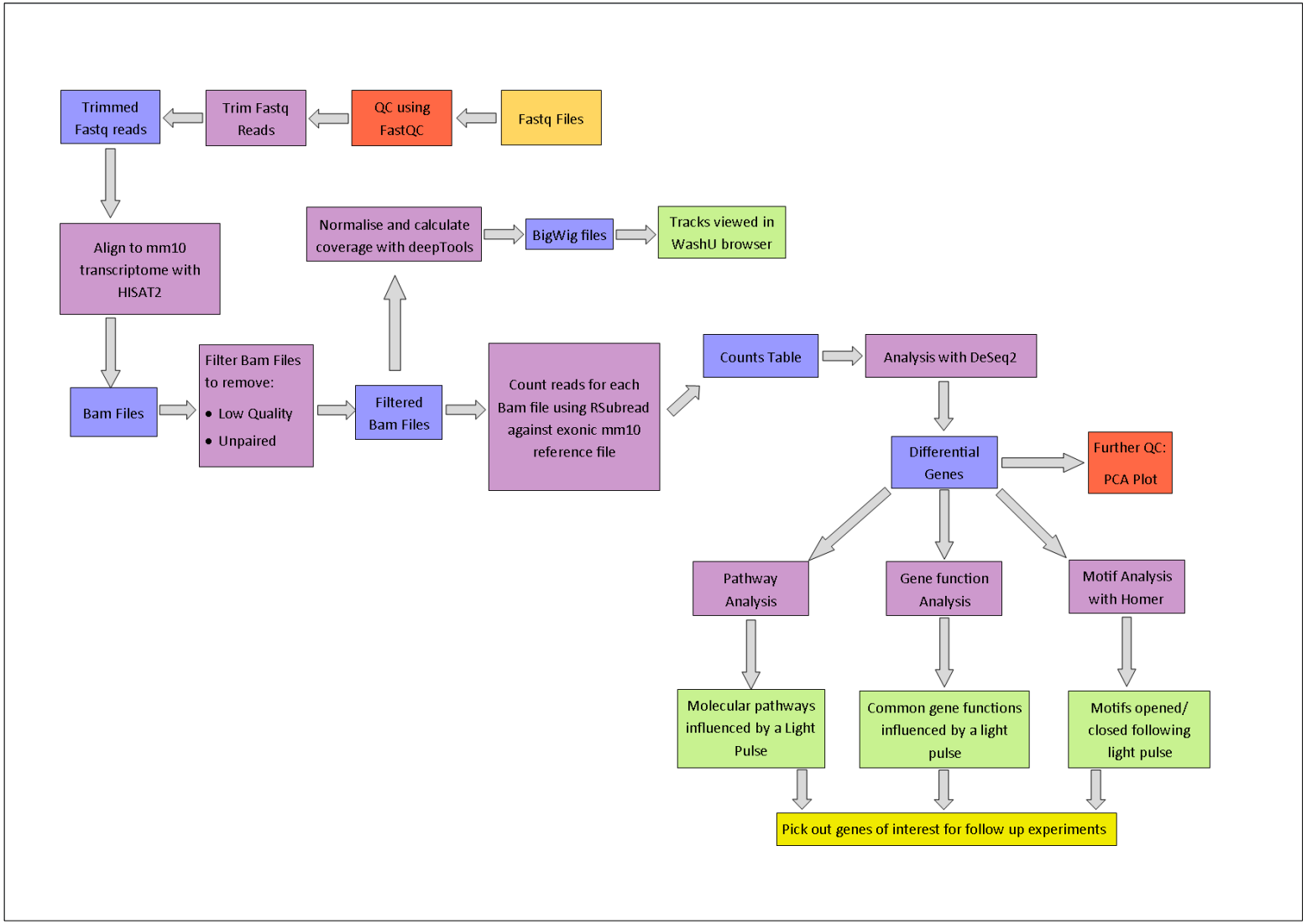


Figure 2.12: The full bioinformatic workflow for RNA-seq.

2.10.1: Initial Quality control of Data

Data for both ATAC-seq and RNA-seq was checked using the Illumina BaseSpace platform, to determine if read distribution was evenly spread across the samples. Following this initial QC, further QC was now carried out using FastQC (Andrews), this allowed for a far more in-depth check of the data, and was applicable for both ATAC-seq and RNA-seq, the following benchmarks were used:

- **Per base sequence quality:** This asses the base quality at each point in the sequencing reads, so this gives a mean, median, and interquartile range for each point, reads were typically 38bp long for the ATAC-seq data.
- **Per Sequence Quality Scores:** This gives the quality score across all sequences; it allows the user to see if subsets of sequences have low quality, giving an indication if there were any problems with sequencing.
- **Per Base Sequence Content:** Plots out the mean proportion of each base, at each position, it is expected in a high quality library that there is an even distribution of each base.
- **Per Sequence GC Content:** This gives the overall distribution of GC content; this is usually expected to be normally distributed.
- **Per Base N Content:** The overall proportion of 'Ns' of across sequence reads, too many Ns indicated there may have been problems with sequencing.
- **Sequence Length Distribution:** In this case, most sequence reads would be expected to be 38bp, because a 75bp cartridge was used in paired end mode.
- **Duplicate Sequences:** This is very important for ATAC data, a large proportion of duplicated sequences indicates that the library has been over amplified, and usually means the library is of low depth.
- **Overrepresented Sequences:** A good quality library will contain a diverse set of sequences, if one or a few sequences are significantly overrepresented, this either means that something very biologically significant has been found, or unfortunately more usually indicates the library has been contaminated.

- **Adapter Content:** This is useful to determine if trimming of sequences is required for the library.
- ***K*-mer Content:** This is a further check, to look for over represented sequences, *k*-mer sequences are sub-sequences of length *k* within a sequence. Partial sequences, and sequences with many errors may be missed by the over-represented sequences check, this additional check will help to resolve this problem.

  For example the *K*mer sequences for 'ATCGATG' are:

K	*K*-mers
1	A, T, C, G, A, T, G
2	AT, TC, CG, GA, AT, TG
3	ATC, TCG, CGA, GAT,ATG
4	ATCG, TCGA, CGAT, GATG
5	ATCGA, TCGAT, CGATG

Once QC had been completed with FastQC, all the data was collated together using MultiQC (Ewels et al., 2016), this program will output figures for each benchmark containing data across all libraries. This allows for a quick and effective comparison of libraries helping to determine if any libraries are of particularly poor quality compared to the others.

Further quality control was carried out specifically for ATAC-seq data, and was not applicable for RNA-seq. A CGAT Ruffus pipeline was used to carry out these checks; the following benchmarks were checked:

- **Fragment length distribution:** This is distinct from sequence length distribution that was checked by FastQC, this looks at the actual fragment lengths cut by the Tn5 enzyme, whereas the sequence length distribution is simply the length of the reads direct from the sequencer. For a good quality ATAC-seq library, a very specific pattern of fragment lengths is expected. This is due to how the Tn5 transposase enzyme will cut open DNA between nucleosomes; there will be a nucleosome free peak of around 80bp, and then a peak approximately every 200bp, because DNA wrapped around a nucleosome is approximately 200bp in length. Figure 2.13 shows the fragment pattern for a good quality ATAC-seq library.

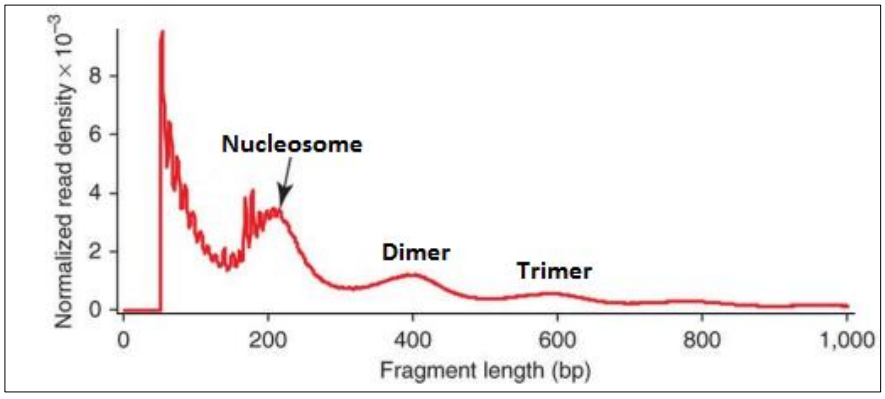


Figure 2.13: **Typical ATAC-seq fragment size distribution.** Peaks can be seen for nucleosome free fragments, then for single, di, and tri nucleosome fragments.

- **Percentage of mitochondrial reads:** Mitochondrial DNA contamination is common for ATAC-seq libraries, although it can simply be filtered out, a high number of mitochondrial reads indicates the library will be of low depth, and the library preparation may need to be further refined.
- **Ratio of mono, di and tri nucleotide reads:** This check helps determine if the transposase enzyme is cutting the DNA as expected, ideally in a good library, the ratio of these should mirror what is shown in Figure 2.61.

2.10.2: Alignment, filtering and visualisation of coverage tracks

For ATAC-seq alignment and filtering were carried out exactly as they were for the pilot experiment, except this time the alignment and filtering programs were run on the cluster using the mapping CGAT Ruffus pipeline (CGAT, Mapping). For RNA-seq HISAT2 was used instead of Bowtie2 for alignment, RNA-seq reads had to be aligned to the mm10 transcriptome rather than the mm10 genome, this is to take account for post translational modification such as splicing. HISAT was run with default settings, for stranded kits strandedness was set to most appropriate, to maximise alignments. Only non-paired and low quality reads were filtered out for RNA-seq, duplicated reads could not be removed because of the possibly of multiple copies already occurring with the transcriptome. For both methods filtered .bam files were produced, and as for the pilot, the read coverage of the .bam files was calculated using deepTools, and .bw files were output. However, this time a small Ruffus pipeline was written to automate the process of generating the .bw files as many more .bam files needed to be read into deepTools. The code is shown below in Figure 2.14:

```
import sys
import os
from ruffus import *
from cgatcore import pipeline as P

# load options from the config file
PARAMS = P.get_parameters(
    ["%s/pipeline.yml" % os.path.splitext(__file__)[0],
     "../pipeline.yml",
     "pipeline.yml"])

# Specific pipeline tasks
@transform('*.bam',
           suffix(".bam"),
           ".bw")
def bam_to_bw(infile, outfile):
    '''Transform bam to bw'''

    cwd = os.getcwd()
    # the command line statement we want to execute

    statement = '''bamCoverage -p 6 -bs 10 --normalizeUsing BPM -b %(infile)s -o %(outfile)s''' % locals()

    P.run(statement)

  # ---------------------------------------------------

def main(argv=None):
    if argv is None:
        argv = sys.argv
    P.main(argv)

if __name__ == "__main__":
    sys.exit(P.main(sys.argv))

# ---------------------------------------------------
```

CGAT and Ruffus plugins imported

Input and output file types

deepTools command

Figure 2.14: Python code to automate converting .bam files .bw files. The code uses the Ruffus python plugin, as well as a CGAT plugin, so that it can be run on the CGAT computer cluster, it takes in all .bam files in a folder and converts them to .bw using bamCoverage from the DeepTools package.

The .bw files were visualised using the WashU browser for both datasets, but this time a track hub was written so the files could be accessed remotely, this method allowed the coverage tracks to be shared far more easily with my supervisors and collaborators. The code for the track hub is shown below in Figure 2.15:



Figure 2.15: Code for the track hub for the pilot experiment. Written in .json file format, this code allows external access to the read coverage tracks stored on the CGAT computer cluster, websites such as the WashU browser can use this code to display the read coverage tracks.

2.10.3: Peak Calling

Following read alignment, for ATAC-seq peaks were called on the .bam files, this process looks for where the peaks in the number of reads are located within the genome, these peaks represent areas where the genome is more open. Peaks were called using MACS2 (Zhang et al., 2008), a CGAT Ruffus pipeline was used to carry out peak calling (CGAT, Peak Calling). The results of the peak calling were visualised in the IGB along with the coverage tracks, to check that the peaks had been called in the correct place, and to determine which parameter setting performed best. Testing indicated all parameter settings gave very similar results, therefore the default MACS2 parameters were used for all analysis to ensure there was no bias between experiments.

Following Peak calling a master peak file was generated; this file is needed for read counting and contains the location of any possible peak across all the different samples. The master peak file was generated by first merging all the filtered .bam files using Samtools, and then carrying out peak calling on the single merged .bam file to create a file with sufficient coverage to increase peakcalling accuracy. The numbers of reads are subsequently counted at each peak position; this is required for differential analysis to determine which samples have the greater number of reads at each peak. Again the merged peak file was visualised in the IGB alongside the coverage tracks, to check the master peak file had been generated correctly.

2.10.4: Read Counting

For ATAC-seq the number of reads contained within the .bam files located at each peak location within the genome were counted, an R (Team, 2017) plugin called Rsubread (Liao et al., 2013) was used to carry out this analysis. For RNA-seq reads were counted against the mm10 annotation file, with 'exon' being used as the feature type instead of 'peak'. An R markdown script was first written deNovo, to carry out this analysis using the Rsubread plugin, the code is shown below in Figure 2.16:

```
library(stringr)
arg1 <- commandArgs(trailingOnly=T)[1]
bamfile <- strsplit(arg1,',')[[1]]

samplename <- bamfile.labels <- gsub(".bam", "", basename(bamfile))

samplename <- str_split_fixed(samplename, "trimmed-", 2)
samplename <- str_split_fixed(samplename[,2], ".b", 2)
samplename <- samplename[,1]

gffFile1 <- commandArgs(trailingOnly=T)[2]
gffType <- strsplit(gffFile1,".")
gffType <- gffType[length(gffType)-1]

if (length(bamfile) != length(samplename)) {
  print("Number of sample names does not equal number of bam files")
  #knit_exit("Number of sample names does not equal number of bam files, the program may crash!!!")
}
library(Rsubread)
stats_table <- list()
counts_table <-list()
for (i in 1:length(bamfile)){
  counts1 <- featureCounts(files=bamfile[i], annot.ext=gffFile1, isGTFAnnotationFile=TRUE,
  nthreads=10,isPairedEnd = TRUE,GTF.featureType="PEAK",GTF.attrType = "peakid")
  counts_table[[i]] <- counts1$counts
  stats_table[[i]] <- counts1$stat
}

peak_table <- data.frame(Peaks=row.names(counts_table[[1]]))
for (i in 1:length(bamfile)){
  peak_table <- cbind(peak_table,Counts=as.numeric(counts_table[[i]]))
  names(stats_table)[i] <- samplename[i]
}
for (i in 1:length(bamfile)){
  colnames(peak_table)[i+1] <- samplename[i]
}
workingdir <- getwd()
peakcsv <- "Steve_NP_peaks_Cortex.csv"
statscsv <- "Steve_NP_Cortex_subread-stats.csv"
peakcsv <- paste(workingdir,peakcsv, sep = "")
statscsv <- paste(workingdir,statscsv, sep = "")
write.table(peak_table,peakcsv,sep=",",row.names = FALSE)
write.table(stats_table,statscsv,sep=",",row.names = FALSE)
```

Figure 2.16: R code written to carry out read counting using the Rsubread R plugin. The code will take in the .bam alignment files, and the master peak file in .gft format, it will then output a counts table, showing how many reads are located at each peak location.

Unfortunately the .bed peak files output from MACS2 are not compatible with RSubread, so a python script was written to convert the .bed files into .gtf format; the code is shown below in Figure 2.7

```
#!/usr/bin/env python
# This code exists in 2 places: ~/datatypes/converters and ~/tools/filters
from __future__ import print_function
import sys
assert sys.version_info[:2] >= (2, 6)

def __main__():
    input_name = sys.argv[1]
    output_name = sys.argv[2]
    skipped_lines = 0
    first_skipped_line = 0
    out = open(output_name, 'w')
    out.write("##gff-version 2\n")
    out.write("##bed_to_gff_converter.py\n\n")
    i = 0
    for i, line in enumerate(open(input_name)):
        complete_bed = False
        line = line.rstrip('\r\n')
        if line and not line.startswith('#') and not line.startswith('track') and not line.startswith('browser'):
            try:
                elems = line.split('\t')
                if len(elems) == 12:
                    complete_bed = True
                chrom = elems[0]
                if complete_bed:
                    feature = "mRNA"
                else:
                    try:
                        feature = elems[3]
                    except Exception:
                        feature = 'feature%d' % (i + 1)
                start = int(elems[1]) + 1
                end = int(elems[2])
                try:
                    score = elems[4]
                except Exception:
                    score = '0'
                try:
                    strand = elems[5]
                except Exception:
                    strand = '+'
                try:
                    group = elems[3]
                except Exception:
                    group = 'group%d' % (i + 1)
                if complete_bed:
                    out.write('%s\t%d\tPEAK\t%d\t%d\t%s\t%s\t.\tpeakid "peak%d" %s\n' % (chrom, i, start, end, score, strand, feature, i))
                else:
                    out.write('%s\t%d\tPEAK\t%d\t%d\t%s\t%s\t.\tpeakid "%s"\n' % (chrom, i, start, end, score, strand, feature))
                if complete_bed:
                    # We have all the info necessary to annotate exons for genes and mRNAs
                    block_count = int(elems[9])
                    block_sizes = elems[10].split(',')
                    block_starts = elems[11].split(',')
                    for j in range(block_count):
                        exon_start = int(start) + int(block_starts[j])
                        exon_end = exon_start + int(block_sizes[j]) - 1
                        out.write('%s\tbed2gff\tPEAK\t%d\t%d\t%s\t%s\t.\texon %s;\n' % (chrom, exon_start, exon_end, score, strand, group)
            except Exception:
                skipped_lines += 1
                if not first_skipped_line:
                    first_skipped_line = i + 1
        else:
            skipped_lines += 1
            if not first_skipped_line:
                first_skipped_line = i + 1
    out.close()
    info_msg = "%i lines converted to GFF version 2.  " % (i + 1 - skipped_lines)
    if skipped_lines > 0:
        info_msg += "Skipped %d blank/comment/invalid lines starting with line #%d." % (skipped_lines, first_skipped_line)
    print(info_msg)
```

Figure 2.17: Code written to covert .bed peak files into .GTF peak files. Output from MACS2 is not compatible with RSubread, so required conversion for input into RSubread.

Following successful implementation of Rsubread, the process was again automated using a Ruffus pipeline, this code was written de-novo and is shown below in Figure 2.18:

```
import sys
import os
from ruffus import *
from cgatcore import pipeline as P

# load options from the config file
PARAMS = P.get_parameters(
    ["%s/pipeline.yml" % os.path.splitext(__file__)[0],
     "../pipeline.yml",
     "pipeline.yml"])

@merge(('*.bam',
       suffix(".bam")),
       "_counts.csv")
def atac_counting(infiles, outfile):
     '''Read counting with rsubread'''
     cwd = os.getcwd()
     peak_file =  PARAMS['peak_file_gtf']
     infiles2 = []

        for x in infiles:
         infiles2.append(x)

    length = len(infiles2)

        for x in range(length):
            infiles2[x] = cwd + "/" + str(infiles2[x])
            print (infiles2[x])

        infiles2. pop()
        infile_str = ','.join(infiles2)

     statement = '''R -e "rmarkdown::render('/ifs/projects/proj078/1_atac_seq/
     10_run2vs3/2_counting/rsubread_multiple_reanalysis.Rmd',output_file='%(cwd)s/rsubread.htm')"
     --args "%(infile_str)s"  "%(peak_file)s" "%(cwd)s/"''' % locals()

     P.run(statement,job_condaenv="proj078-R.1")
def main(argv=None):
     if argv is None:
         argv = sys.argv
     P.main(argv)

if __name__ == "__main__":
     sys.exit(P.main(sys.argv))
```

Figure 2.18: Code written to automate the read counting process. The pipeline takes in all .bam alignment files from the current folder, a .gtf peaks file, and carries out read counting to output a counts table against either the master peak file for ATAC-seq, or mm10 annotation file for RNA-seq.

A typical counts table is shown below in Figure 2.19:

Peak ID	Sample 1	Sample 2	Sample 3	Sample 4
1	1	1	2	8
2	5	12	6	2
3	0	1	1	1
4	11	19	13	1
5	18	13	15	2
6	2	6	6	0
7	2	3	3	0
8	2	1	1	0
9	0	2	0	4
10	0	0	0	0
11	3	1	1	0
12	4	9	8	0
13	4	7	9	5
14	12	15	6	0
15	0	4	3	0
16	2	4	2	2
17	1	0	0	3
18	0	1	5	0
19	0	1	5	1
20	5	4	3	0

Figure 2.19: A typical counts table for sequencing data. Peak 14 can be seen to have very different counts for samples 1 and 2, compared to samples 3 and 4, differential analysis would determine if this is a statistically significant difference.

2.10.5: Differential Analysis

Following generation of a counts table from the .bam alignment files and for ATAC-seq a master peak file, differential analysis was carried out using DESeq2 (Love et al., 2014). This process is the same for ATAC-seq or RNA-seq, and will associate values of statistical significance to any differences in the number of reads located at genes across all sample conditions. Therefore any genes that have been identified from the .bw tracks as being more open or closed for ATAC-seq or transcribed differently for RNA-seq can now be validated as to whether any difference observed is actually statistically significant.

To give greater control over the analysis process, an R-script was written de-novo to carry out the differential analysis using DESeq2, instead of using an integrated package such as DiffBind. The counts table is loaded into R; no normalisation was carried out on the data at this point, because this is a requirement of DESeq2, which implements its own in-build sample normalisation.

From the counts table the rows denote each gene '*i* 'and the columns each sample '*j'*, for every gene a generalized linear model (GLM) is fitted. The following is modified from the published DESeq2 paper (Love et al., 2014); the counts from the counts table are modelled as K_{ij} as following a negative binomial distribution with mean μ_{ij} and dispersion α_i. The mean is taken as a quantity q_{ij}, proportional to the concentration of cDNA fragments from the gene in the sample, and is scaled by a normalization factor s_{ij} , i.e., $\mu_{ij} = s_{ij}\, q_{ij}$. Therefore for many applications, the same constant s_j can be used for all genes in a sample; this can then be taken to account for any differences in sequencing depth between samples. GLMs with a logarithmic link:

$$\underset{2}{\log}\, q_{ij} = \sum_r x_{jr}\beta_{ir}$$

with design matrix elements x_{jr} and coefficients β_{ir} being used.

A design matrix is also uploaded into R, and will indicate if a sample *j* is treated or not, it is used as the foundation upon which to carry out the differential analysis. At its simplest, the different analysis will look to give a comparison between two groups; treated and control samples; the GLM fit will return coefficients to indicate the overall expression strength of the gene and the $\log_2$ fold change

between the treated and control samples. Within-group variability, i.e., the variability between replicates, is modelled by the dispersion parameter $\alpha\, i$, which describes the variance of counts via:

$$\mathrm{Var}K_{ij}=\mu_{ij} + \alpha i \mu^{2}_{ij}$$

To facilitate an easier analysis, RMarkdown code was written to break the code into chunks, this code is shown below in Figure 2.20:



Figure 2.20: RMarkdown for differential analysis with DESeq2. R libraries are loaded; files are input, the experimental set up is defined, and differential analysis carried out; code is broken down into smaller chunks to facilitate a simpler analysis.

DESeq2 will use all the different sample conditions to carry out the differential analysis. Once the differential analysis has completed, individual sample conditions can then be contrasted to see how the results differ. For example in the pilot experiment, sample CT16 LP may be contrasted with CT16 SHAM to see what effect the light pulse has on the experiment. Once all sample conditions have been contrasted, the data for each contrast is visualised using an MA plot, which is an application of a Bland–Altman plot for visual representation. When the data is contrasted it is transformed onto M (log ratio) and A (mean average) scales, which are then plotted. Figure 2.21 below shows a typical MA plot.

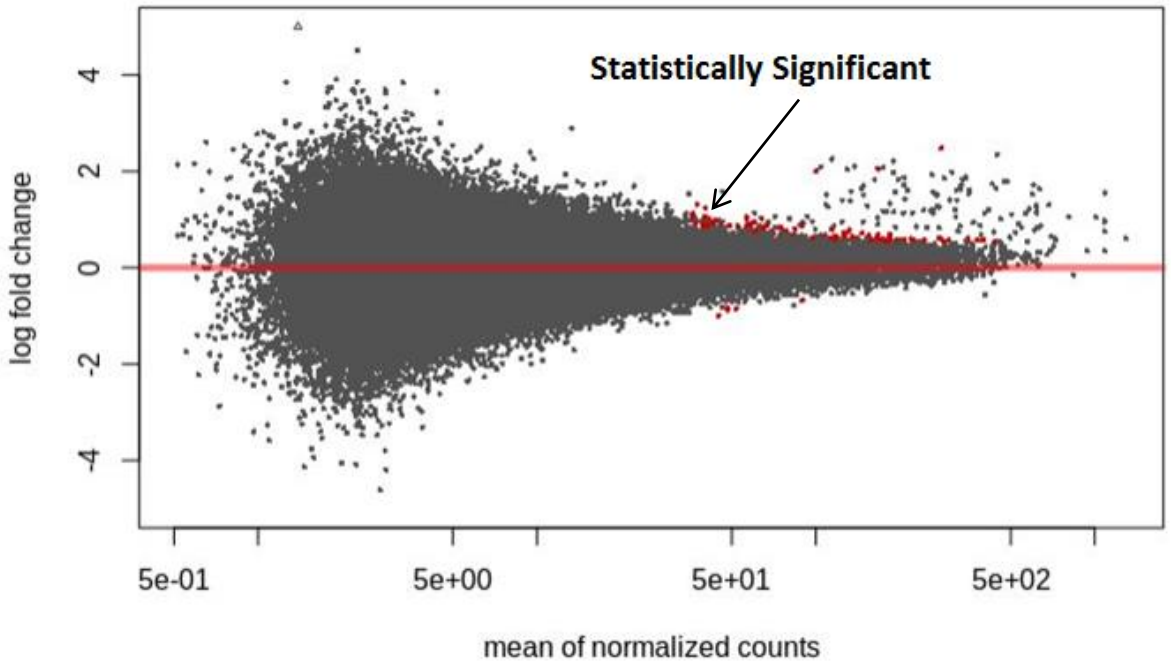


Figure 2.21: A typical Bland–Altman MA plot. Each dot represents a gene and shows the log fold change for the number of counts and therefore any change in 'openness' for an ATACseq experiment. Any dot that is coloured red is statistically significant.

The code to contrast conditions is shown below in Figure 2.22:

```
#All the different contrasts of conditions
```{r}
resCT16_LPvsSHAM <- results(analysis_multi, contrast=c("group", "CT16Yes", "CT16No"))
resCT8_LPvsSHAM <- results(analysis_multi, contrast=c("group", "CT8Yes", "CT8No"))
resCT8vsCT16_SHAM <- results(analysis_multi, contrast=c("group", "CT16No", "CT8No"))
resCT8vsCT16_LP <- results(analysis_multi, contrast=c("group", "CT16Yes", "CT8No"))
```

##Plot MA for data
```{r}
plotMA(resCT16_LPvsSHAM, ylim=c(-5,5))
plotMA(resCT8_LPvsSHAM, ylim=c(-5,5))
plotMA(resCT8vsCT16_SHAM, ylim=c(-5,5))
plotMA(resCT8vsCT16_LP, ylim=c(-5,5))
```

```{r}
getgenes_LP60CT16 <- resCT16_60vs0
ordered_LP60CT16 <- getgenes_LP60CT16[order(getgenes_LP60CT16$pvalue),]
getgenes_LP60CT16 <- subset(ordered_LP60CT16, padj < 0.1)
peaks_LP60CT16 <- as.data.frame(peak)
getgenes_LP60CT16$peak <- row.names(getgenes_LP60CT16)
getgenes_LP60CT16 <- as.data.frame(getgenes_LP60CT16)
colnames(peaks_LP60CT16)[colnames(peaks_LP60CT16)=="V4"] <- "peak"
thepeaks_CT16LP <- merge(x= peaks_LP60CT16[,c("seqnames","start","end","peak","strand")], +
y = getgenes_LP60CT16[,c("peak","pvalue","padj")], by = "peak")
thepeaks_CT16LP <- thepeaks_CT16LP[,c(2,3,4,1,5,6,7)]
thegenes_CT16LP <- merge(x= thepeaks_CT16LP[,c("peak","pvalue","padj")], +
y = anno_table_copy[,c("peak","ENSEMBL","geneId","SYMBOL")], by = "peak")
thegenes_CT16LP <- thegenes_CT16LP[,c(6,5,4,3,2)]
thegenes_CT16LP <- thegenes_CT16LP[order(thegenes_CT16LP$padj),]
```

```{r}
write.csv(thegenes_CT16LP,"ATACseq/6_differential/geneLists")
```
```

Figure 2.22: RMarkdown code to contrast the different experimental conditions. Following differential analysis, the different experimental parameters must be contrasted to determine which genes are differential ready for export.

To determine which gene have differential chromatin opening for ATAC-seq or differential expression for RNA-seq, gene lists are extracted for each contrast, the following result parameters are given by DESeq2:

- **baseMean:** The mean number of counts for the gene across all conditions
- **log2FoldChange:** The log fold change (LFC) that has occurred between the conditions contrasted
- **lfcSE:** The standard error value
- **stat:** Wald statistic, which is the LFC divided by its standard error
- **Wald p-value:** Indicates the statistical strength of any change that has been observed
- **padj:** This is the stricter adjusted p-value, corrected using Banjamini-Hochberg adjustment.

Data is output in .csv format with various adjusted p-value cut offs being used to generate lists of differential genes, for genomics a typical cut-off is 0.05. However, if a more or less strict adjusted p-value is preferred the cut-off can be adjusted accordingly, for example sometimes the adjusted p-value is relaxed to look for genes with a larger LFC. Any differential genes with low adjusted p-values can be visualised in the genome browser tracks, to determine if the genes are indeed differential or not.

2.10.6: Principal Component Analysis (PCA)

Once a counts table had been generated following read counting, for both ATAC-seq and RNA-seq PCA was carried out on the data as a further quality control; this was executed using the in-build R PCA plot function. A PCA is an excellent method to look for patterns from a large dataset, such as a counts table from ATAC-seq data, it is expected that samples from similar experimental conditions will cluster together on a PCA plot. A PCA summarizes the counts table by representing the multivariate data table as a smaller set of variable summary indices, this allows trends, clusters and outliers in the data to be quickly identified. If samples from similar experimental conditions do not cluster together, or there are a large number of outliers, the reliability of the data may be questioned at this point. Figure 2.23, show a typical PCA plot were the data has clustered appropriately.



Figure 2.23: A typical PCA plot. The data can be seen clustering into the different experimental conditions, with some overlap between conditions 2 and 3.

2.10.7: Batch Correction with RUVseq

RUVseq carries out normalization by employing factor analysis using a suitable set of control genes; here published SCN light response data from (Jagannath et al., 2013) and time of day RNA-seq SCN expression data from (Pembroke et al., 2015) were used to create three housekeeping gene lists. The first contained genes not induced or suppressed by light in the SCN, the second genes that do not fluctuate in their expression over the course of a day, and a third that was an amalgamation of the previous two.

Data is analyzed using the in depth bioinformatic methods outlined until read counting section 2.10.4, which generates a counts table; two approaches were carried out to determine the best method to normalize the data. The raw counts table from Rsubread was first used as input for RUVseq without any prior normalization, to test how effective the program works exclusively by itself. Following this a hybrid method was employed; data was normalized for read depth using DeSeq2 as described in section 2.10.5 to generate a new counts table normalized for read depth, this counts table was then used as input for RUVseq.

Data is first loaded into R:

```
```{r}
myfile <- "C:\\Uni\\DPhil\\Thesis\\Chapter\ 6\ Batch\ effects\
qPCR\\RUVseq\\Peak_counts_no-outliers.csv"
counts_matrix <- as.matrix(read.csv(myfile,sep=",",row.names="Peaks"))
counts_matrix2 <- counts_matrix
#counts_matrix2 <- norm_counts
counts_matrix2 <- as.matrix(counts_matrix2)
mode(counts_matrix2) <- "integer"
peakfile <- "C:\\Uni\\DPhil\\Thesis\\Chapter\ 6\ Batch\ effects\
qPCR\\RUVseq\\peakfile.narrowpeak.csv"
```
```

DeSeq2 normalised or non normalised counts table input here

It should be noted data input from DeSeq2 contains decimalized numbers, which cannot be input into RUVseq, so input from DeSeq2 had to be converted to an integer matrix, this creates a small amount of data loss.

The peaks were annotated using ChIPseeker, the experimental factors were saved as a factor variable, and the counts table converted to a RUVseq dataset object

```r
peak <- readPeakFile(peakfile,header=FALSE)
peakAnno <- annotatePeak(peak, tssRegion=c(-3000, 3000),
                         TxDb=txdb, annoDb="org.Mm.eg.db")
anno_table <- as.data.frame(peakAnno)

ExperimentSetup <- as.factor(rep(c(rep("CT16LP30",4),
rep("CT16LP60",4),rep("CT16SHAM",2),
rep("CT8LP30",4),rep("CT8LP60",4),rep("CT8SHAM",4))))

RUVseqDataSet <- newSeqExpressionSet(counts_matrix2)
```

The annotated peak table was used to create a master counts table where the peaks are converted to their nearest gene:

```r
colnames(anno_table)[6] <- "Peaks"
counts_master <- as.data.frame(counts_matrix2)
counts_master$Peaks <- row.names(counts_master)
colnames(anno_table)[colnames(anno_table)=="SYMBOL"] <- "Gene_ID"
counts_master <- merge(x= counts_master, y =
anno_table[,c("Peaks","Gene_ID","annotation")], by = "Peaks")
```

SCN time of day RNA-seq and light pulse micro-array data are uploaded as control genes; only genes which are expressed, but whose expression does not change by light or time of day are included in the final housekeeping gene lists:

```r
myfile2 <- "C:\\Uni\\DPhil\\Thesis\\Chapter\ 6\ Batch\ effects\ qPCR\\RUVseq\\aarti_sik1.csv"
myfile3 <- "C:\\Uni\\DPhil\\Thesis\\Chapter\ 6\ Batch\ effects\ qPCR\\RUVseq\\time_of_day.csv"

light_rna <- as.data.frame(read.csv(myfile2,sep=",", stringsAsFactors=FALSE))
timeofday <- as.data.frame(read.csv(myfile3,sep=",",header=TRUE, stringsAsFactors=FALSE))

light_rna$LP30induction <- (light_rna$LP30 - light_rna$Sham)
light_rna$LP60induction <- (light_rna$LP60 - light_rna$Sham)

hk_time <- filter(timeofday, Dynamic_range < 2, Expression_level > 100, DESeq2.padj > 0.05)
hk_time_genes <- as.data.frame(hk_time$Gene_ID)

hk_light <- filter(light_rna, LP30induction < 0.25, LP30induction > -0.25,
LP60induction < 0.25, LP60induction > -0.25)  ##Rough guess
hk_light_genes <- as.data.frame(hk_light$Gene_Symbol)

colnames(hk_light_genes)[colnames(hk_light_genes)== "hk_light$Gene_Symbol"] <- "Gene_ID"
colnames(hk_time_genes)[colnames(hk_time_genes)== "hk_time$Gene_ID"] <- "Gene_ID"

genes_in_both <- merge(x= hk_time_genes, y = hk_light_genes, by = "Gene_ID")
```

The counts for the control genes are retrieved from the master counts table, and only genes annotated as promoter are included:

```r
hk_light_counts <- merge(x= counts_master, y = hk_light_genes, by = "Gene_ID")
hk_time_counts <- merge(x= counts_master, y = hk_time_genes, by = "Gene_ID")
hk_both_counts <- merge(x= counts_master, y = genes_in_both, by = "Gene_ID")

hk_light_counts_prom <- filter(hk_light_counts, grepl("Promoter",annotation))
hk_time_counts_prom <- filter(hk_time_counts, grepl("Promoter",annotation))
hk_both_counts_prom <- filter(hk_both_counts, grepl("Promoter",annotation))
```

The counts for housekeeping genes are used by RUVseq to normalize the dataset, three datasets are calculated, for light, time of day and both housekeeping genes:

```r
NormalisedLightHK <- RUVg(RUVseqDataSet, hk_light_counts_prom$Peaks, k=1)
NormalisedTimeHK <- RUVg(RUVseqDataSet, hk_time_counts_prom$Peaks, k=1)
NormalisedbothHK <- RUVg(RUVseqDataSet, hk_both_counts_prom$Peaks, k=1)
```

PCA plots are generated for each new dataset, to analyze how the dataset has been modified:

```r
plotPCA(NormalisedLightHK, col=colors[ExperimentSetup], cex=1.2,
main = "PCA: light HK genes", pch=19, labels=FALSE)
par(mar=c(10.5, 4.1, 4.1, 2.1),xpd=FALSE)
legend("bottom",col=c(colors[ExperimentSetup][2],
colors[ExperimentSetup][5],colors[ExperimentSetup][10],
colors[ExperimentSetup][14],colors[ExperimentSetup][18],
colors[ExperimentSetup][22]), legend = c(as.character(ExperimentSetup[2]),
as.character(ExperimentSetup[5]),as.character(ExperimentSetup[10]),
as.character(ExperimentSetup[14]),as.character(ExperimentSetup[18]),
as.character(ExperimentSetup[22])),pch = 19,bty='n',cex=0.75,
xpd = TRUE,horiz = TRUE, inset = c(-0.41,-0.41))
```

2.10.8: Motif Searching

Background and motif searching program methodology

DNA sequence motifs are short, recurring patterns in DNA which often have a biological function, and they are frequently located in regulatory components of DNA such as promoters and enhancers. Here they often act as cis-regulatory elements, acting as binding sites for trans-regulatory elements such as transcription factors. The program Homer (Heinz et al., 2010) was used to carry out motif searching; when carrying out motif searching, Homer will take a target input dataset from either:

- A gene list such as the differential gene list obtained from DESeq2, the program will search in regulatory regions located close to these genes for enrichment of specific motifs.
- Alternatively a peak file may be used; in this case the program will search only in regions of the genome found within the peaks for motif enrichment.

Homer will first employ a 'known' motif search, and then use this as the basis for a 'de-novo' motif search. In the initial stage, the program will use an in-built database of known motifs, comprised of motifs from the following sources:

- **ChIP-Seq Transcription Factor Motifs:** The vast majority of motifs in the database are based on the analysis of published ChIP-Seq data.
- **Promoter Motifs:** Due to the importance of promoter regions, the database contains a library of motifs found by analysing promoter regions vs. random genomic regions.
- **General Factors X & Y:** These are binding regions that we essentially do not know what binds to them, but they do however appear in many papers, so are therefore included.

Homer will use a background dataset either given by the user, or its own in-build background dataset comprised of confident, non-regulated promoters. The program will then look for motifs from within the known motifs database that are over-represented in the target set relative to the background set.

Once a list of known motifs has been obtained the program will then carry out the de-novo search; this is broken down into several steps:

- **Extract the DNA sequences from the input;** if a list of genes is used the reference genome will be used to extract DNA sequences located near to the gene, as default the program looks 200bp up and down stream. While if a peak file is used, the program will convert the peaks into DNA sequences, again from the relevant reference genome.
- **Create an oligo table from the DNA sequences;** The DNA sequences will then be broken down into smaller DNA fragments, by default the program uses lengths of 8, 10 and 12bp, which are parsed into an oligo table; this is now essentially a motif table to search against.
- **Global search and motif optimization:** Homer will perform a search for enriched oligos based on the oligo table; any enriched oligos will then be transformed into simple position specific probability matrices. These matrices are much like the ones found within other motif databases such as Jaspar.
- **Mask and Repeat:** The enriched oligos having been transformed into probability matrices will now be masked and the sequences bound by the motif are removed from the analysis, i.e if the Motif was ATGCW, ATGCT and ATGCA would be removed from the oligo table, preventing the same motif being optimized again. This continues with each of the most enriched oligos until the user defined number of de-novo motifs has been reached.
- **Comparison with known motifs:** Finally de-novo motifs are identified from their best match in the JASPAR and the known motif database.
- **Output a list of identified de-novo motifs**

Running Homer for ATAC-seq and RNA-seq data analysis

As previously stated, Homer will take input as either a gene list, or peak list; using either of these methods has their merits, but for ATAC-seq data an amalgamation of both was used. Following differential analysis, a list of genes whose chromatin is either more or less open between different sample conditions will have been obtained. However, to simply use this list would not make best use

of the data that was available because from the ATACseq alignment .bam files, it is known exactly where chromatin within the DNA of regulatory elements of the differential genes is open or closed. Therefore a list of peak locations for areas of chromatin near to genes identified as having differential chromatin nearby was followed generated as follows:

- Annotate peaks from the master peak file.
- Identify all annotated peaks whose corresponding gene is also in the differential gene list.
- Generate a .csv peak file with containing only differential peaks.
- Use differential peak list as input into Homer.

For ATAC-seq the following command was then used:

findMotifsGenome.pl <DifferentialPeakLocationList> <mm10> </outputdir> -size given

For RNA-seq the differential gene list was simply used as input, with the following command:

findMotifs.pl <DifferentialGeneList> <mm10> </outputdir> -start -400 -end 100

The genome is set to mm10; for ATAC-seq the size parameter ensures the program searches within the peaks given; this parameter is used because a broad peak file was used, so the full range of the peak region is given. For RNA-seq the –start and –end parameters determine how far up and downstream of genes will be searched for differential motifs.

Search for occurrences for specific motifs

The genes containing enriched motifs or motifs of particular interest can be identified using Homer. The differential gene list is used as input along with the motif of interest; this will output a gene list, identifying where the motif of interest has differential chromatin accessibility. The program requires a .motif file, in a Homer specific format; if an enriched motif is used, this can be obtained from the

output of the Homer motif enrichment search. However, to search for occurrences of circadian related motifs of interest, Jaspar (Fornes et al., 2020) formatted motifs had to be converted to Homer formatted motifs.

For CLOCK



Jaspar Format

>MA0819.1 CLOCK

A [	226	429	11	429	16	40	3	0	1	89
C [	35	85	429	5	429	0	5	3	152	128
G [	93	89	2	15	0	429	5	429	110	39
T [	75	13	2	0	28	0	429	9	429	174

Each column has to be totalled and the chance of each base is calculated in decimal form, the data is then transposed into 4 columns, rather than 4 rows, as below:

For the first base:

Total	429
A [	226 / 429 = 0.53
C [	35 / 429 = 0.08
G [	93 / 429 = 0.22
T [	75 / 429 = 0.17

Homer Format

>AACACGTGTT CLOCK 8.05

A	**C**	**G**	**T**
0.53	0.08	0.22	0.17
0.7	0.14	0.14	0.02
0.02	0.97	0	0
0.96	0.01	0.03	0
0.03	0.91	0	0.06
0.09	0	0.91	0
0.01	0.01	0.01	0.97
0	0.01	0.97	0.02
0	0.22	0.16	0.62
0.21	0.3	0.09	0.4

Homer has the additional field in this case ‘8.05’; this is the Log odds detection threshold, used to determine bound vs. unbound sites. This is essentially how perfect the match has be between the motif and whatever is binding the sequence.

When searching for specific motifs, the following thresholds were used:

- Stringent = 8
- Moderate = 5
- Relaxed = 2

2.10.9 Gene set enrichment analysis (GSEA)

For both ATAC-seq and RNA-seq this process is the same; once a gene list has been generated either by annotation of genes found close to or within differential chromatin, or generated from genes shown to be differentially expressed (DE), GSEA can be carried out. Lists of genes located near to enriched motifs can also be used as input for GSEA. Analysis looks for genes with common individual functions, and genes that are found within similar molecular pathways. Several databases are used to determine gene functions, and what pathways they are involved in, including:

For gene function ontology:

- **Gene Ontology (GO) knowledgebase:** World's largest database of gene functions including commonly used GO biological function (GO BF) and Go molecular function (GO MF).
- **Clusters of Orthologous Groups of proteins (COG):** Phylogenetic classification of the proteins encoded by 21 complete genomes including mouse

For Pathway analysis:

- **Kyoto Encyclopedia of Genes and Genomes (KEGG) Pathway**: Maps of the molecular interaction, reaction and relation networks.
- **Reactome:** An open-source, open access, manually curated and peer-reviewed pathway database.
- **BioCarta:** Provides Online maps of metabolic and signalling pathways

If many genes share a common and unexpected function, new hypotheses can be devised relating to circadian entrainment and the effects of SD. Genes within these clusters can then be looked at more closely to see what their function is in other tissues, and may provide candidates for further investigation in follow up experiments.

Enrichment analysis must always be viewed with a note of caution that the functions assigned to many genes may be tissue specific, and are often determined from the first tissue from which the gene was discovered to be expressed within. It is therefore possible that the function for genes given from

the databases above may not be a true reflection of what function the gene is carrying out in the SCN. Indeed, it is often in fact chromatin structural changes across the genome in different tissues that allows genes to carry out different functions (Felsenfeld et al., 1996).

Final Quality Control from pathway analysis:

As SCN tissue will be assayed for chromatin accessibility and gene transcription, it is expected that genes from the differential lists would be enriched for functions related to circadian biology, and pathways involved with entrainment. Although this may provide few useful new insights, it is a good indicator that the experiment has been carried out correctly, and is an important final step for the quality control process. If few genes are identified which are involved in circadian biology it may indicate the experiment was not carried out correctly, or that the brain punches taken were not wholly or only partially SCN tissue.

Database for Annotation, Visualization and Integrated Discovery (DAVID) gene function and pathway enrichment analysis:

The online webserver DAVID was used to carry out both a gene function and pathway analysis. The differential gene list was first converted in Entrez IDs, this conversion is carried out because Entrez IDs are species specific, unlike gene symbols, and therefore prevent any incorrect gene function and pathway assignment. The Entrez gene list was then uploaded into DAVID, and the webserver searched against all the above databases, GO terms output from DAVID were then visualised using Revigo (Supek et al., 2011). No axis are included for figures generated using Revigo, because distances within the plot have no intrinsic meaning, semantically similar GO terms are merely plotted close together within the figure.

Chapter 3

Refinement of nuclei extraction for ATAC-seq and subsequent pilot ATAC-seq experiments

The key hypothesis underpinning this project is that chromatin structure, and therefore accessibility, changes within the SCN on a daily basis and in response to external stimuli, such as light exposure and SD. Therefore, to investigate this hypothesis, ATAC-seq was used to assay chromatin accessibility. As the isolation of intact nuclei forms the starting point for ATAC-seq, the first objective of this project was to successfully isolate intact nuclei from SCN neuronal tissue. Since the published ATAC-seq protocol is optimized to take cultured cells as the input material, the first approach was to simply dissociate the SCN tissue to individual cells, thus allowing the published ATAC protocol to be followed. A second approach (which ultimately proved to be the most effective) was also adopted in which the SCN was dissociated directly to intact nuclei, but this method required further refinement of the published protocol to start with nuclei rather than dissociated cells.

The tissue had to be extracted and single cells or nuclei obtained within a strict timeframe for this experiment; ideally library preparation should take place a maximum of 30 min following SCN removal. This is to minimise the possibility of chromatin changes occurring after the mice had been culled, and for this reason tissue could not be frozen either. To examine which method gave the highest quality material for input into the ATAC assay a number of different methods were investigated to achieve this goal; including enzymatic digestion, sonication, and douncing using a pestle and mortar. Following successful single cell isolation by enzyme digestion and direct nuclei extraction by douncing; two pilot experiments were subsequently carried out to determine which method gave the highest quality ATAC-seq libraries.

3.1: SCN dissociation using enzymatic digestion and sonication

It was hypothesised enzymatic digestion would provide a reliable and controllable method to isolate dissociated cells. It was hoped by finding the most favourable enzyme and optimising the length of incubation, a protocol would be established where by the enzyme would break down tissue and digest the extracellular matrix, but leave cells intact. Following tissue dissociation, the lysis buffer and procedures from the published ATAC seq protocol could be then be carried out in the same manner, as it would for cultured cells. To test this hypothesis several different enzymes with varying levels of activity, were investigated, using different lengths of incubation and agitation.

Bead homogenization of SCN tissue was also identified as a possible method of tissue dissociation, and it was hypothesised that it may provide a much quicker method, more in line with the tight time frames of any future circadian experiment. It was anticipated that bead homogenization may be able to dissociate tissue straight into nuclei, rather than into individual cells first, and this would decrease the time taken before ATAC seq library preparation could begin. This approach used the Bertin Precellys24 tissue homogenizer to physically break up the cells, and the instrument was used for varying lengths of time, homogenising tissue at different pulse frequencies, to determine the best combination for nuclei isolation.

In both of the approaches detailed above, a trade-off had to be achieved between quickly breaking down tissue to individual cells, but without causing too much damage. Ideally to follow the published ATAC seq protocol, a dissociated cell suspension had to be achieved. However it was hypothesised that if sonication gave intact nuclei, ATAC-seq would still work, but the protocol may have to be refined, with the cellular lysis step not being required.

To begin to examine the effectiveness of enzymatic tissue disaggregation, the protocol from chapter 2.1 was used to test each enzyme, and all enzymatic digestions were carried out at 37°C, for 5, 10, 15 and 20 min, with or without gentle agitation. The following enzymes were investigated:

Trypsin: A fairly aggressive digestive enzyme, which cuts at the carboxyl side of the amino acids lysine or arginine.

Collagenase: An enzyme which breaks the peptide bonds in collagen, it is produced by bacteria and often by the body as part of its normal immune response

Dispase: A protease which cleaves fibronectin, it is considered a milder treatment than both trypsin and collagenase.

Papain: A cysteine protease, and is the mildest of the 4 enzymes that were tested.

Following testing nuclei isolation using enzymes, sonication was then investigated following the protocol outlined in chapter 2.1.1. The pulse frequencies and durations that were used are outlined in table 3.1 below:

Pulse Frequency (rpm)	Duration (s)	Pulse Frequency (rpm)	Duration (s)
5000	5	5200	5
5000	2 × 5	5200	2 × 5
5000	10	5200	10
5000	2 × 10	5200	2 × 10

Table 3.1: Pulse frequencies and durations used to dissociate SCN tissue with the Bertin Precellys24 Instrument.

In both cases the viability of nuclei was checked under a microscope using ThermoFisher NucBlue™ Live ReadyProbes™ Reagent; with all viable nuclei fluorescing brightly upon exposure to ultraviolet light. To test the different approaches cortex brain punches were used to approximate SCN punches, ensuring only a minimum number of mice were sacrificed.

Papain treatment with gentle agitation was found to be the most effective method for attaining the greatest quantity of viable single cells within 30 min. Although several of the other enzymes were able to dissociate the brain tissue faster; they resulted in cell death, with very few viable cells remaining. The results are shown in Table 3.2, and Figure 3.1 shows an example of viable nuclei under a microscope following papain digestion and cellular lysis with the ATAC lysis buffer.

Enzyme	Agitation	Dissociate 5 mins	Dissociate 10 mins	Dissociate 15 mins	Dissociate 20 mins	Viable Nuclei
Trypsin	No	×	×	×	✓	×
Collagenase	No	×	×	×	✓	×
Dispase	No	×	×	×	×	×
Papain	No	×	×	×	×	×
Trypsin	Yes	×	×	✓	✓	×
Collagenase	Yes	×	×	✓	✓	×
Dispase	Yes	×	×	×	✓	×
Papain	Yes	×	×	×	✓	✓

Table 3.2: Cellular dissociation of neural tissue using a variety of enzymes. None of the enzymes gave viable cells without agitation; papain with agitation was the only enzyme that both dissociated the tissue and preserved cell viability.

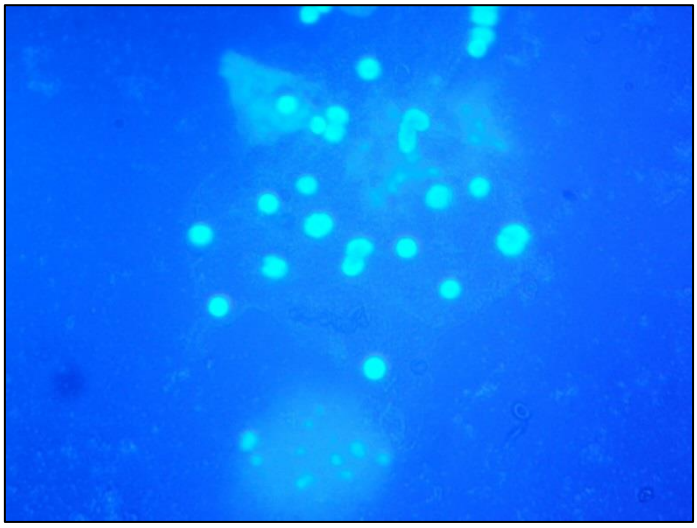


Figure 3.1: Nuclei luminescing under ultraviolet light. Nuclei were stained with ThermoFisher NucBlue™ Live ReadyProbes™ Reagent and visualised using a fluorescence microscope following brain tissue dissociation with papain and gentle agitation through a pipette.

Homogenization using the Bertin Precellys24 at all pulse frequencies and durations resulted in total loss of viable cells and nuclei; no nuclei could be seen under a microscope either straight homogenization, Table 3.3 summarises the results.

Pulse Frequency (rpm)	Duration (s)	Viable Cells after sonication	Viable Cells after lysis
5000	5	×	×
5000	2 × 5	×	×
5000	10	×	×
5200	2 × 10	×	×
5400	2 × 10	×	×
5600	2 × 10	×	×
5800	2 × 10	×	×

Table 3.3: Cellular homogenization using a Precelleys24. At all pulse frequencies and durations viable cells could not be attained.

Enzymatic digestion using papain was the only method that successfully resulted in intact nuclei being obtained, with sonication of SCN tissue using a Precelleys24 determined to be an unsuitable method for either dissociating tissue into individual cells, or straight to nuclei. Subsequent counting of nuclei following enzymatic digestion with papain, and cell lysis with the ATAC seq lysis buffer revealed sufficient nuclei had been obtained for ATAC seq library preparation.

3.2: Pilot Experiment 1: ATAC sequencing using nuclei extracted with papain

Following successful isolation of intact dissociated cells using enzymatic digestion with papain, a pilot experiment was carried out to determine if cells extracted with papain could be used to generate ATAC-seq libraries. This experiment aimed to test the full ATAC seq Library preparation protocol, including sequencing of the samples, and bioinformatic analysis of the data obtained.

ATAC seq libraries were prepared with nuclei extracted from cortex tissue following the protocol outlined in chapter 2.2; the libraries were quantified as outlined in chapter 2.4, and sequenced using the protocol detailed in chapter 2.5. All animal procedures were strictly adhered to as outlined in

chapter 2.6, with an emphasis on following the 3R (Replacement, Reduction and Refinement) methodology of research using live animals. Two ATAC-seq libraries were generated, to minimise the number of mice being used, each sample consisted of 4 brain punches from the cortex, approximating the size of 4 SCN punches. A basic bioinformatic analysis was carried as out described in section 2.9.

To begin to assess the success of ATAC library preparation, initial QC was conducted by quantifying the concentrations of the libraries (Table 3.4) as described in section 2.4, with sample 1 displaying a high DNA concentration, whilst sample 2 had the opposite result, with a very low DNA abundance, below the minimum ideally required for sequencing.

Experiment	Sample name	Concentration (nM)
Pilot 1	Sample 1	5.22
Pilot 1	Sample 2	0.12

Table 3.4: The concentrations of the first ATAC-seq libraries generated.

Following sequencing of the two libraries it was clear that the library preparation had been unsuccessful. Further QC from the distribution of sequencing reads found that the two libraries from the first pilot experiment only accounted for a very small number of the reads from the sequencing run (Figure 3.2 - with this run also including the libraries from the second pilot experiment detailed below). Read coverage across the genome was visualised using The Integrated Genome Browser (IGB) (Freese et al., 2016), and very few reads were seen for sample 2, likely because the DNA concentration was so low (Figure 3.3). For sample 1, reads can be seen throughout the genome indicating that the Tn5 enzyme had not only cut out and tagged areas of open DNA, but had also tagged DNA fragments from all over the genome, as reads can be seen aligning to both exons, introns, and throughout areas of the genome between genes. This strongly suggested that free DNA must have been present from all parts of the genome, regardless of accessibility within chromatin, most likely due to excessive nuclei lysis and the subsequent release and chromatin structure loss of all DNA.

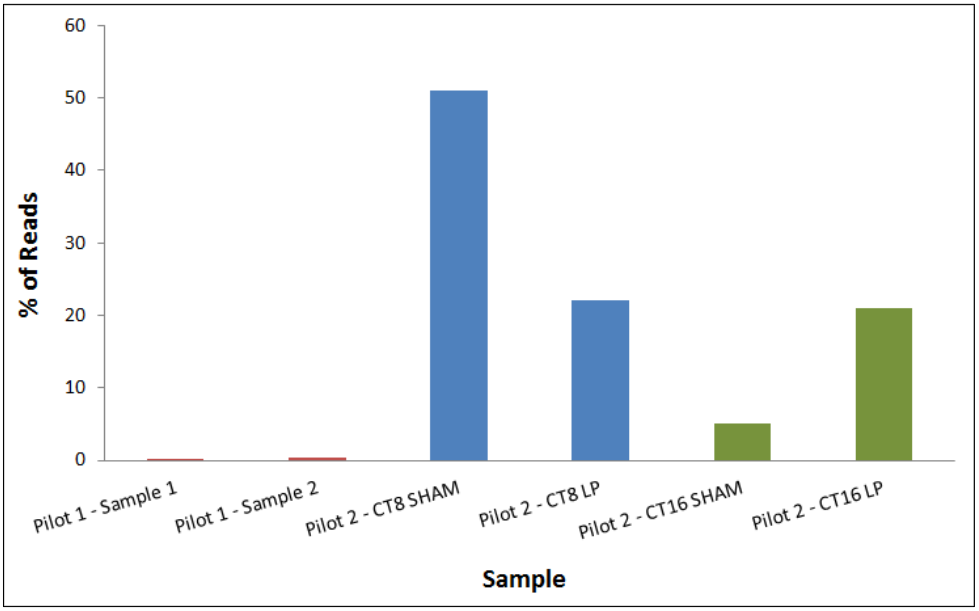


Figure 3.2: Read distribution across the different samples for both pilot experiments. Samples were sequenced together.

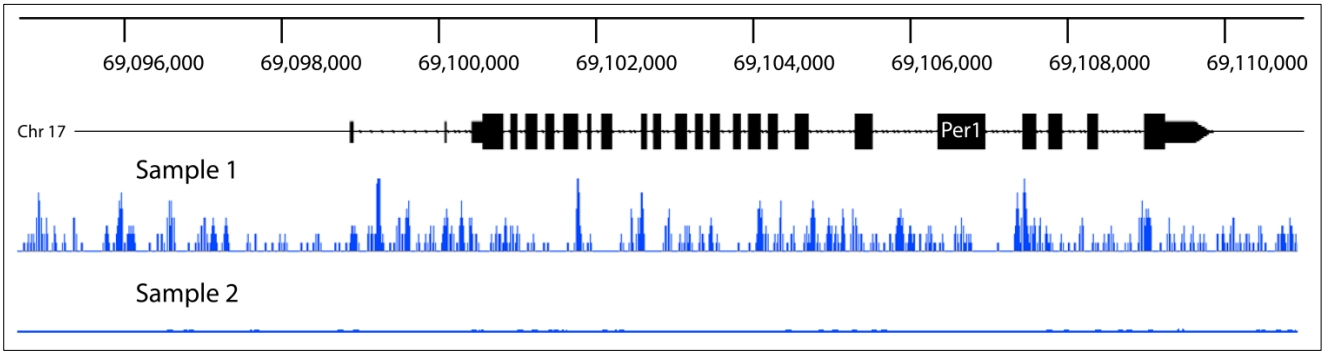


Figure 3.3: Read coverage for Pilot 1 near to *Per1*. In sample 1, reads were found across the genome in areas of chromatin expected to be closed; in sample 2, very few reads were sequenced, indicating poor library preparation.

This result clearly demonstrated that papain was not a successful method for extracting intact cells and assaying chromatin structure using the Tn5 enzyme. Despite papain being the least aggressive enzyme, and giving intact nuclei as shown in Figure 3.1, it likely still caused the lysis of some nuclei, which released sufficient free DNA to create so much background noise that any signal from DNA tagged from within intact nuclei was lost. This technique for nuclei isolation was therefore dismissed as a possible method, and instead a more suitable approach for rapidly generating isolated nuclei was sought.

3.3: Pilot Experiment 2 using nuclei isolated with a Dounce homogenizer

Following the poor initial QC of samples from pilot experiment 1, and concerns about the long time period papain treatment took to dissociate tissue; a faster more effective method of nuclei isolation was needed. A new nuclei isolation protocol from (Su et al., 2017) was identified as a potentially more effective method; this protocol had already successfully been used to isolate intact nuclei from DG tissue. Subsequent efficacious application of this protocol enabled a second more in depth pilot experiment to be carried out, therefore allowing the ATAC-seq library preparation method to be examined, and also the hypothesis that chromatin changes state following light input to the SCN to be tested. The experiment aimed to investigate the proposed hypothesis without culling excessive numbers of mice, again in line with the 3Rs approach to research using live animals. If the pilot proved successful, data generated could be incorporated into future experiments, meaning fewer animals would have to be sacrificed, and if the protocol and/or hypothesis are completely fallacious, a minimum number of animals would have been used.

This project hypothesises that chromatin changes occur within the SCN over the course of a day and in the presence of a light pulse. Mice were culled at two time points; CT8; a non-light responsive time point, and CT16; a light responsive time point, with or without a 30 minute light pulse (400 lux) (Minors et al., 1991). Although ATAC-seq is able to work effectively with only a small number of cells; the SCN in mice contains as little as 10,000 cells (Welsh et al., 2010), so the quantity of starting material may provide a potential problem. Therefore even by pooling SCN from several mice, the number of cells available for ATAC-seq may still be quite low; as it is inevitable that not all cells will remain intact following initial SCN removal, and subsequent homogenization. Initial testing using cortex punches, and nuclei counting with a haemocytometer determined at least three brain punches would be required, but it was uncertain if this would be sufficient for optimum library preparation. To determine the ideal number of SCN tissues punches, samples were prepared using a varying number. If the experiment proved successful, read counts would be normalised within the analysis to account for varying starting material.

The pilot experiment was broken down into four experimental groups as shown in Table 3.5:

Sample name	Group	Time of Day	Light Pulse (LP)	SCN
CT8 SHAM	Group 1	CT8	None	5
CT8 LP30	Group 2	CT8	30 min	4
CT16 SHAM	Group 3	CT16	None	3
CT16 LP30	Group 4	CT16	30 min	3

Table 3.5: The four experimental groups for the second pilot experiment.

The experiment aimed to answer the following questions:

- How many SCN are needed for optimum ATAC seq library preparation?
- Are there time of day chromatin changes between CT8 and CT16?
- What chromatin changes (if any) are occurring in the presence of a light pulse at either CT8 or CT16?

Light pulsing of mice was carried out as described in chapter 2.6 in line with the experimental objectives outlined above. All animal procedures were strictly adhered to as defined throughout section 2.6, and nuclei were isolated using a Dounce homogenizer with a pestle and mortar as detailed in chapter 2.1.2. ATAC libraries were then generated following chapter 2.2, with the subsequent libraries quantified as outlined in chapter 2.4, and sequenced using the protocol in chapter 2.5. Bioinformatic analysis of data obtained was carried out using the basic methodology described in chapter 2.9. Initial examination of nuclei following douncing revealed a large quantity of intact nuclei had been obtained, providing a promising start to this protocol, under ultraviolet light following treatment with NucBlue™ reagent numerous nuclei were seen fluorescing (Figure 3.4).

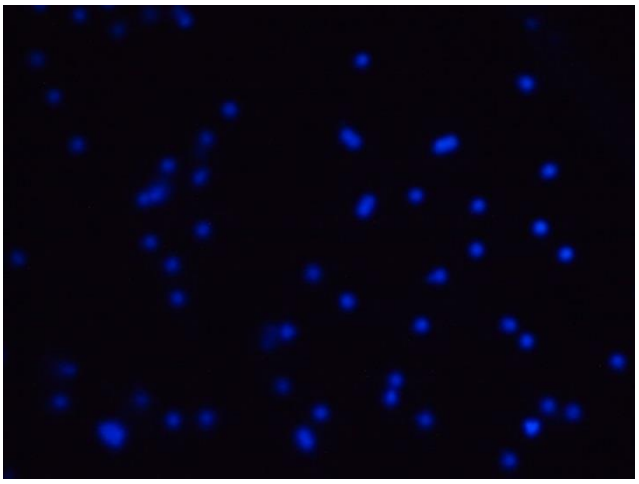


Figure 3.4: Nuclei fluorescing under ultraviolet light following extraction using a Dounce homogenizer. Numerous nuclei successfully extracted using a Dounce homogenizer luminesce under ultraviolet light following treatment with NucBlue™ reagent.

However, following library preparation and sequencing, the read distribution across the different samples was not spread evenly (Figure 3.3), either due to incorrect library dilution or incorrect dilution of other samples run on the cartridge at the same time. However, all libraries gave concentrations above the minimum amount to sequence of 0.5 nM (Table 3.6), and the samples accounted for 35.24% of all the reads run on the cartridge for this sequencing run. Table 3.6 also gives the actual percentage of reads each sample accounts for, along with their concentration before dilution. It can be seen that the original concentration did not always correlate with the eventual number of reads. This therefore further suggested that errors were made during library dilution, or that the other reads on the column affected the read distributions for each sample.

Experiment	Sample name	Concentration (nM)	% of reads
Pilot 2	CT8 SHAM	4.01	51.4
Pilot 2	CT8 LP30	4.2	22.2
Pilot 2	CT16 SHAM-1	0.55	4.6
Pilot 2	CT16 LP30-2	10	21.1

Table 3.6: The concentration of DNA libraries from the second pilot experiment prior to sequencing. Also shown are the percentage of reads each sample accounted for within the sequencing run.

Following initial QC, all samples from the second pilot appeared to have a reasonable sequencing depth, although sample CT16 LP was lower than the other samples. Reads were successfully aligned to the mouse mm10 reference genome, and examination of the sequencing coverage was indicative of successful ATAC-seq library preparation, as reads were found in areas of the genome that would be expected to be open, such as promoters and enhancers, whilst few reads were found in areas such as exons, which would usually be closed (Figure 3.5).

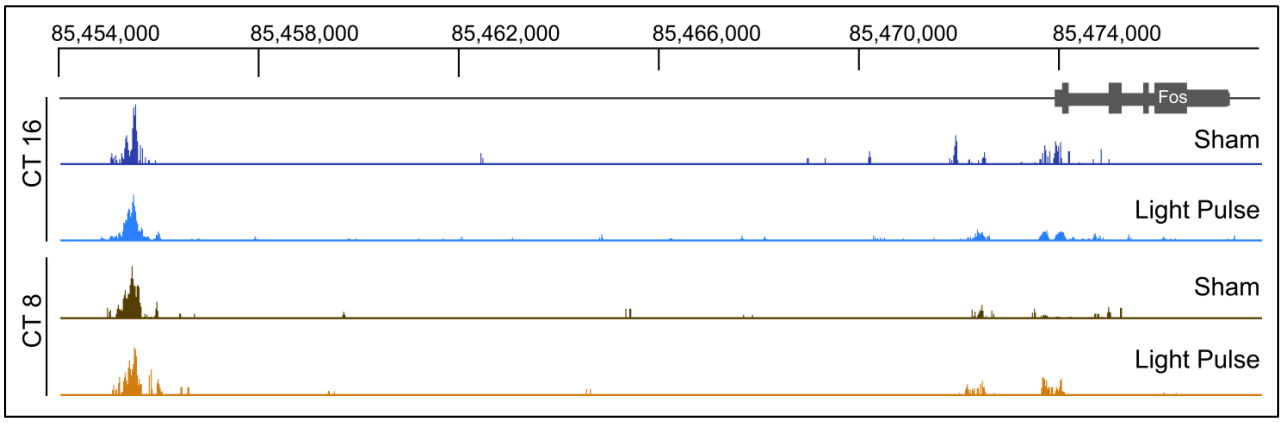


Figure 3.5: Coverage tracks for *Fos* from the pilot experiment 2. Data visualized using the IGB; reads can clearly be seen located at the upstream enhancer, indicating it is open. Whilst few reads can be seen in inter-gene regions and at exons, indicating the chromatin is closed as expected in these regions.

Although a full differential analysis could not be carried out due to a lack of replicates, several key molecular clock genes were viewed in the IGB, and changes in chromatin could be seen for several genes. Figure 3.6 shows the coverage tracks for the genome located near *Csnk1e*, a core clock gene responsible for marking of PER1 and CRY for degradation and translocation back into the nucleus. Fewer reads can clearly be seen at the promoter for the sample light pulsed at CT16, indicating that chromatin was more closed at this region; therefore it can be speculated this may lead to the gene being switched off.

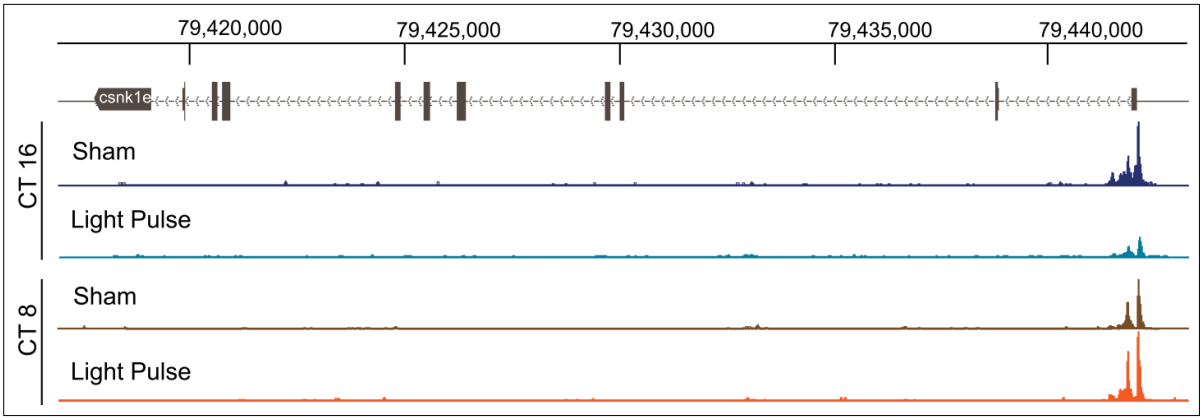


Figure 3.6: Coverage tracks for *Csnk1e* from the pilot experiment 2. Data visualized using the IGB; chromatin was observed to be closed following a light pulse at CT16.

3.4 Chapter 3 Discussion

Following visualisation of chromatin accessibility in the IGB for the second pilot experiment, it was determined that a viable method for carrying out ATAC-seq library preparation on mouse SCN tissue had been achieved. This therefore enabled a much larger ATAC-seq experiment to be planned with multiple replicates for each experimental condition. Although a full differential analysis could not be carried out for the pilot experiment due to lack of replicates, the differential peaks identified for several circadian genes gave promise that my original hypothesis may be correct. Chromatin was clearly closing following a light pulse only at CT16 for several key circadian genes. This would therefore indicate that a light pulse may be switching off these genes, only at light responsive times of day, and therefore this is highly likely to have an effect on an individual's circadian rhythm.

Chapter 4

The effect of light on chromatin structure within SCN

4.1 Introduction

The SCN is entrained to the external light/dark cycle by photoreceptors located almost solely within the retina of the eye, and synchronisation is mainly co-ordinated by photosensitive retinal ganglion cells (pRGCs), which utilise the photopigment melanopsin (OPN4) (Freedman et al., 1999). Although pRGCs provide the primary input to the SCN, rods and cones provide further input, and either one can compensate for the other (Hattar et al., 2003). Following activation of pRGCs, the neurotransmitters glutamate and PACAP are released at the terminals of the retinohypothalamic tract (RHT) (Hannibal et al., 2000), which projects directly to the SCN leading to an influx of Ca^{2+} into SCN neurones. Although light input to the SCN, via the RHT is well understood, downstream molecular mechanisms leading to entrainment within the SCN are still relatively unknown.

Initial influx of Ca^{2+}, has been shown to have an regulatory effect on a number of kinases, including PKA, PKG, CaMK, and MAPK (Meijer and Schwartz, 2003), but their actual role within entrainment remains elusive. Despite this, through phosphorylation, all these pathways are all thought to converge on the cAMP response element binding protein (CREB) (Ginty et al., 1993, Gau et al., 2002). Following phosphorylation, CREB is able to bind cAMP response elements (CRE) in the promoters of genes regulated by light in the SCN, such as *Per1* (Schwartz et al., 2011), *Fos* (Lin et al., 1997) and *Egr1* (Porterfield et al., 2007).

Despite the induction of light inducing molecular clock genes such as *Per1* within the SCN, the effect of light on molecular clock proteins is not consistent across the whole day. Indeed, entrainment occurs primarily in the morning (Khalsa et al., 2003) for diurnal individuals such as humans, with any light received in the evening having a very diminished effect. Furthermore, even when light is encountered at the optimum time of the day, entrainment is limited to around two hours (Vitaterna et al., 2006),

and any light received after this will not result in further shifts to the molecular clock (Minors et al., 1991). However, previous work from our laboratory has already began to resolve some of these unanswered questions; adenosine accumulation throughout the wakeful period has been identified as the most likely explanation for diminished entrainment in the evening (Jagannath et al., 2021).

Moreover, the kinase *Sik1*, has been shown to act as the brake on entrainment after approximately one of hour of light activation within the retina (Jagannath et al., 2013). Following light exposure the same molecular pathways which activate *Per1* through CRE binding motifs, also activate *Sik1* (Figure 1.4). Continuous light exposure therefore not only up regulates PER1 production to shift the molecular clock, but also facilitates the gradual accumulation of SIK1 within the nucleus. SIK1 following a delay of around an hour subsequently phosphorylates and deactivates CRTC, and therefore blocks further CREB driven PER1 production. This as a result limits further shifting of the molecular clock through activation of *Per1* via CREB, in vivo knockdown of SIK1 production has been shown to speed up phase shifting following light exposure at CT14 (Figure 4.1).

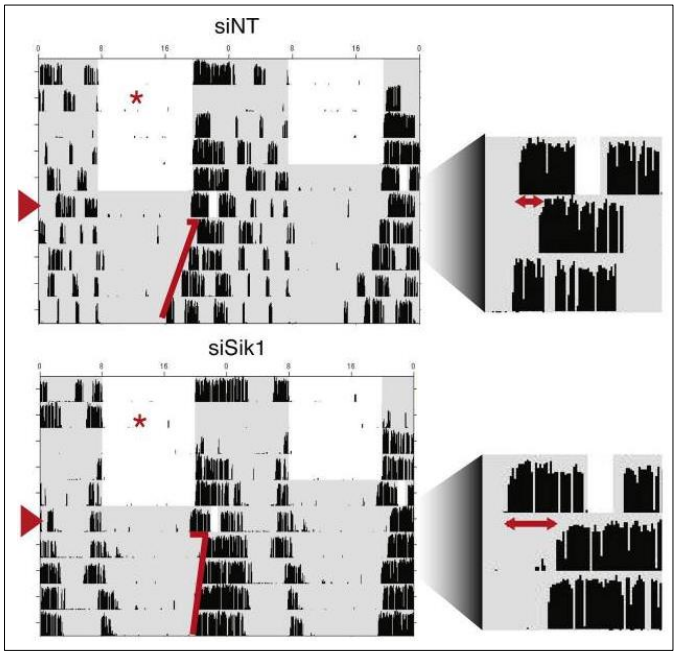


Figure 4.1: In Vivo Knockdown of Sik1 results in enhanced phase shifting. Actograms from light-pulsed animals receiving siNT (top) or siSik1 (bottom), animals which received siSik1 showed a greater phase shift in their circadian rhythm the next day compared to those with *Sik1* intact. Figure taken from Jagannath et al. (2013).

In spite of these recent advances in our understanding of entrainment within the SCN, there still remain many unanswered questions. Microarray analysis of the mouse transcriptome within the SCN following light pulsing of varying lengths shows that many genes are both up and down regulated in response to light (Jagannath et al., 2013), yet we have little idea of what role many of these genes are playing within the SCN. SIK1 is able to switch off increased PER1 production by repressing its activation through CRE, but *Per1* expression should still be driven by binding of SRE and AP-1 motifs (Figure 1.4), but does not appear to be the case if shifts to the molecular clock are limited to 2 hours. This projects aims to investigate how light impacts the regulation of genes within the SCN at the level of chromatin accessibility, with the objective to determine why so many genes have such diverse expression following light exposure.

Chromatin structure plays an important role in regulating when genes are switched on and off (Felsenfeld et al., 1996), to further investigate the expression of genes within the SCN, ATAC-seq was employed to assay chromatin accessibility. Chromatin modifiers have already been shown to induce chromatin structural changes on a daily basis in a circadian manner within the brain through Methyl CpG binding protein 2 (MeCP2) (Martínez de Paz et al., 2015). MeCP2 is a chromatin bound protein that alters chromatin by acting through a number of other chromatin modifiers (Ausió et al., 2014), it is very abundant in the brain and is under circadian control (Martínez de Paz et al., 2015). Furthermore, chromatin structure has been shown to change within the DG following electrical stimulation and subsequent activation of neurons (Su et al., 2017). It is therefore sensible to hypothesise that chromatin structure in neurons of the SCN will be regulated in a circadian manner, and may change following activation by light via the RHT. This project aims to determine if this is indeed the case by investigating the chromatin state near to and within genes expressed in the SCN, at differing times of a day. Chromatin state will also be investigated to determine if it is also changing near to and within genes which are induced in the SCN following a light pulse. This will help determine additional genes whose expression may be changing in the SCN following a light pulse, and investigate if the chromatin structure of genes known to change their expression after light input is changing to facilitate their up or down regulation.

Changes to chromatin accessibility are frequently accomplished through epigenetic modification, by methylation of DNA, and alteration of histone proteins. Epigenetic modification of DNA has been shown to play a crucial role in the regulation of neuronal gene expression, with changes to epigenetic markers leading to differences in neuronal activity (Gräff et al., 2011). Using ATAC-seq this project will investigate if changes in chromatin accessibility are occurring where known epigenetic markers bind and alter DNA, and to determine if the chromatin structure of epigenetic modifiers such as DNA methyltransferases is also changing to facilitate their up-regulation.

Very little is currently known about how chromatin accessibility in the SCN changes over the course of a day and in response to light, despite chromatin structure already been shown to change across other brain regions. It is probable that a number of different external factors will influence chromatin modification within the SCN, including recent sleep history, and the time of and length of light exposure. Within the SCN, it is highly likely epigenetic modification is frequently occurring, and will result in changes to chromatin structure. The effects of external influences on the neuroepigenetic landscape, will then integrate with epigenetic modifications that occur in a daily manner in line with an individual's circadian rhythm. The combination of regular circadian epigenetic modification and those occurring due to external influences such as light will determine the accessibility of chromatin throughout the genome. ATAC-seq will enable the accessibility of chromatin across the whole genome within the SCN to be determined at different times of the day, with and without light input to the SCN.

Light signalling to the SCN is known to cause wide ranging changes in gene expression, but how these changes might be controlled through modifications of chromatin accessibility remains elusive. This project aims to assay the accessibility of chromatin in an attempt to determine exactly how chromatin structure may change over the course of a day and following a light pulse, in an attempt to fill this gap in our current knowledge.

The project aims to answer the following questions:

- Is chromatin accessibility changing over the course of the day within the SCN?
- Does differential chromatin remodelling occur in response to light input, and are any changes restricted to one particular time of the day?
- Following changes in chromatin accessibility within the SCN, does this predispose the clock to be more or less light responsive at certain times of day?
- Are genes expressed within the SCN already known to change their expression at certain times of the day and following light exposure, also showing changes in chromatin structure?
- Can new genes and pathways that regulate circadian entrainment be identified by daily variations in chromatin structure?

4.2: ATAC-seq Experiment 1 Results

Within the SCN, how chromatin accessibility changes over the course of the day, and in response to light input remains unknown. Therefore ATAC sequencing was employed to investigate changes in chromatin accessibility within the murine SCN at different times of the day and following a light pulse.

Mice were light pulsed for 30 and 60 mins (400 lux) at two time points (CT8 /CT16), as described in section 2.6. Nuclei were isolated as defined in section 2.1, and ATAC-seq libraries prepared as per section 2.2. The libraries were subsequently quantified and sequenced as outlined in section 2.4 and 2.5 respectively; data was analysed using in-depth bioinformatic techniques set out in section 2.10.

For each sample group, 4 replicates were taken, and collectively samples were comprised of 4 SCN punches, therefore each sample condition was comprised of 16 mice, Table 4.1 summarises the different sample conditions:

Sample Name	Circadian Time	Light Pulse?
CT8 Sham	CT8	None
CT8 LP30	CT8	30 min
CT8 LP60	CT8	60 min
CT16 Sham	CT16	None
CT16 LP30	CT16	30 min
CT16 LP 60	CT16	60 min

Table 4.1: The experimental groups for the sleep deprived experiment. Each sample condition comprised 4 replicates.

4.2.1: Quality Control

After obtaining sequencing data, all samples were subjected to a range of quality control checks, to understand overall sample quality. This included checks to determine sequencing quality, the standard of ATAC-seq library preparation, robustness of alignments, and initial indication of overall data quality.

Initial read count analysis across different sample groups appeared fairly even, indicating good quality libraries and a successful dilution before pooling, however no reads were identified for two CT16 sham samples as these did not sequence correctly (Figure 4.2).

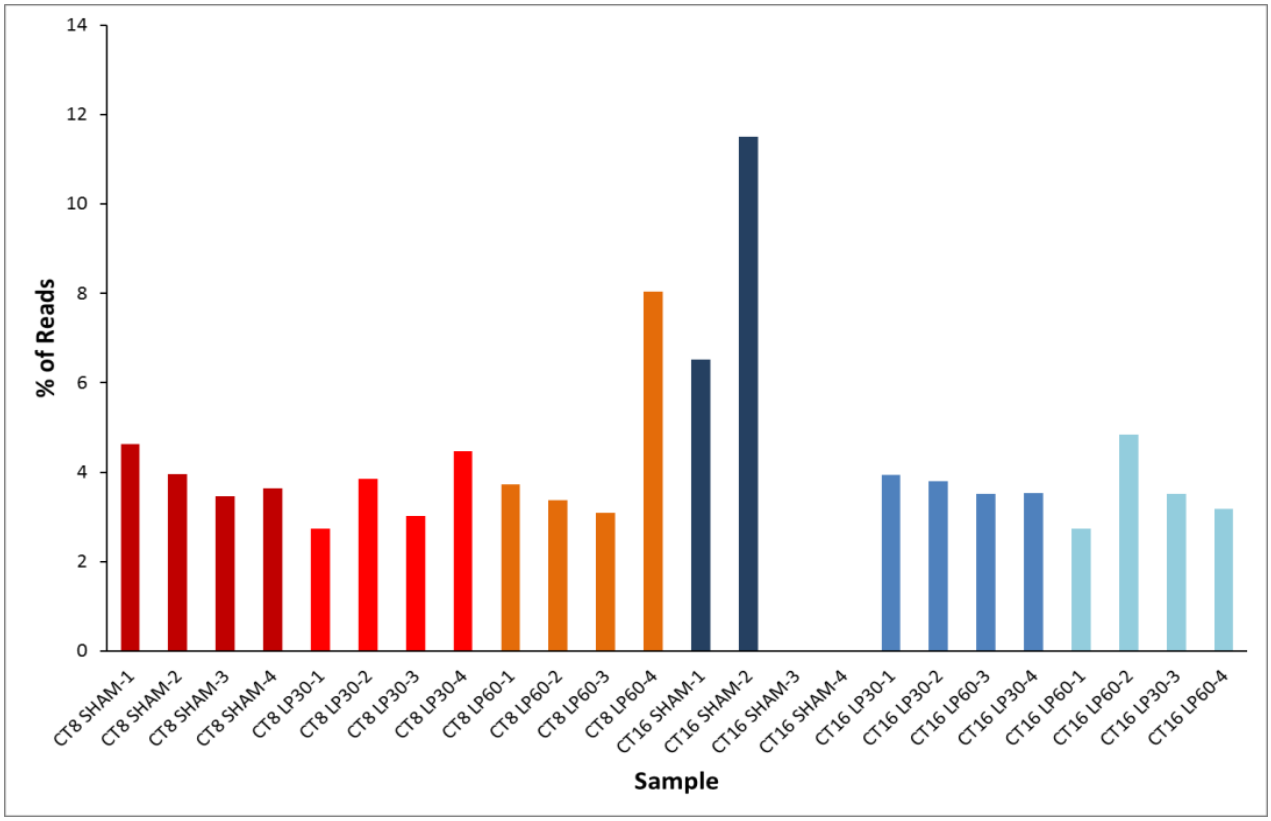


Figure 4.2: Read distribution across the different ATAC-seq different samples for light pulse experiment 1. An even distribution can be seen, indicating good library preparation, and even dilution of samples prior to sequencing, two CT16 SHAM libraries did not sequence.

The overall read quality was high, with all samples displaying good per sequence (Figure 4.3A) and per base (Figure 4.3B) quality scores. The average GC content was higher than predicted (Figure 4.3C), and a degree of sequence duplication was detected (Figure 4.3D), however this is not unexpected for ATAC-seq libraries.

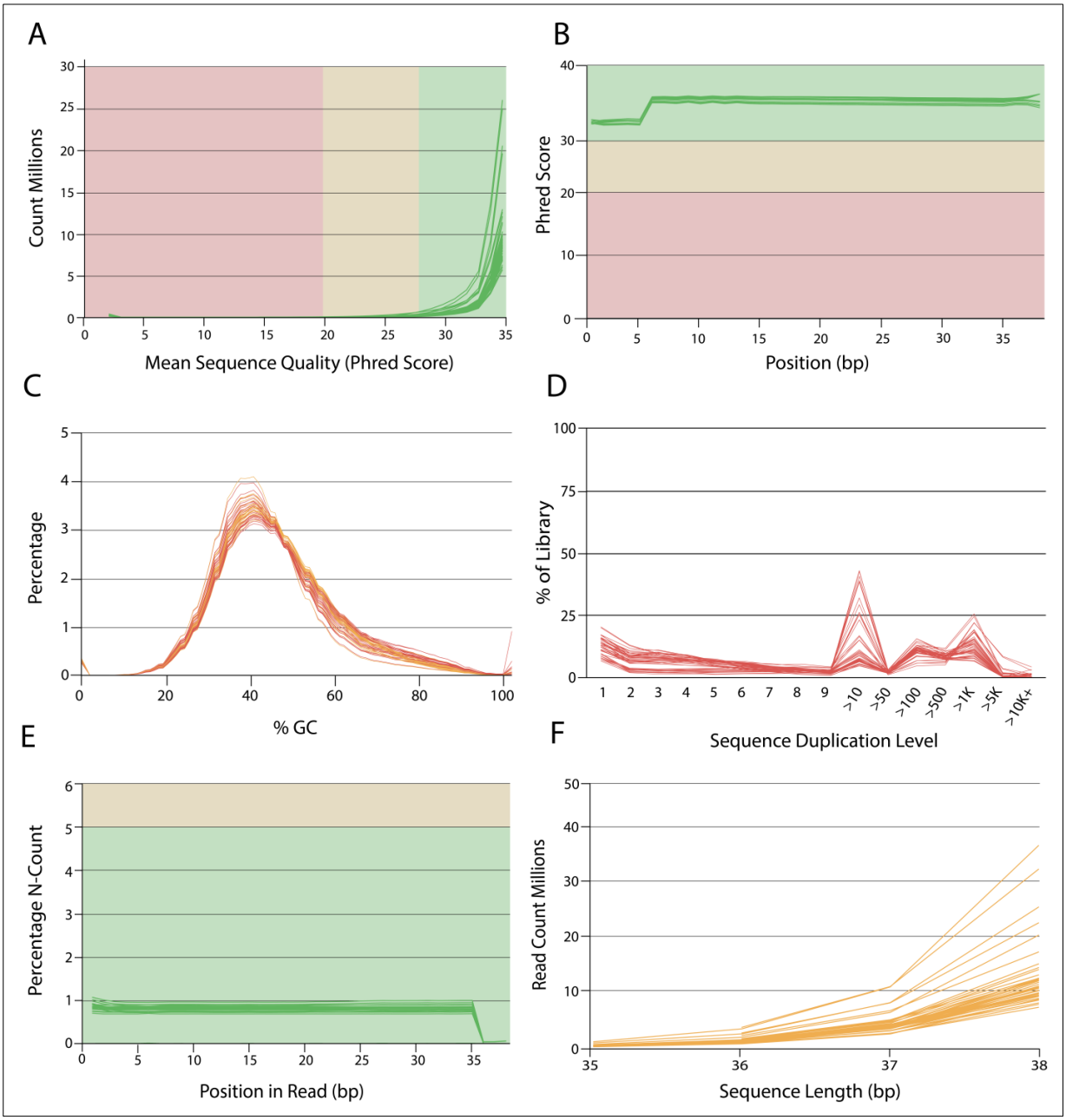


Figure 4.3: Quality control for the light pulse experiment 1 sequencing run. A: Per sequence quality score is excellent indicating good quality sequencing has been obtained; **B:** Mean quality score is excellent, showing there were no subsets of sequences of poor quality; **C:** GC content shows some bias; **D:** Overall sequence duplication level is poor; **E:** Per base N Content is excellent, indicating the sequencer successfully discerned the base at each point in the sequence; **F:** Read length distribution shows most sequences are 37/38bp as expected, read trimming has resulted in a small number of shorter sequences.

Data fragment size distribution was in line with what is expected from successful ATAC-seq library preparation, with all samples displaying a mononucleosome peak around 200bp, and a smaller dinucleosome peak near 400bp (Figure 4.4A). Duplicated reads were detected in all samples, likely due to the small amount of starting material (Figure 4.4B). However, by having four replicates for each sample condition there were sufficient unique reads to carry out a full analysis on the data, with these evenly spread across the samples (Figure 4.4B). Additionally, read mapping analysis demonstrated that almost all reads mapped to the mm10 genome with a high percentage of uniquely mapped reads (Figure 4.4C).

The range of QC metrics conducted on the ATAC-seq libraries demonstrated that they were of sufficient quality to proceed with downstream processing; although duplication was high this was not unexpected.

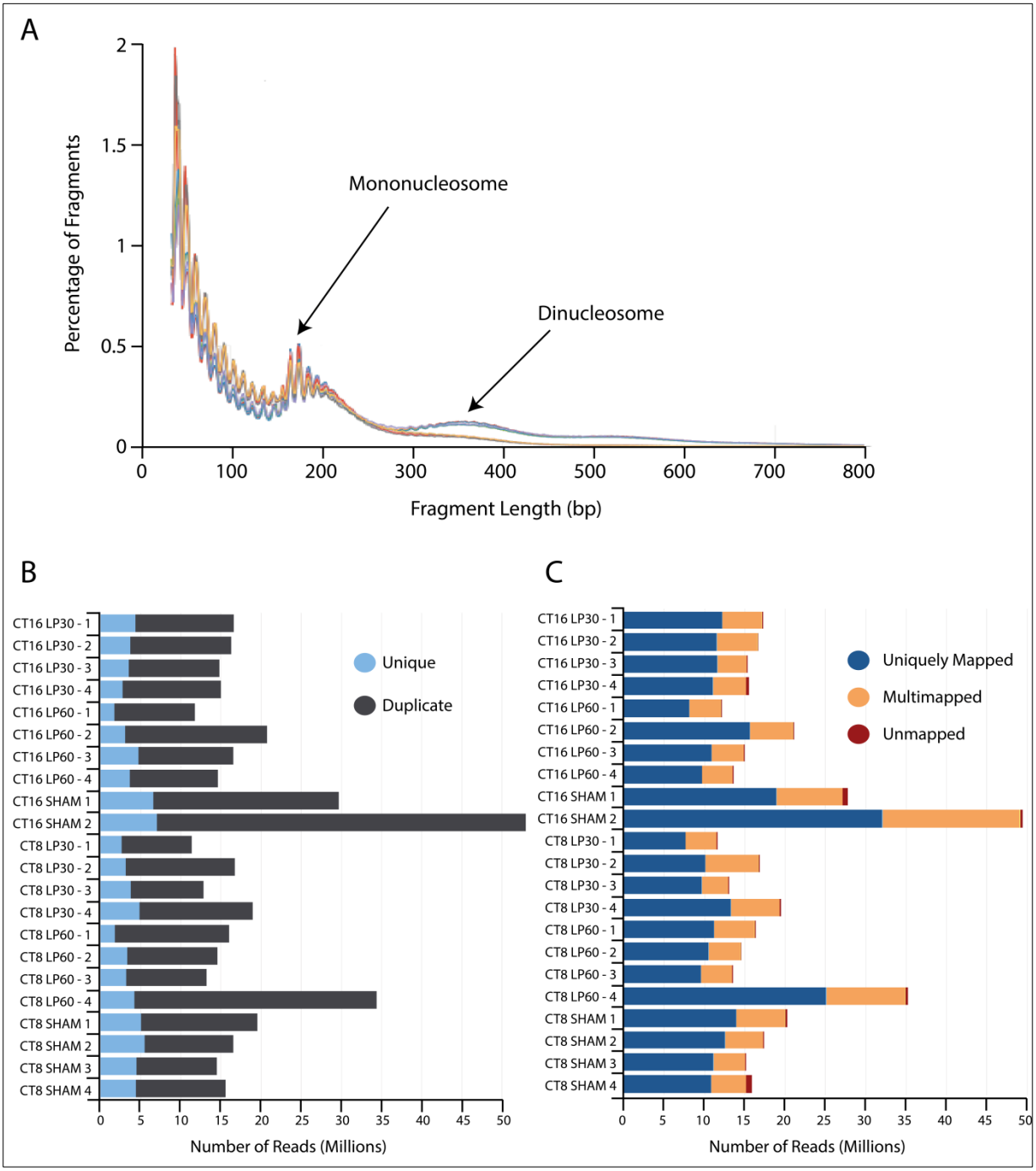


Figure 4.4: Further quality control for light pulse experiment 1. A: Fragment size distribution shows both a mono and dinucleosome peak indicating good ATAC-seq library preparation; **B:** Number of duplicates are high, but most samples still contain sufficient unique reads and each sample condition with the exception of CT16 SHAM consists of 4 replicates; **C:** Read mapping statistics to mm10 genome are excellent, with most reads mapping uniquely.

Principal component analysis (PCA) was carried out to summarize the distance matrix between samples, to record distances between each combination of samples. This helps to determine if samples cluster into their respective sample conditions and whether any outliers are present. For this experiment samples were expected to cluster by the time of day samples were taken, and if they had been light pulsed or not.

PCA revealed the samples generally clustered well, with only 3 outliers; samples taken at CT8 and the two CT16 sham samples cluster on the left, while the light pulse samples taken at CT16 clustered more on the right (Figure 4.5A -outliers highlighted in red). Following removal of the 3 outliers, based on results from the pilot experiment samples with the smallest expected change in chromatin clustered on the left, and those with the greatest expected change on the right (Figure 4.5-B).

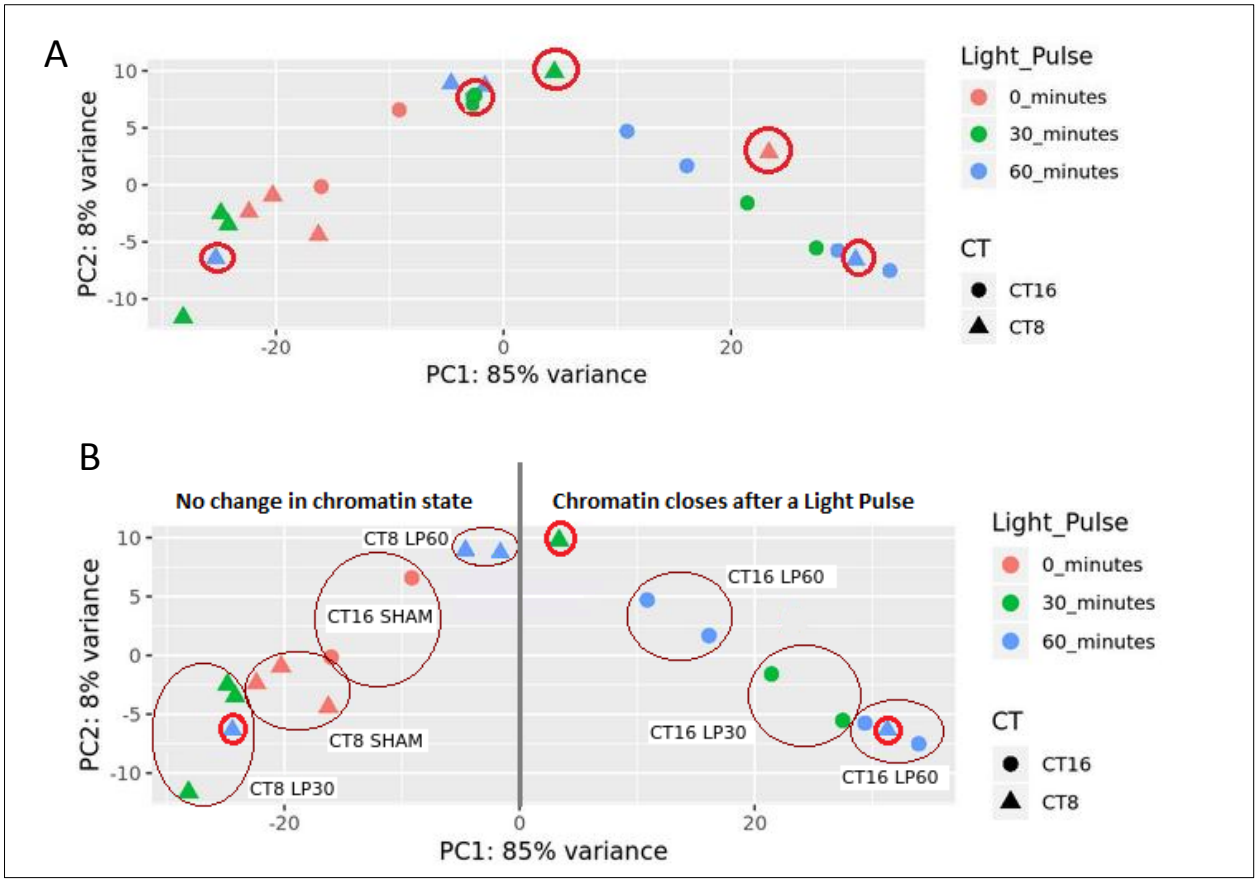


Figure 4.5: PCA plot for read counts from the first ATAC-seq light pulse experiment; A: Six outliers that do not follow the general pattern of clustering have been highlighted; **B:** Without 3 of the outliers, the samples are labelled by their patterns of clustering, and expected change in chromatin state based on the pilot experiment.

4.2.2: Differential analysis of aligned reads

Following PCA, a number of samples were identified which were highly likely to be outliers (Figure 4.5); the outliers did not cluster with the other samples by PCA and had the lowest data quality from initial QC. The outlier samples were CT16 LP30 1 and 2, and CT8 sham 4, and were removed prior to differential analysis, to more robustly identify areas of differential chromatin accessibility. Following removal of outliers, a far greater number of genes with differential chromatin nearby were identified, to verify areas were indeed differential, chromatin accessibility data was viewed visually using the IGB, and differential chromatin was clearly visible (Figure 4.8-B).

To determine chromatin state within the SCN at different times of day and following light pulsing, differential analysis of aligned reads was carried out using DeSeq2. Areas of the genome where a statistically significant greater or smaller number of reads aligned across sample conditions indicates where chromatin had changed state. A large number of genes were identified whose nearby chromatin closed following both a 30 minute and 60 minute light pulse, including several interesting candidate genes for further downstream analysis.

The greatest variation in chromatin state was observed between samples taken at CT16 that were light pulsed, and those that were not light pulsed. 466 genes were found to have statistically differential chromatin nearby (adjusted P-value < 0.05) following a 30 minute light pulse (Figure 4.6-A), and 513 genes following a 60 minute light pulse (Figure 4.6-B). Chromatin accessibility showed little difference between samples light pulsed at CT16 either for 30 or 60 min (Figure 4.6-C); this indicates chromatin accessibility changes regardless of the length of the light pulse. A statistically differential (adjusted P-value < 0.05) change in chromatin state was observed for a small but notable list of 51 genes over the course of the day shown (Figure 4.6-D), these samples were taken without a light pulse at CT8, and CT16. At CT8, very few areas of chromatin showed a statistically significant (adjusted p-value < 0.05) change following either a 30 or 60 minute light pulse (Figure 4.6-E/F). Chromatin state between samples light pulsed at CT8 and CT16 was scrutinised and revealed extensive areas differential chromatin (Figure 4.6-H/I), these are therefore the areas of chromatin which have changed state at CT16 but not at CT8.

The results obtained are in agreement with the pilot experiment which showed chromatin state only changes at CT16 following light pulsing, and very few changes take place following a light pulse at CT8. Following gene annotation of differential chromatin, results were visualised using a volcano plot (Figure 4.7), chromatin changes were identified for several interesting genes following light pulsing at CT16. Areas of differential chromatin were visualised using the IGB, chromatin near to *Csnk1e* appeared to close following a light pulse (Figure 4.8-B) at CT16, CSNK1E plays an important role in the molecular clock, and is responsible for marking PER1 for degradation, and as a result the resetting of the molecular clock (Harms et al., 2004). Chromatin close to *Nfil3* also closed following light pulsing (Figure 4.9-A), NFIL3 is part of the ancillary molecular clock (Figure 1.2), and binds D-boxes which can alter *Per1/2* expression, therefore having an effect on the core molecular clock. A further gene *MafG* had a substantial amount of chromatin which closed close-by (Figure 4.9-B), discovery of differential chromatin near to *MafG* provides a very remarkable discovery because there is little current association of MAFG with circadian rhythms and sleep, yet MAFG is able to bind both the AP-1 and CRE motifs (Kerppola and Curran, 1994).

Chromatin changes occurring between CT8 and CT16 without light pulsing were a little more difficult to discern when viewed visually using the IGB, but for a number of genes, changes had no doubt taken place. Chromatin close to *Rora* changes over the course of a day, it is closed at CT16, but open at CT8, and light pulsing has no effect (Figure 4.10). RORA has an important role in the ancillary molecular clock (Figure 1.2), and can bind the ROR motif within the promoter region of *Cry1*, therefore having an effect on the core molecular clock too.

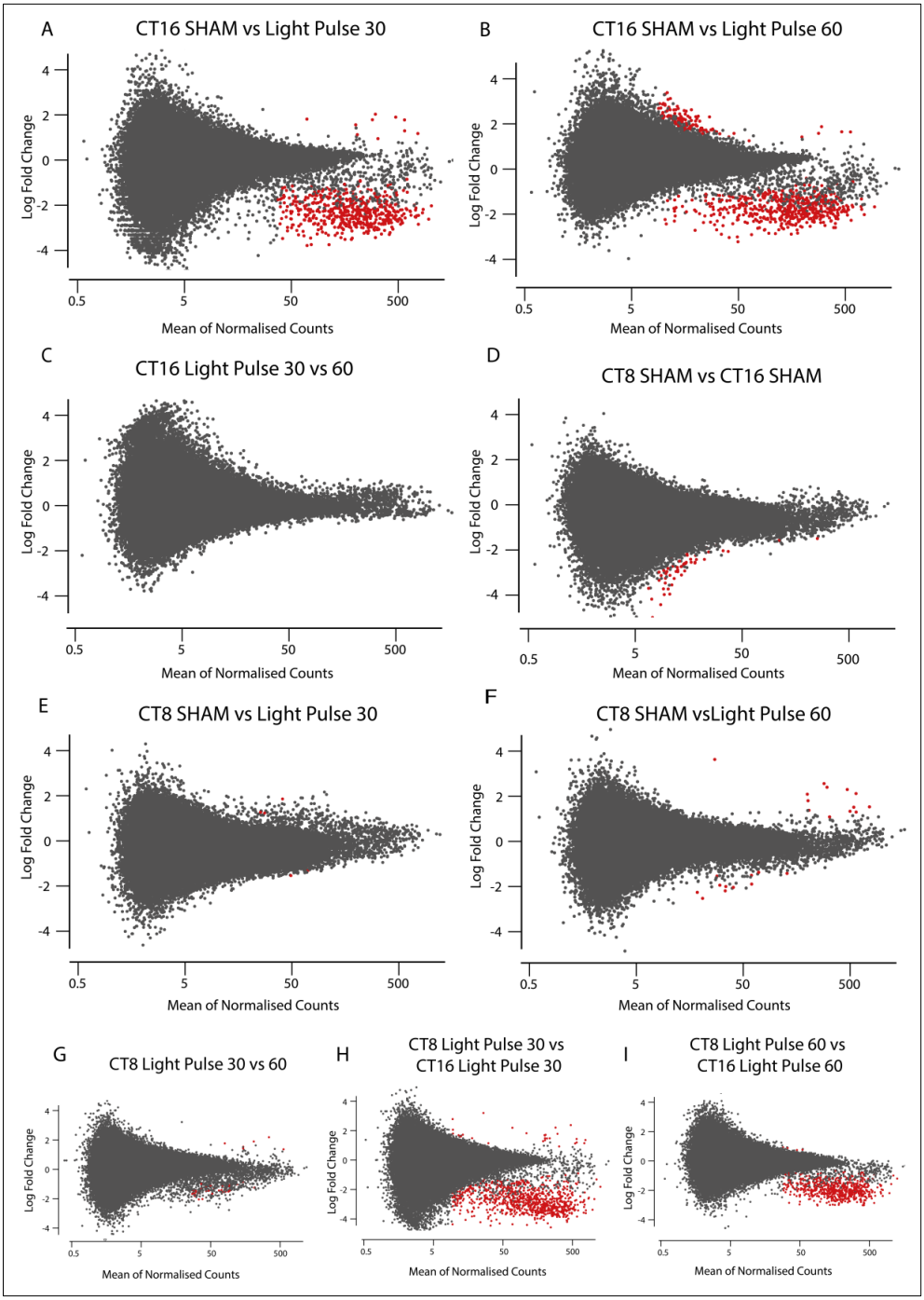


Figure 4.6: Bland–Altman plots for ATAC-seq light pulse experiment 1. Dots represent genes which have differential chromatin nearby; statistically significant genes with an adjusted P-value below 0.05 are highlighted in red. **A:** CT16 SHAM vs CT16 LP30, chromatin mainly closes following light pulsing; **B:** CT16 SHAM vs CT16 LP60, chromatin mainly closes following light pulsing, with a small amount opening; **C:** CT16 LP30 vs CT16 LP60, no statistically significant changes in chromatin state are observed; **D:** CT8 SHAM vs CT16SHAM, a small number of areas of statistically significant changes in chromatin state are observed; **E:** CT8 SHAM vs CT8 LP30, no statistically significant changes in chromatin state are observed; **F:** CT8 SHAM vs CT8 LP60, a very small number of areas of statistically significant changes in chromatin state are observed; **G:** CT8 LP30 vs CT8 LP60, very few areas of statistically significant change in chromatin state are observed; **H:** CT8 LP30 vs CT16 LP30, chromatin state is very different, showing where chromatin has changed after a 30 minute LP at CT16 but not at CT8.; **I:** CT8 LP16 vs CT16 LP60, chromatin state is very different, showing where chromatin has changed after a 60 minute LP at CT16 but not at CT8.

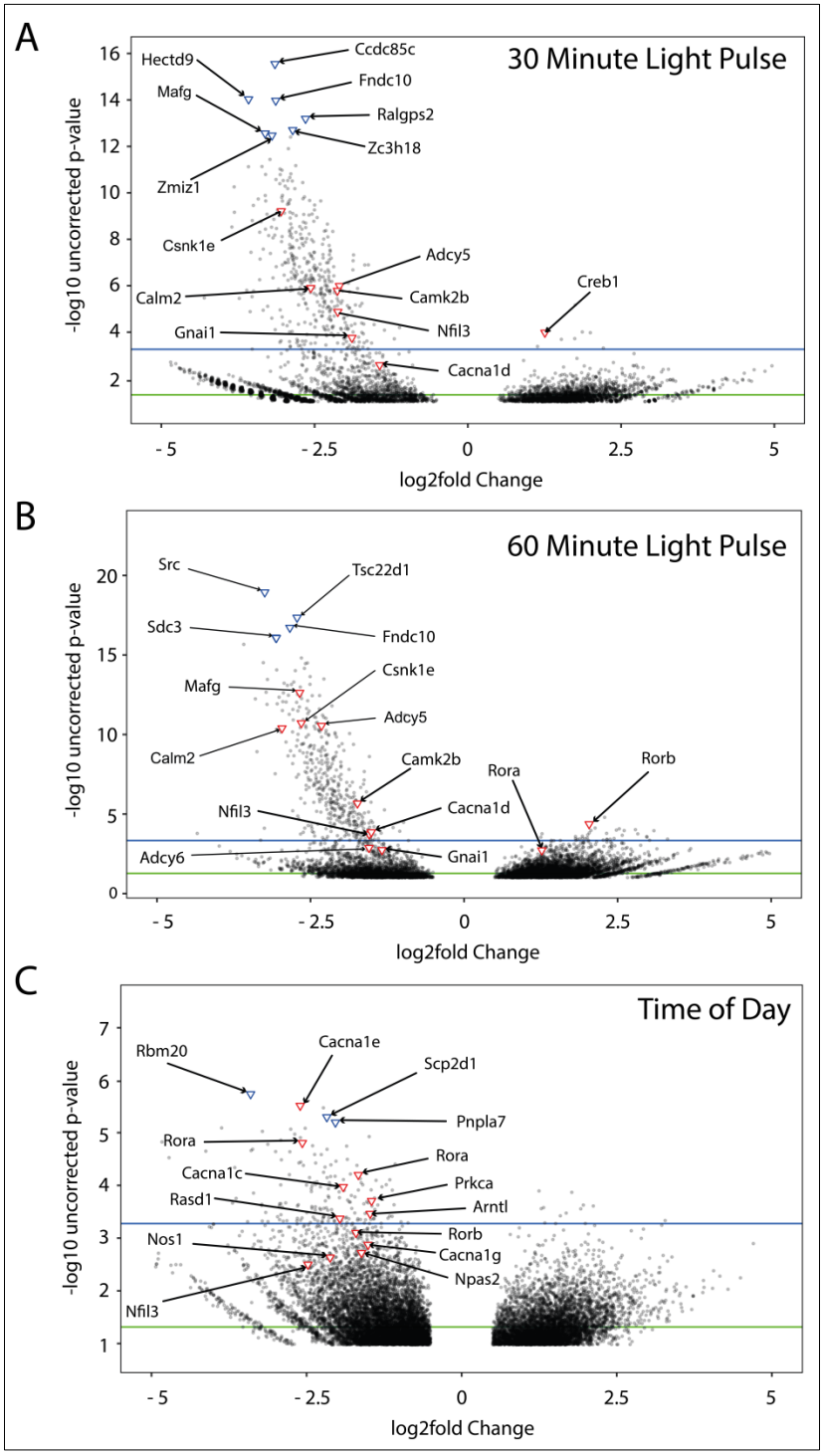


Figure 4.7: Volcano plots for ATAC-seq light pulse experiment 1. Points signify an areas of differential chromatin, blue triangles indicate the most statistically significant areas of differential chromatin, red triangles show areas of differential chromatin near to genes known to be associated with the molecular clock and entrainment, labels indicate the nearest gene to the differential chromatin, the green line is the uncorrected P-value threshold of 0.05, and the blue line is the adjusted P-value threshold of 0.05. **A:** 30 minute light pulse at CT16; a large number of statistically significant areas of differential chromatin are near genes associated with the molecular clock and entrainment, including *Csnk1e* and *Nfil3*. **B:** 60 minute light pulse at CT16; a large number of areas of differential chromatin above the unadjusted p-value threshold are near genes associated with the molecular clock and entrainment, including *Calm2* and *Adcy5*. **C:** Time of day chromatin changes; several statistically significant areas of differential chromatin are near genes associated with the molecular clock and entrainment, including *Arntl* and *Rora*.

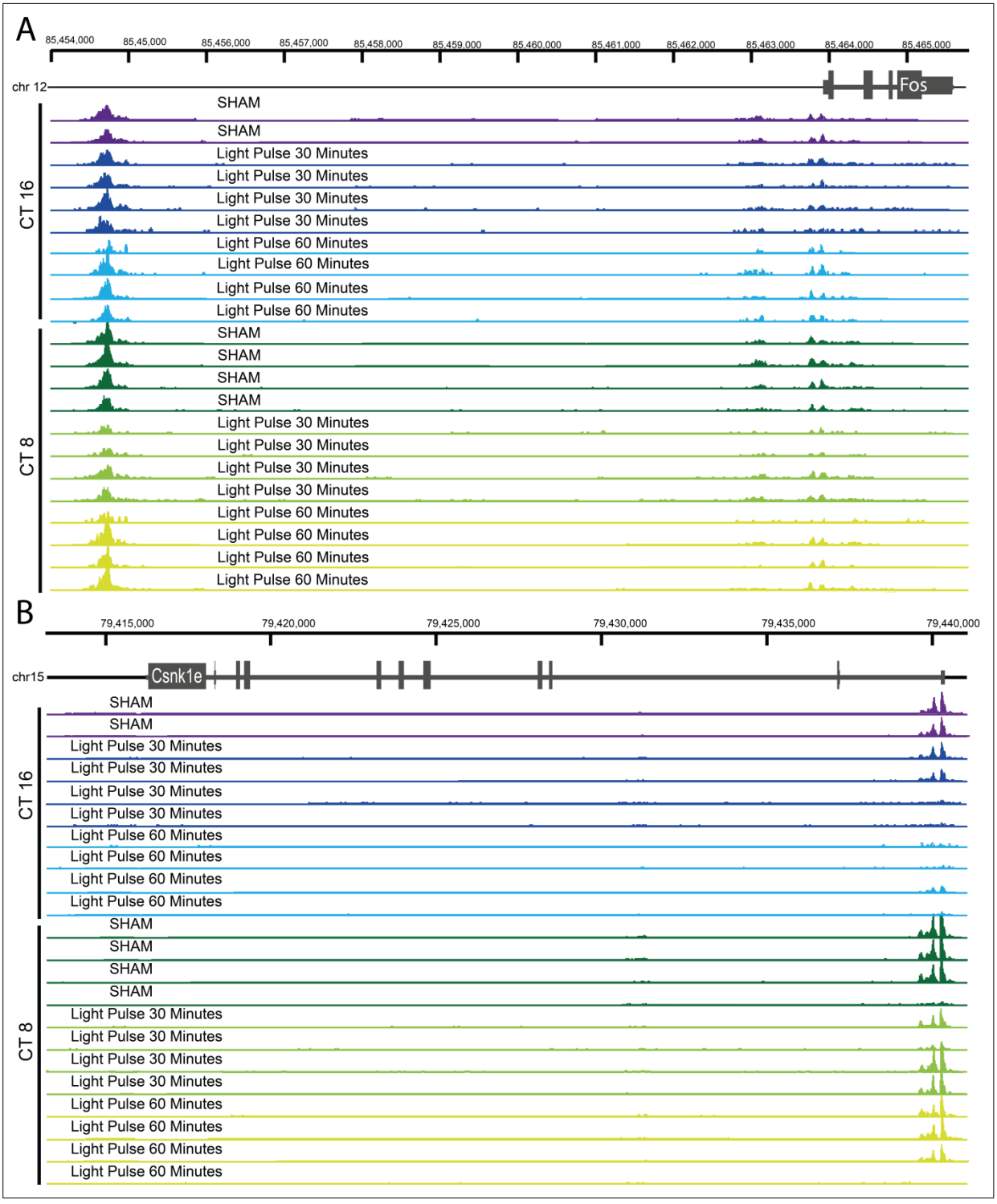


Figure 4.8: **Read coverage tracks for *Fos* and *Csnk1e* from light pulse ATAC-seq experiment 1; A:** Chromatin located near to and upstream of *Fos* remains open across all sample conditions; **B:** Chromatin near to *Cnsk1e* closes following a light pulse at CT16.

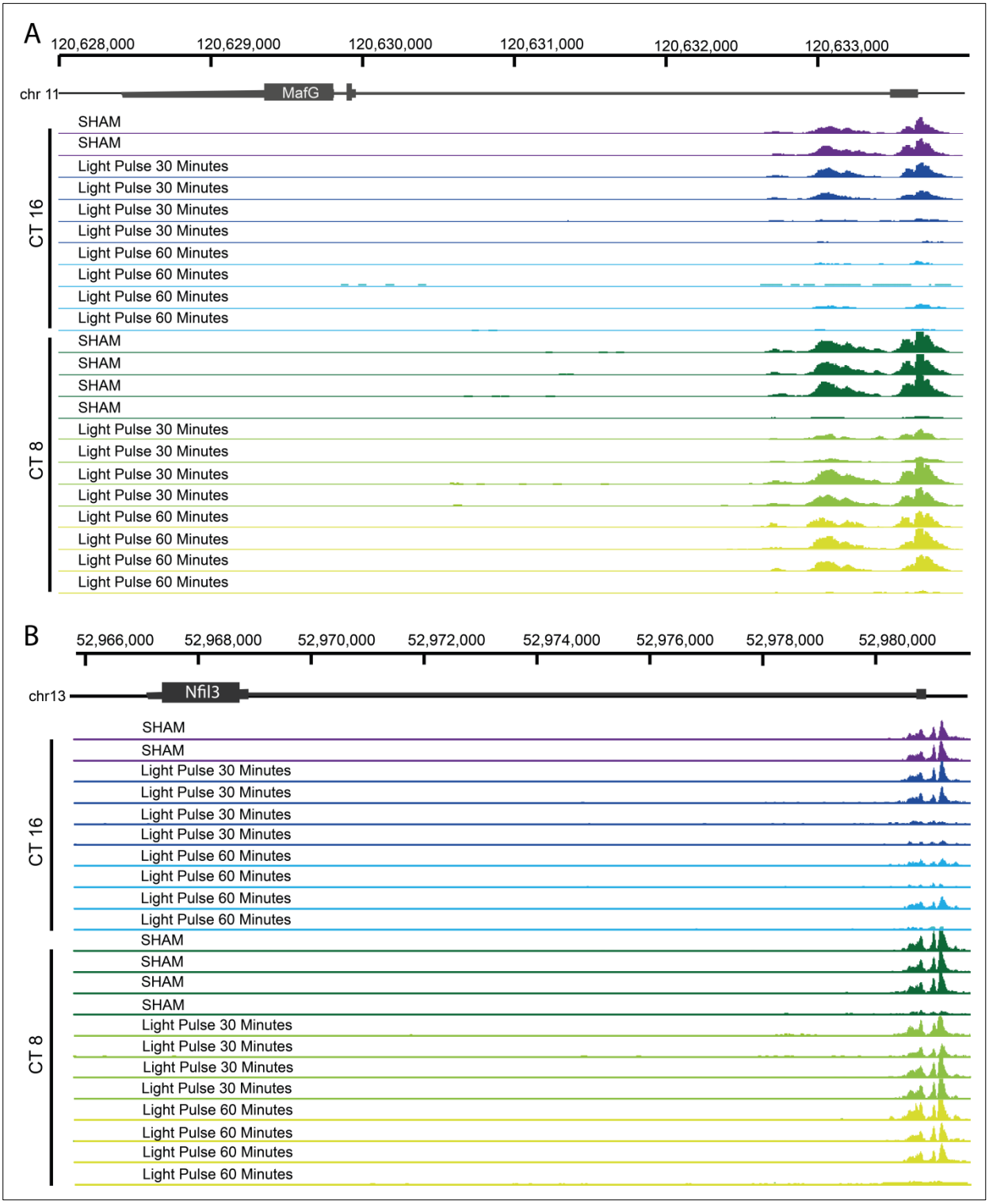


Figure 4.9: **Read coverage tracks for *MafG* and *Nfil3* from ATAC-seq light pulse experiment 1**. **A:** Chromatin near to *MafG* closes following a light pulse at CT16; **B:** Chromatin near to *Nfil3* closes following a light pulse at CT16.

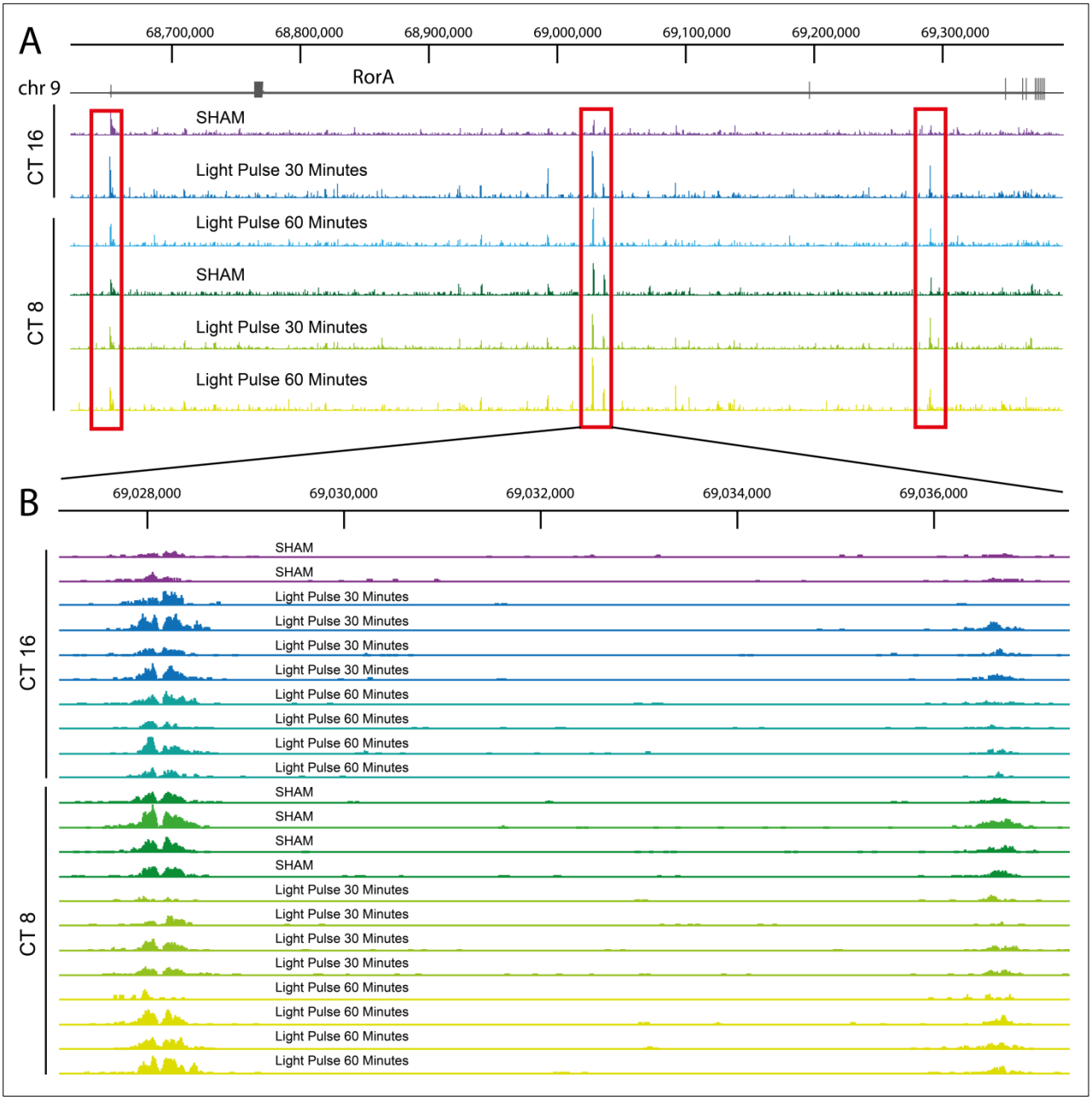


Figure 4.10: Read coverage tracks for *Rora* from ATAC-seq light pulse experiment 1; **A:** Coverage across the whole of the gene, three areas of open chromatin have been highlighted; **B:** Close-up coverage tracks for one area of open chromatin, here chromatin has closed without a light pulse at CT16, but remains open across all other samples, indicating a time of day chromatin change.

Changes in chromatin following both a 60 and 30 minute light pulse at CT16 occur near to similar genes, this is a further indicator that the experiment has identified potentially interesting genes, and gives some reassurance following removal of two samples replicates from CT16 with a 30 minute light pulse. 596 genes are found in both differential gene lists for samples subjected to a 30 or 60 minute light pulse at CT16 (Figure 4.11), a small number of genes which change following a light pulse at CT16 also change over the course of a day.

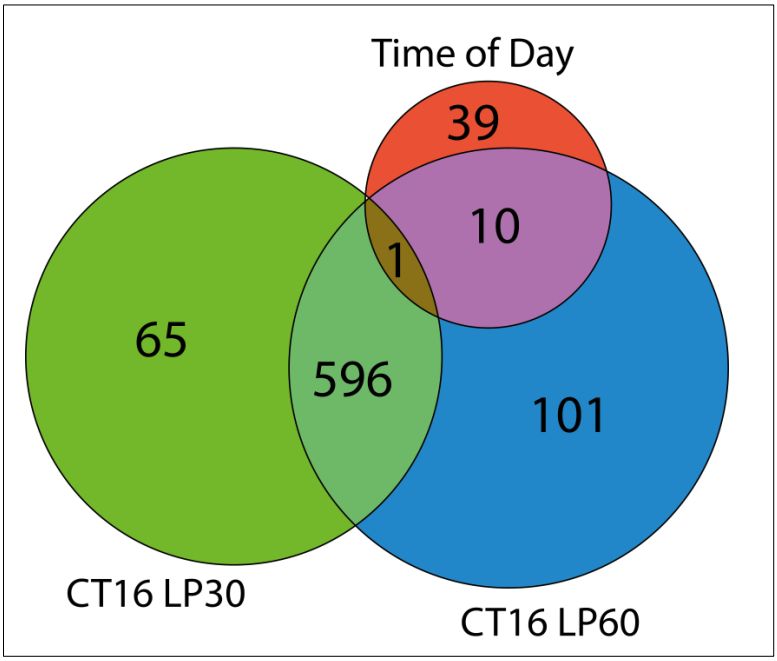


Figure 4.11: Venn plot for the number of statistical significant areas chromatin (adjusted p-value < 0.05). Areas of differential chromatin identified following a 30 and 60 minute light pulse at CT16 show the greatest similarity, with only a small overlap observed with areas of chromatin changing at different times of the day.

Differential analysis has revealed a large number of areas of statistically significant (adjusted p-value < 0.05) changes in chromatin accessibility occurring following a 30 and 60 minute light pulse at CT16, but not at CT8. A small number of areas of differential chromatin were identified to change over the course of a day; downstream analysis will now be carried out on the gene lists obtained following annotation of these areas of differential chromatin.

4.2.3: Gene and Pathway Enrichment Analysis

To determine if genes whose nearby chromatin changed following a light pulse or over the course of a day share common functions, or are involved in similar biological pathways, GO biological process (GO BP) and GO molecular function (GO MF) enrichment analysis was used. This was carried out for genes identified near to statistically significant differential chromatin at CT16 following a 30 minute (Figure 4.12-A/B) and a 60 minute light pulse (Figure 4.12-C/D), and for genes whose nearby chromatin changed across the day without light pulsing (Figure 4.12-E/F) . No analysis was carried out for samples light pulsed as CT8, because few differential genes had been identified.

Genes from the differential lists for samples light pulsed for 30 and 60 min at CT16 showed good clustering for both GO MF, and GO BF, with both containing numerous different clusters. In both light pulsed groups similar GO terms clustered, indicating a longer light pulse did not lead to a different effect. In response to light several interesting GO MF and BP terms clustered well including cAMP activity, chromatin modification, kinase activity, ubiquitination and circadian regulation of gene expression. All GO terms identified make sense in the context of our current understanding of how entrainment occurs within the SCN following light exposure; cAMP (CREB) binding is integral to entrainment through its induction of PER1. Kinases have numerous roles within entrainment, and include *Sik1* already described here, which is an important brake on phase shifting the molecular clock. Ubiquitination is essential for re-setting of the molecular clock to ensure clock proteins such as PER1 and CRY are degraded, and their differential regulation through chromatin accessibility could likely affect the period of circadian rhythms. Clustering of GO terms associated with chromatin modification indicates that genes actively involved in the modification of chromatin are themselves being modified following light pulsing, further demonstrating the importance of chromatin modification following light pulsing.

Less clustering was observed for time of day chromatin alterations, this was not unexpected because the pool of genes was much smaller, but several interesting GO terms still clustered. GO terms in these groups appeared to be more associated with transmembrane activity and ion transport,

potentially indicating that the cellular membrane changes in anticipation, or to regulate incoming stimulus such as light via the RHT at different times of the day.

Overall enrichment analysis supports the hypothesis that changes in chromatin structure are important for circadian entrainment, the analysis suggests chromatin reorganization impacts many pathways that we know are essential for light signaling and entrainment within the SCN.

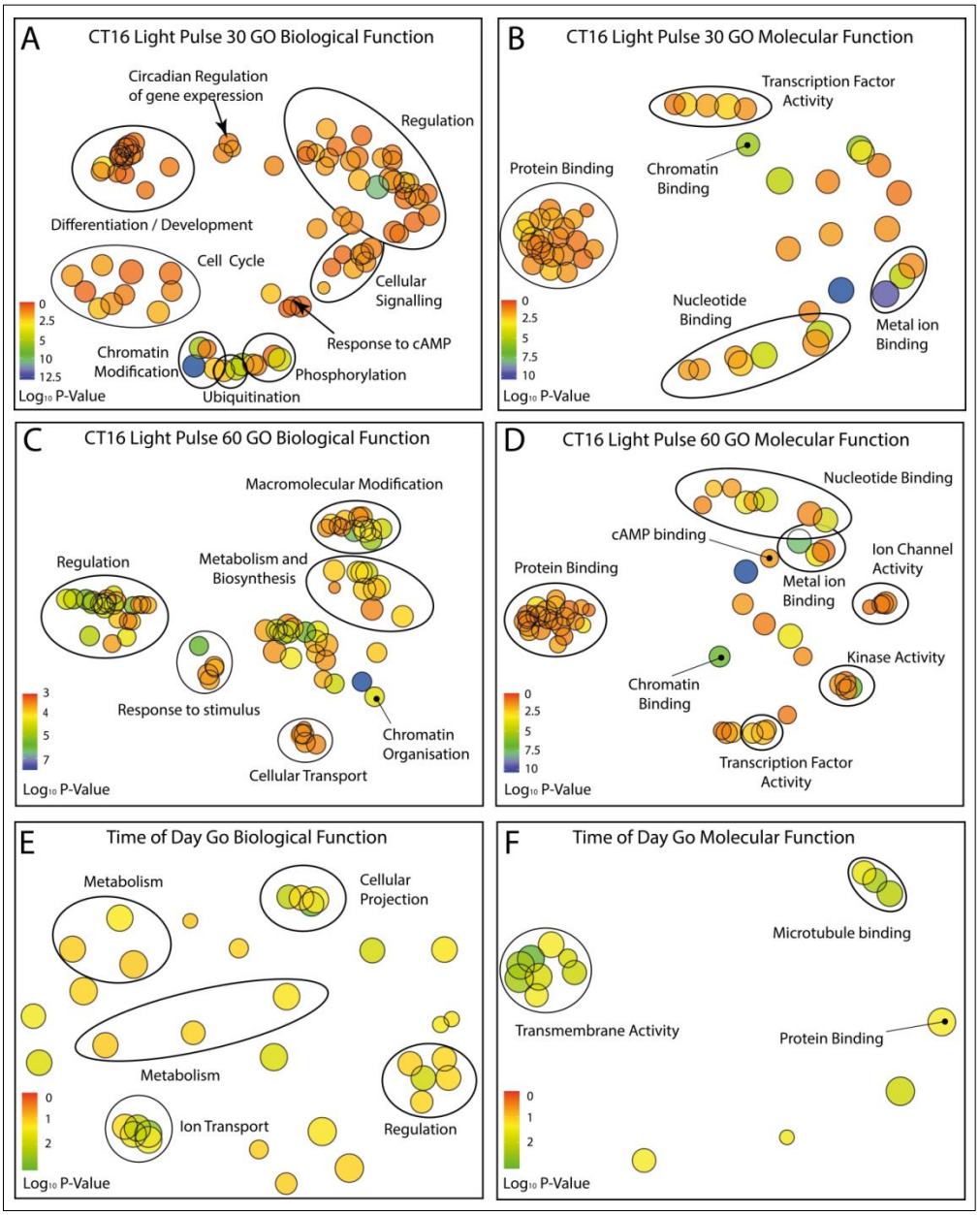


Figure 4.12: Clustering of GO BO/MF terms for ATAC-seq experiment 1 visualised using Revigo.; **A:** GO BF for samples light pulsed at CT16 for 30 min; **B:** GO MF for samples light pulsed at CT16 for 30 min; **C:** GO BF for samples light pulsed at CT16 for 60 min; **D:** GO MF for samples light pulsed at CT16 for 60 min; **E:** GO BF for samples taken with a light pulse at different times of the day; **F:** GO BF for samples taken with a light pulse at different times of the day.

4.2.4: Motif Enrichment Analysis

Differential gene functional enrichment analysis indicated GO terms associated with transcription factor activity were enriched amongst genes with differential chromatin following light pulsing at CT16. To investigate if chromatin structural changes were occurring within transcription factor binding sites, motif enrichment analysis was carried out for all genomic areas containing statistically differential chromatin (adjusted p-value < 0.05) at CT16 following light pulsing for 30 and 60 min. Motif enrichment analysis was also carried out for statistically differential areas of chromatin identified between CT8 and CT16 without light pulsing.

Following initial differential gene and pathway analysis it was expected that motifs identified from both the CT16 light pulsed samples would be very similar, but this did not turn out to be the case. Far more enriched motifs were identified from the differential chromatin list for samples light pulsed at CT16 for 60 min, rather than just 30 min. Following a 30 minute light pulse, initial known motif analysis only identified 3 enriched motifs (Table 4.2), and the majority of the DeNovo motifs identified appeared as less reliable hits (Table 4.3). In contrast, motifs enriched within differential chromatin after a 60 minute light pulse were more extensive, and more plausible in the context of a circadian experiment. Known motif results (Table 4.4), include AP-1, which has an extensive role in entrainment, and ROR, a critical component of the ancillary molecular clock (Figure 1.2). Additionally, the DeNovo motifs results (Table 4.5) give an assortment of interesting motifs, again including AP-1, but also Early growth response protein 1 (EGR), Activator protein 2 (AP-2), Zinc finger X-chromosomal (ZFX), and Nuclear respiratory factor 1 (NRF1). EGR is known to play an important role with the molecular clock, with mice lacking EGR1 having impaired BMAL1 Oscillation (Riedel et al., 2018), and AP-2 has previously been shown to be over represented in the promoter regions of CCGs (Bozek et al., 2009), but no function has been determined with relation to entrainment.

Enriched motifs were also found within chromatin changing state between CT8 and CT16 without light pulsing, enriched known motifs (Table 4.6) again include AP-1 and ROR, and several notable

DeNovo motifs were also identified (Table 4.7) with the top two by P-value being FOS and CREB, which both play an important role in entrainment within the SCN.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-2	-5.848e+00	0.43%	0.02%	NPC-Brn2
	1e-2	-5.509e+00	1.07%	0.22%	NPC-OCT6
	1e-2	-5.091e+00	0.86%	0.15%	PGR

Table 4.2: Known motifs identified within differential chromatin following a 30 minute light pulse at CT16.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-9	-2.257e+01	54.08%	39.53%	EGR2
	1e-8	-2.057e+01	0.86%	0.00%	MYB
	1e-7	-1.680e+01	10.52%	4.50%	FOXD1
	1e-7	-1.672e+01	3.86%	0.79%	ELEF3
	1e-5	-1.309e+01	3.22%	0.72%	Srf
	1e-5	-1.277e+01	5.15%	1.72%	TCF3
	1e-5	-1.250e+01	14.81%	8.42%	THAP1
	1e-4	-1.107e+01	9.87%	5.06%	Sox2
	1e-4	-1.090e+01	6.22%	2.59%	NRF1
	1e-2	-5.097e+00	13.09%	9.45%	MafA
	1e-1	-3.973e+00	1.07%	0.32%	Egr1

Table 4.3: De-novo motifs identified within differential chromatin following a 30 minute light pulse at CT16. Motifs include EGR whose TF is important within the molecular clock, and MAFA which likely binds MAFG whose nearby chromatin closed following a light pulse.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-15	-3.628e+01	4.87%	0.53%	AP1/Jun
	1e-14	-3.357e+01	6.43%	1.12%	Fra1
	1e-13	-3.101e+01	5.26%	0.80%	Fos
	1e-12	-2.982e+01	6.04%	1.13%	JunB
	1e-12	-2.940e+01	6.43%	1.31%	BAFT
	1e-12	-2.825e+01	5.65%	1.05%	Fra2
	1e-12	-2.790e+01	6.24%	1.30%	Atf3
	1e-11	-2.648e+01	6.43%	1.46%	Ap-1
	1e-9	-2.086e+01	3.51%	0.54%	Bach2
	1e-4	-1.057e+01	1.36%	0.16%	NF-E2
	1e-2	-6.617e+00	0.97%	0.15%	Bach1
	1e-2	-6.270e+00	0.97%	0.17%	ROR
	1e-2	-5.833e+00	0.97%	0.19%	Nrf2
	1e-2	-5.616e+00	11.11%	7.70%	Sox3
	1e-2	-5.599e+00	2.34%	0.93%	MafK

Table 4.4: Known motifs identified within differential chromatin following a 60 minute light pulse at CT16. Enriched motifs include AP-1 whose transcription factors are important for entrainment in the SCN and ROR an important binding motif in the ancillary molecular clock.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-17	-3.971e+01	6.82%	1.04%	AP1/JUN
	1e-11	-2.616e+01	5.26%	0.99%	MEIS2
	1e-9	-2.300e+01	0.97%	0.01%	Sox3
	1e-8	-1.998e+01	7.02%	2.20%	AP-2
	1e-8	-1.886e+01	1.17%	0.03%	HRE
	1e-7	-1.662e+01	8.19%	3.23%	Zfx
	1e-7	-1.627e+01	13.06%	6.57%	NRF1
	1e-5	-1.272e+01	0.58%	0.01%	Foxh/ KLF3
	1e-5	-1.249e+01	3.90%	1.16%	Foxj
	1e-4	-1.088e+01	1.17%	0.10%	Irf5
	1e-3	-8.272e+00	2.14%	0.58%	E2F
	1e-3	-7.794e+00	23.59%	17.66%	ZNF669/ HEK293
	1e-3	-7.483e+00	8.19%	4.77%	KLF10
	1e-3	-7.394e+00	4.87%	2.35%	EGR/SP4
	1e-3	-7.032e+00	18.71%	13.68%	Srf2

Table 4.5: De-novo motifs identified within differential chromatin following a 60 minute light pulse at CT16. Enriched motifs include AP-1 and EGR, which are both important binding sites for TFs involved with the molecular clock and entrainment.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-12	-2.828e+01	41.18%	6.06%	Atf3
	1e-12	-2.778e+01	31.37%	2.95%	Fos
	1e-11	-2.758e+01	37.25%	4.83%	Fra1
	1e-11	-2.740e+01	35.29%	4.21%	Fra2
	1e-11	-2.718e+01	27.45%	2.06%	AP1/Jun
	1e-10	-2.452e+01	35.29%	5.01%	JunB
	1e-10	-2.361e+01	37.25%	6.08%	BATF
	1e-10	-2.314e+01	39.22%	7.08%	AP-1
	1e-6	-1.583e+01	35.29%	8.69%	Sox2
	1e-6	-1.473e+01	37.25%	10.45%	Sox15
	1e-5	-1.176e+01	45.10%	18.04%	Sox3
	1e-3	-8.651e+00	37.25%	15.86%	Sox6
	1e-3	-8.167e+00	25.49%	8.58%	Sox4
	1e-3	-8.037e+00	9.80%	1.17%	ROR
	1e-3	-7.644e+00	37.25%	17.12%	Sox10

Table 4.6: Known motifs identified within differential chromatin changing across the day. Enriched motifs include AP-1 whose transcription factors are important for entrainment in the SCN and ROR an important binding motif in the ancillary molecular clock.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-12	-2.966e+01	39.22%	4.94%	Fos
	1e-10	-2.309e+01	35.29%	5.48%	CREB
	1e-9	-2.166e+01	27.45%	3.14%	ZF
	1e-8	-2.053e+01	13.73%	0.39%	E2F7
	1e-7	-1.640e+01	45.10%	13.94%	Sox2
	1e-6	-1.547e+01	7.84%	0.09%	Nr5a2
	1e-6	-1.439e+01	11.76%	0.58%	Nkx2
	1e-5	-1.194e+01	21.57%	4.16%	Sox8
	1e-4	-1.012e+01	11.76%	1.24%	EBF

Table 4.7: De-Novo motifs identified within differential chromatin changing across the day. Motifs include FOS and CREB, both important binding sites for TF that induce PER1 following light input to the SCN.

Searching for specific Motifs

Following successful identification of a number of enriched motifs within differential chromatin following light pulsing at CT16, further motif analysis was carried out to identify exactly which genes enriched motifs were located near to. Motif specific searches were carried out against the AP-1, and ZFX motifs, due to their importance in circadian rhythms and entrainment, AP-1 is important in the up regulation of *Per1* via CREB, and other zinc finger transcription factors such as ZFHX have been shown to modify circadian function (Parsons et al., 2015). This analysis intended to determine if genes containing either the AP-1 or ZFX motif shared common functions. Analysis found 41 genes with an AP-1 motif and 36 genes with a ZFX motif nearby, gene enrichment analysis did not yield any significant clustering, but several interesting genes were identified. These included dystrobrevin-binding protein 1 (*Dtnbp1),* which contained a ZFX motif within chromatin that closed following light pulsing at CT16 (Figure 4.13), *Dtnbp1 is* a very important gene relating to schizophrenia (Williams et al., 2005). Chromatin within an intronic region of *Schip1* contained several AP-1 motifs that opened following light pulsing at CT16, *Schip1* has been shown to reduce circadian period when mutated (Forni et al., 2014).

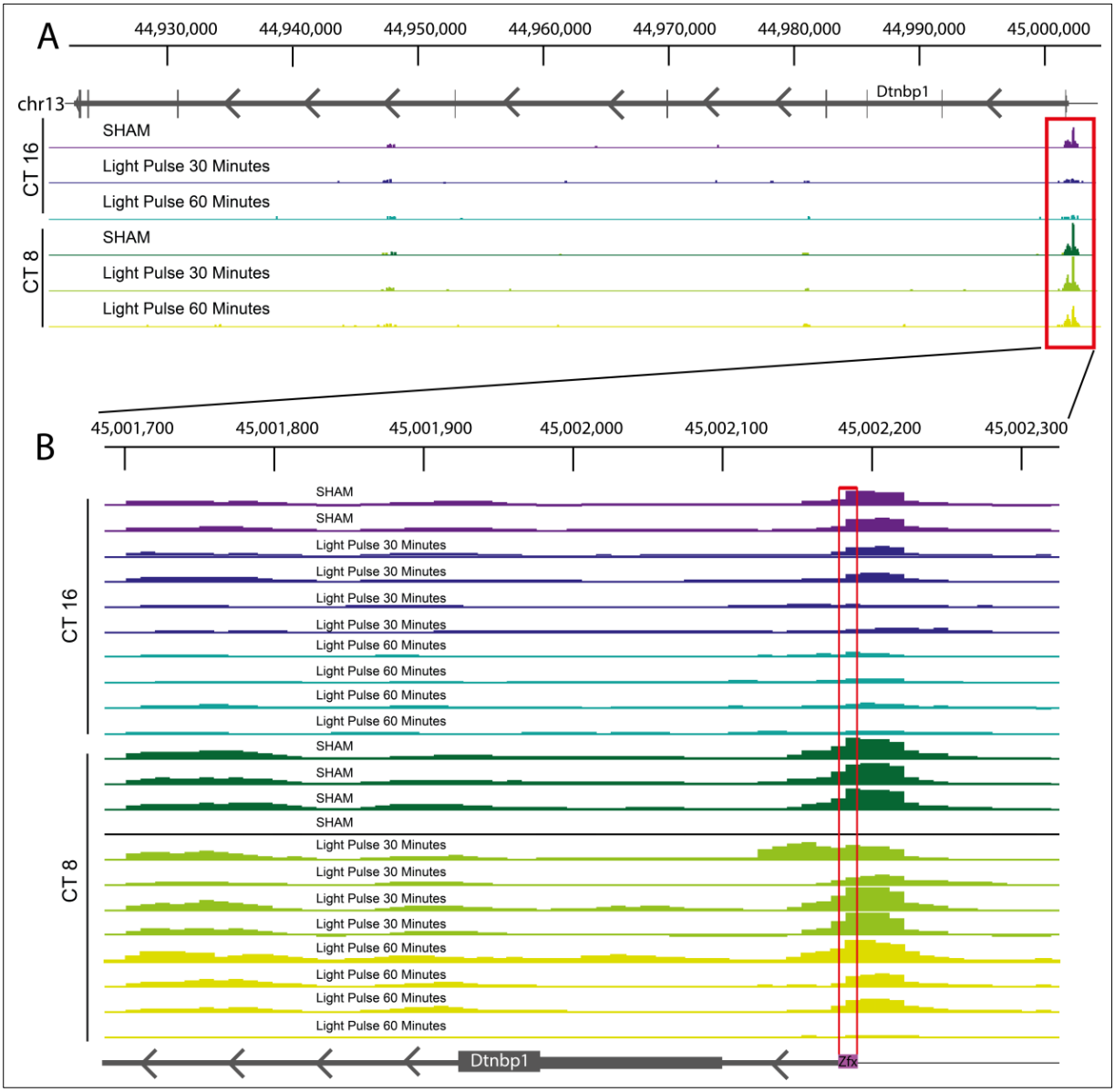


Figure 4.13: Read coverage tracks for *Dtnbp1* from light pulse ATAC-seq experiment. **A:** Chromatin closes in the promoter region at CT16 following light pulsing; **B:** Close up of the differential area of chromatin, a ZFX motif is located within this peak.

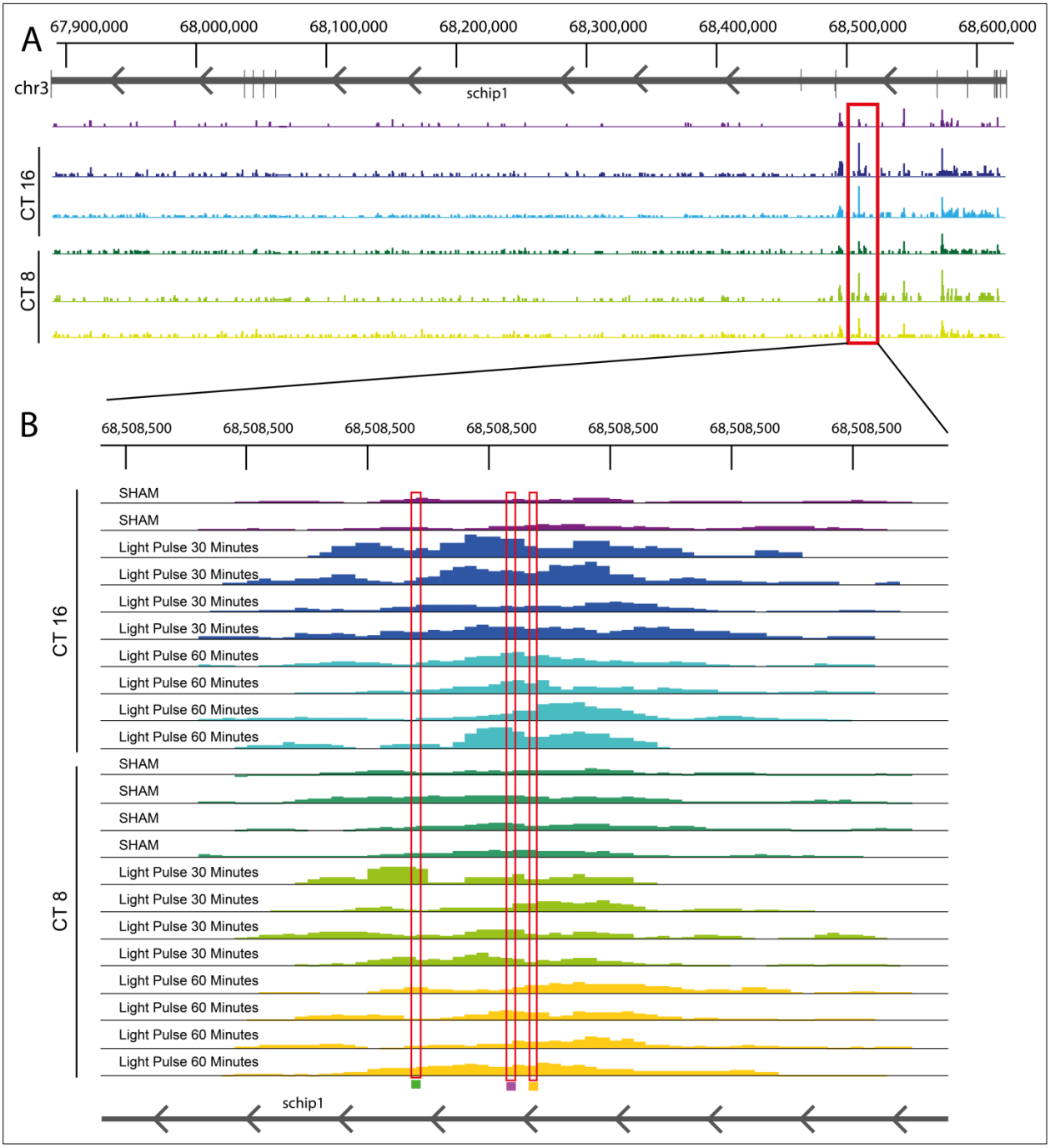


Figure 4.14: Read coverage tracks for *Schip1* from light pulse ATAC-seq experiment; **A:** Chromatin opens within an intron at CT16 following light pulsing; **B:** Close up of the differential area of chromatin , several AP-1 motifs are located within this region and have been highlighted.

4.2.5: Batch Correction with RUVseq

Although a large number of statistically significant areas of differential chromatin were found following light pulsing at CT16, in an attempt to further improve data quality, and remove unwanted variation from ATAC-seq data for subsequent differential analysis with DeSeq2, a Bioconductor package called Remove unwanted variation seq (RUV-seq; Risso et al., 2014) was investigated. Normalization methods employed by default when carrying out differential analysis with DeSeq2, take account of differences in sequencing depth, but do not account for other technical anomalies, such as those occurring during library preparation. RUV-seq employs factor analysis using a suitable set of control genes, and aims to give far better estimates of expression fold-changes. It was anticipated that by using RUV-seq, DeSeq2 would give enhanced estimates about differential expression and accessibility.

SCN light response data from Jagannath et al. (2013) and time of day RNA-seq SCN expression data from Pembroke et al. (2015) were used to create three housekeeping control gene lists. The first contained genes not induced or suppressed by light in the SCN, the second genes that do not fluctuate in their expression over the course of a day, and a third that was an amalgamation of the previous two.

ATAC-seq data was analyzed using the bioinformatic methods outlined in section 2.11 until a counts table was generated; the counts table was then corrected using RUVseq as outlined in section 2.12. Two approaches were carried out to determine the best method to normalize the data; the raw counts table from Rsubread was first used as input for RUVseq without any prior normalization, to test how effective the program works exclusively by itself. Following this a hybrid method was employed; data was normalized for read depth using DeSeq2 to generate a new counts table normalized for read depth, this counts table was then used as input for RUVseq.

Results from the correction of data not already normalized for read depth appeared positive; prior to normalization (Figure 4.15-A) PCA revealed very little clustering could be observed between samples. Following data processing using light or time of day housekeeping genes data quality improved, and following PCA samples separated by time of day, but clustering remained poor (Figures 4.15-A/B). Further normalization using both sets of housekeeping genes together gave the

greatest improvement in data quality, following PCA samples now generally clustered by both time of day, and whether they had been light pulsed or not (Figure 4.15-D).

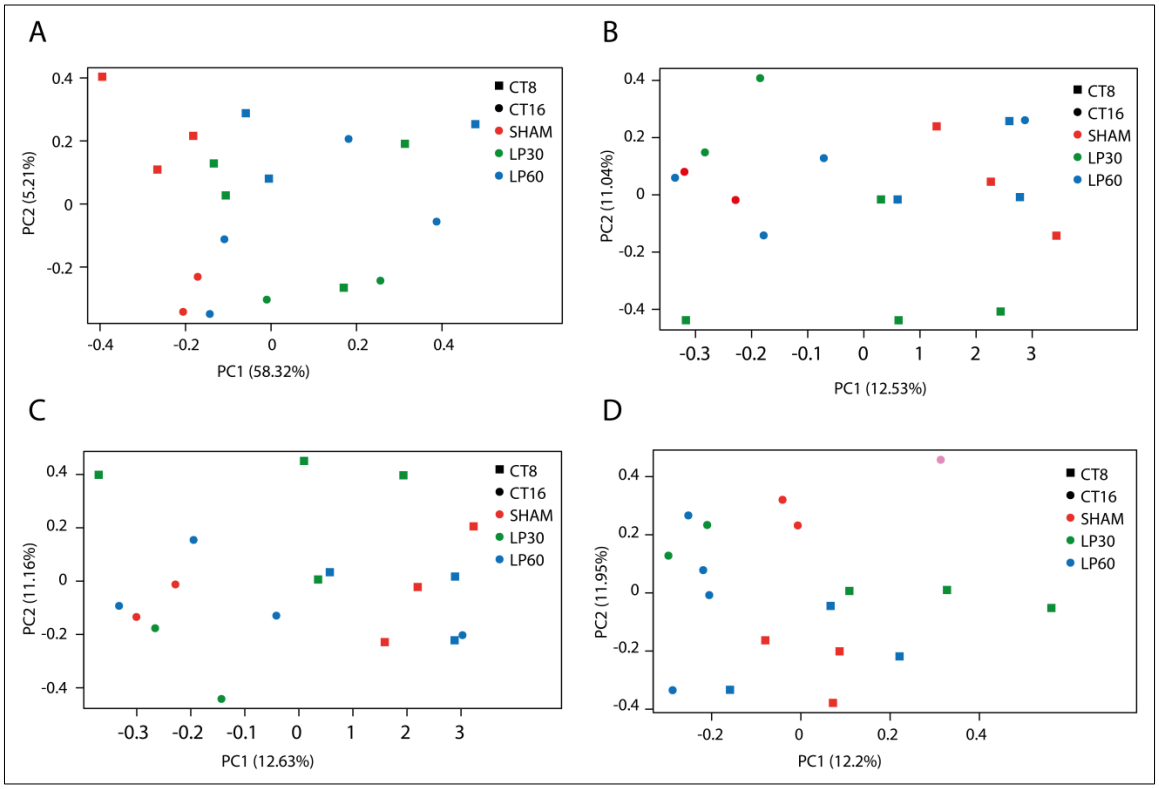


Figure 4.15: RUV-seq normalization using non-normalized data as input; A: PCA plot before normalization shows little clustering between samples; **B:** PCA plot following normalization with light housekeeping gene counts, most samples separate by time of day, but clustering remains poor overall; **C:** PCA plot following normalization with time of day housekeeping gene counts, most samples separate by time of day, but clustering remains poor overall; **D:** PCA plot following normalization with both light and time of day housekeeping genes, most sample now cluster loosely by the time of the day and by treatment.

PCA of data normalized for read depth by DESeq2 prior to further correction with RUVseq revealed tight clustering by time of day, but with poor separation (Figure 4.16-A). This pattern of clustering is less robust than previously observed by PCA (Figure 4.5), and is likely due to data loss when the data was converted to an integer matrix prior to further correction with RUVseq. Further normalization using RUVseq, independently with either light or time of day housekeeping genes (Figures 4.16-B/C) did not improve data quality any further, and normalization with both housekeeping genes together appeared to reduce data quality (Figure 4.16-D), leading to a loss of clustering following PCA.

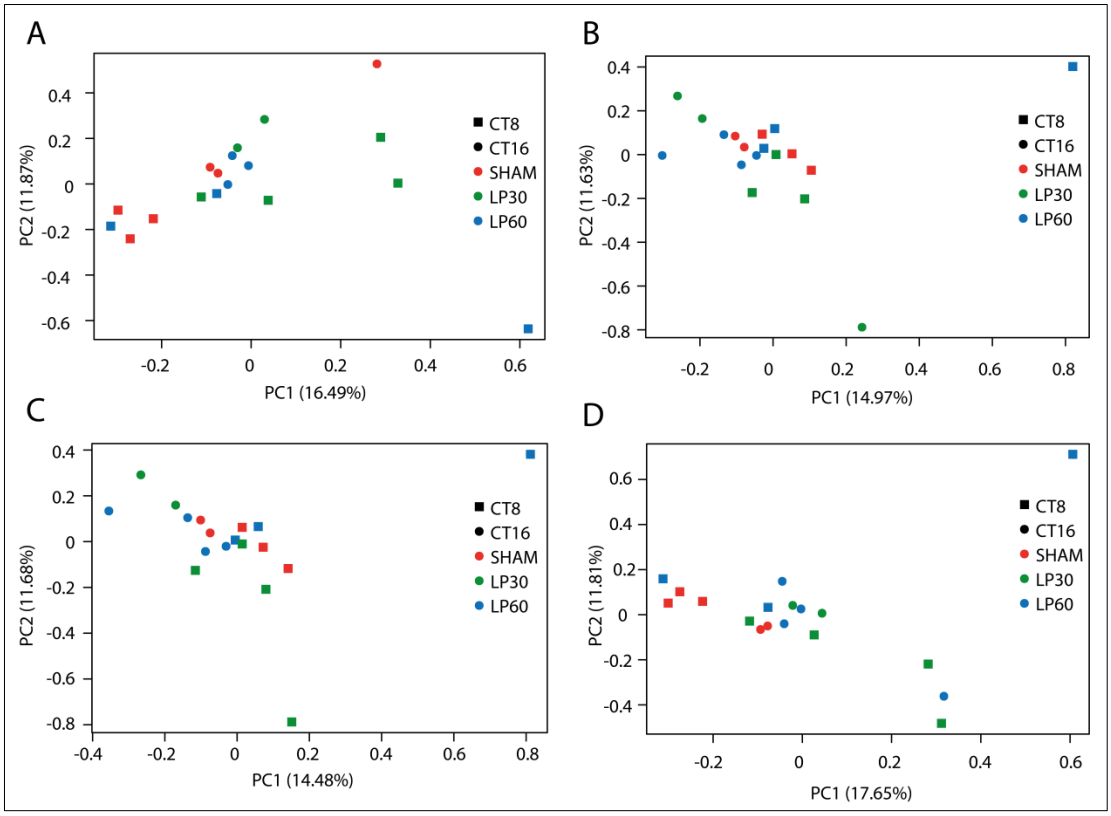


Figure 4.16: RUV-seq normalization using DESeq2 normalized data as input; A: PCA plot before normalization shows some clustering by time of day; **B:** PCA plot following normalization with light housekeeping gene counts, most samples cluster by time of day, but separation between samples is poor overall; **C:** PCA plot following normalization with time of day housekeeping gene counts, most samples cluster by time of day, but separation between samples is poor overall; **D:** PCA plot following normalization with both light and time of day housekeeping genes, very little clustering can now be observed.

Overall RUV-seq did not produce a significant improvement in data quality; when raw data counts were input to RUVseq, and corrected using both light and time of day housekeeping gene together data quality did improve slightly (Figure 4.15-D). However, when data previously normalized for sequencing depth with DESeq2 was further normalized using RUVseq, none of the housekeeping gene sets were able to improve data quality any further, in fact when both housekeeping gene sets were used together data quality appeared to diminish (Figure 4.16-D). Normalized counts from DESeq2 had to be converted to an integer matrix before they could be input to RUVseq; this resulted in a loss of data, and likely contributed to the reduction in data quality.

In conclusion, to test improvements in data quality using only RUV-seq without prior normalization, data corrected using both housekeeping gene sets together was used for differential analysis with DESeq2. Following analysis this did not result in a greater number of differential genes being obtained, this is likely because DESeq2 will have still applied its own normalization prior to differential analysis. Batch correction using RUVseq has clearly demonstrated that there are limits on how far data quality can be further improved using bioinformatic techniques, and combining two or more different techniques together does not necessarily result in better data quality. This demonstrates how data and its subsequent analysis to a large part are ultimately constrained by the initial data quality obtained when samples are sequenced.

4.2.6: Comparison of genes with differential chromatin to published micro-array data

In order to understand how chromatin structure and gene expression in the SCN following light pulsing are linked, genes identified near to differential chromatin following light pulsing at CT16 were compared to published micro-array data from Jagannath et al. (2013), where gene expression in the SCN was investigated following a light pulse at CT16 for varying lengths of time. The aim of this comparison was to determine the impact of changes in chromatin accessibility on downstream gene expression. As the majority of genes with differential chromatin showed reduced accessibility in their promoter regions, I hypothesise that this will result in gene expression being switched off for these genes because TFs will be unable to bind to initiate transcription.

From the microarray data, genes shown to change their expression following a 60 minute light pulse were compared to genes whose nearby chromatin showed a statistically significant change in chromatin state following a 60 minute light pulse. Comparison of the two lists found 22 genes present in both (Figure 4.17), a fisher exact test demonstrated that these 22 genes were unlikely to appear in both lists by chance (Table 4.8).

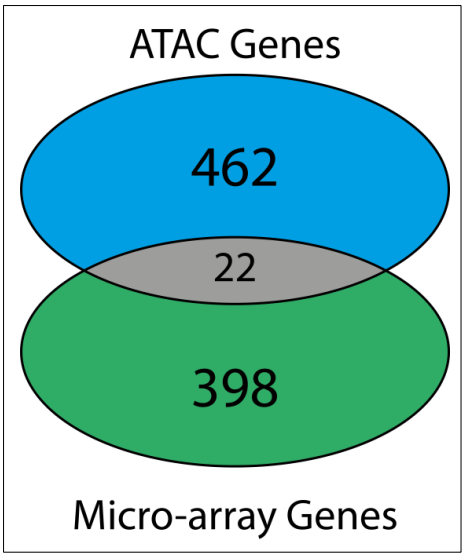


Figure 4.17: Venn diagram for comparison of microarray and light pulse experiment 1data. Genes identified from ATAC-seq to have differential chromatin nearby following a 60 minute light pulse at CT16, were compared to genes identified from the microarray data to be DE following a 60 minute light pulse at CT16; 22 genes appeared in both lists.

This is therefore a good indication that chromatin accessibility is having some effect on gene transcription. However, although 22 is statistically significant, this was a lower number than anticipated, and demonstrates the complexity of transcriptional regulation and indicates that many genes may have accessible chromatin, but that chromatin may already be bound by trans-acting regulatory factors, which prevent transcription.

	Non differential Micro-array genes	**Differential Micro-array genes**
Non differential ATAC genes	24,000	398
Differential ATAC genes	462	22

Overlapping P-Value = 3e-05 *Odds Ratio* = 2.8

Table 4.8: Contingency table for fisher exact test to compare the transcriptome with chromatin accessibility after light pulsing. The number of genes found in the gene lists for genes with differential chromatin accessibility, and genes with differential gene expression, following a 30 minute light pulse at CT16.

4.2.7: Comparison of chromatin accessibility with publicly available data for epigenetic marks

To evaluate the link between chromatin accessibility and other epigenetic markers/modifications, areas of differential chromatin were compared to chromatin immunoprecipitation sequencing (CHIP-seq) coverage tracks for common epigenetic marks. The methylation of DNA or modification of histone proteins will have a direct effect on chromatin structure; therefore areas of the genome identified by ATACs-seq to alter in chromatin state were anticipated to contain epigenetic markers which have enabled these changes to take place.

Data was taken from the Encode consortium for mouse mid-brain CHIP-seq (Davis et al., 2018), data included marks for histone methylation protein H3K4me3 (Encode ID: ENCSR203KIB), and histone acetylation proteins H3K9ac (Encode ID: ENCSR286OKN) and H3K27 (Encode ID: ENCSR088UKA). Epigenetic marks were found to be abundantly present by ATAC-seq where light pulsing has been shown to change chromatin state, including near to *MafG* and *Csnk1e* following a light pulse at CT16 (Figure 4.18-A/B). In contrast chromatin close to *Fos* where chromatin did not change under any sample condition investigated by ATAC-seq showed less abundant epigenetic marks (Figure 4.18-C).

This clearly demonstrates that areas of the genome where chromatin accessibility is dynamic and changes state following external stimuli such as a LP, are frequently bound by epigenetic marks such as histone acetylation proteins likely to facilitate this change.

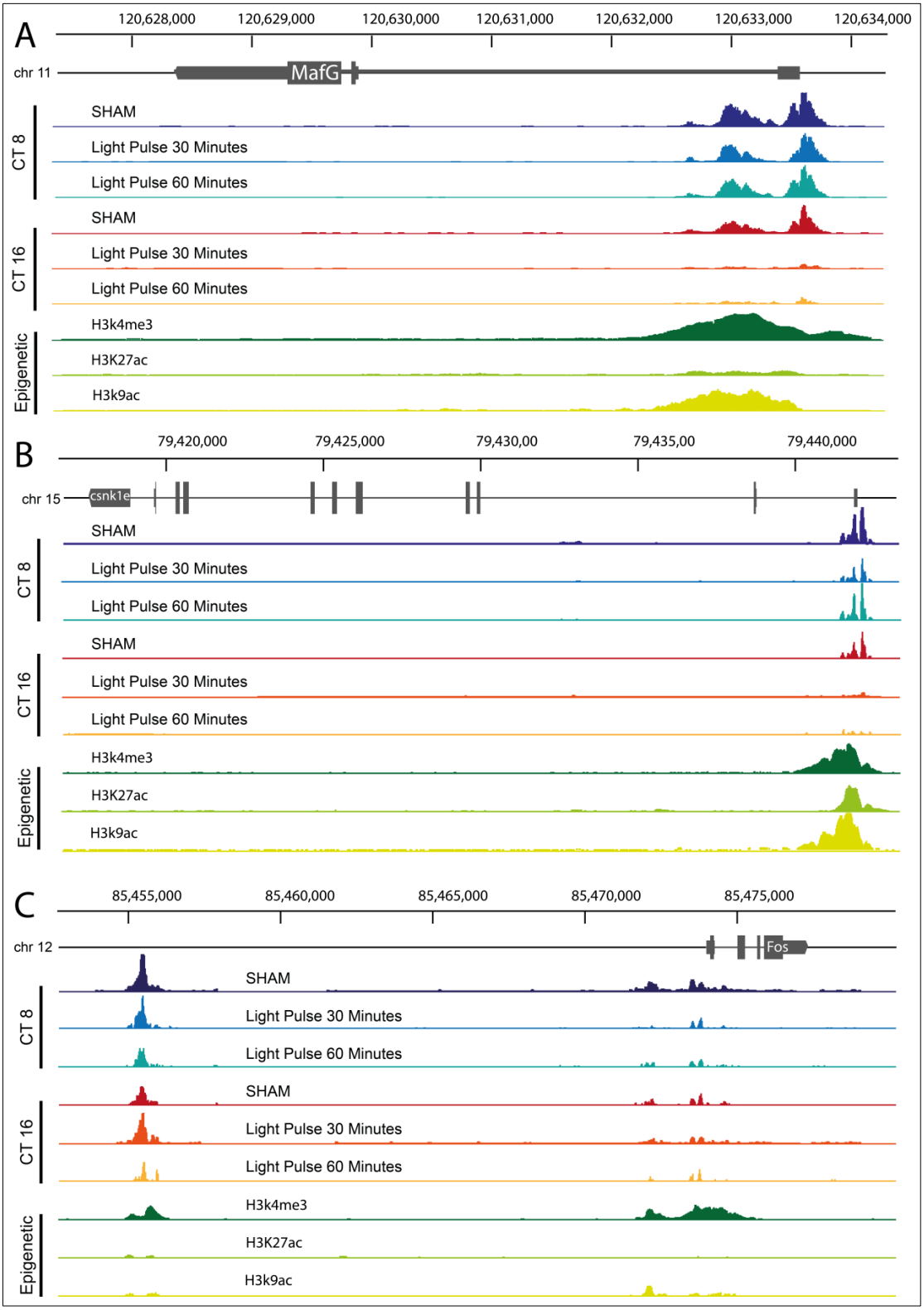


Figure 4.18: Comparison of ATAC-seq read coverage with epigenetic CHIP-seq read coverage. Read coverage tracks for ATAC-seq light pulse experiment 1 along with chip-seq coverage tracks for epigenetic marks; H3k4me3, H327ac and H3k9ac, ATAC-seq and epigenetic tracks are scaled separately. **A:** Epigenetic marks are located where chromatin near to *MafG* closes following a light pulse at CT16; **B:** Epigenetic marks are located where Csnk1e closes following a light pulse at CT16; **C:** Epigenetic marks are present where chromatin is open near to *Fos* but far less numerous.

4.3: ATAC-seq Light Pulse Experiment 2

Following the first successful light pulse ATAC-seq experiment, a further repeat was carried out to validate the results obtained, with the experiment being carried out in the same manner as the first. Furthermore, in an effort to further improve the data quality several minor modifications were made to the nuclei isolation protocol.

Firstly, a pluriStrainer Mini 40 µm (pluriSelect cat no. 43-10040-40) allowed direct nuclei straining into a 1.5ml Eppendorf tube (as opposed to the larger 40 µm cell strainer, which was placed in a 50ml falcon tube), and aimed to reduce loss of material when samples were transferred from the falcon tube to the Eppendorf tube. Secondly, following the first nuclei centrifugation, 50µl of supernatant was left behind as well as the nuclei pellet, and again aimed to reduce any loss of material.

Mice were light pulsed as detailed in section 2.6, but with only a 60 minute light pulse. It was determined that similar genes showed differential chromatin accessibility following light pulsing, and no genes showing differential chromatin accessibility after a 30 minute light pulse subsequently reverted back to the control state during the addition 30 min of light pulsing. Furthermore, motif analysis revealed a greater number of interesting motif binding sites changed chromatin state following 60 min of light pulse, rather than just 30 min. Each replicate again comprised 4 SCN, so each sample condition remained representative of 16 mice.

4.3.1: Quality Control

After obtaining sequencing data, all samples were again subjected to rigorous quality control checks, to understand overall sample quality. These checks in line with the first experiment included sequencing quality assessment, checks to determine the standard of ATAC-seq library preparation robustness of alignments, and initial indication of overall data quality. For this experiment data quality appeared to have improved and looked excellent across all samples, read numbers were spread evenly across all samples (Figure 4.19), giving an immediate indication that the samples were diluted correctly and of good quality.

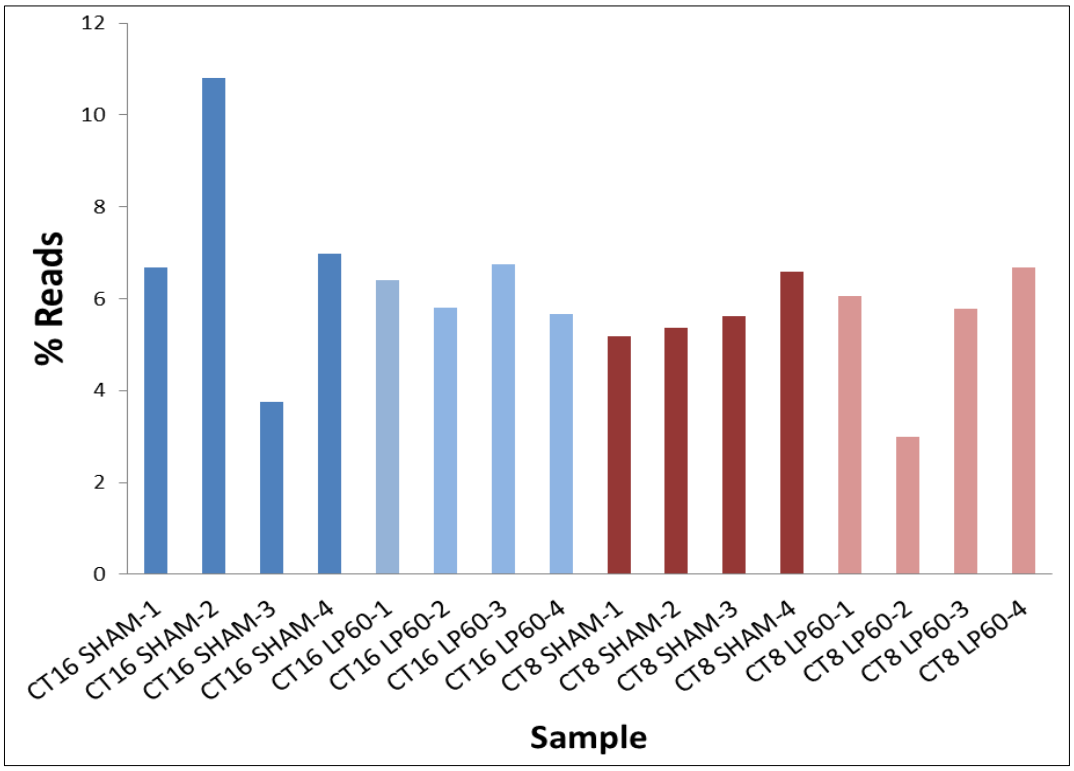


Figure 4.19: Read distribution across the different ATAC-seq different samples for light pulse experiment 2. An even distribution can be seen, indicating good library preparation, and even dilution of samples prior to sequencing, all samples sequenced correctly.

Tests carried out by FastQC to determine sequencing quality indicated data to again be of high quality, with GC content improving compared to the first run (Figure 4.20-C) and sequencing duplication still present (Figure 4.20-D), but at much lower levels than in the previous experiment.

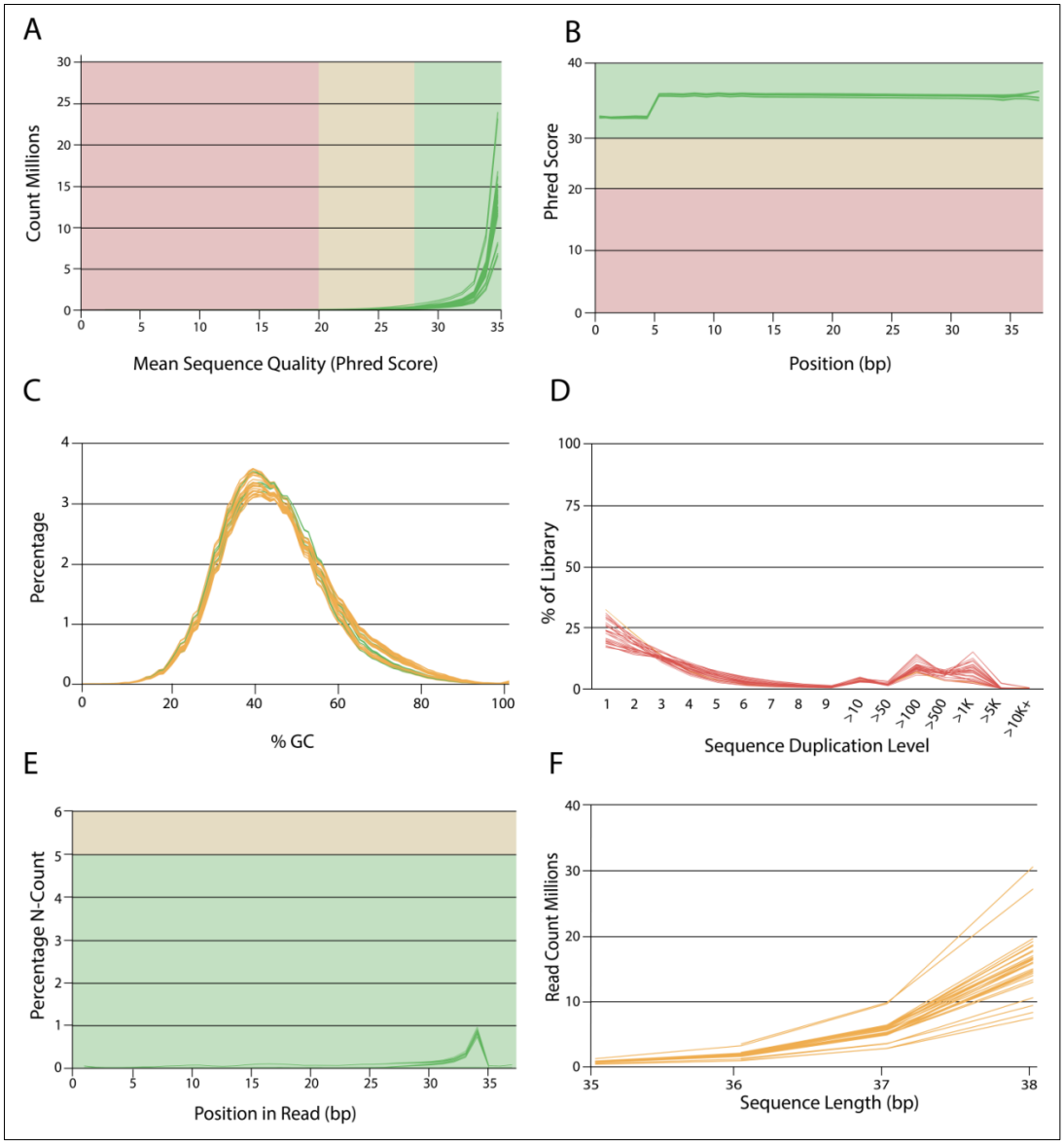


Figure 4.20: Quality control for the light pulse experiment 2 sequencing run. A: Per sequence quality score is excellent indicating good quality sequencing has been obtained; **B:** Mean quality score is excellent, showing there were no subsets of sequences of poor quality; **C:** GC content has improved from the first ATAC-seq experiment; **D:** Overall sequence duplication remain high, but has improved over the first ATAC-seq experiment; **E:** Per base N Content is excellent, indicating the sequencer successfully discerned the base at each point in the sequence; **F:** Read length distribution shows most sequences are 37/38bp as expected, read trimming has resulted in a small number of shorter sequences.

Data fragment size distribution was again in line with what is expected from successful ATAC-seq library generation (Figure 4.21-A). All samples had a mononucleosome peak around 200bp, and a smaller dinucleosome peak near 400bp. Whilst clearly still present, the number of duplicates was significantly lower, with some libraries now containing only 50% duplicates, indicating that more reads of higher quality were available for alignment and subsequent differential analysis (Figure 4.21-B. Almost all reads mapped to the mm10 genome with most libraries having around 15M uniquely mapped reads (Figure 4.21-C).

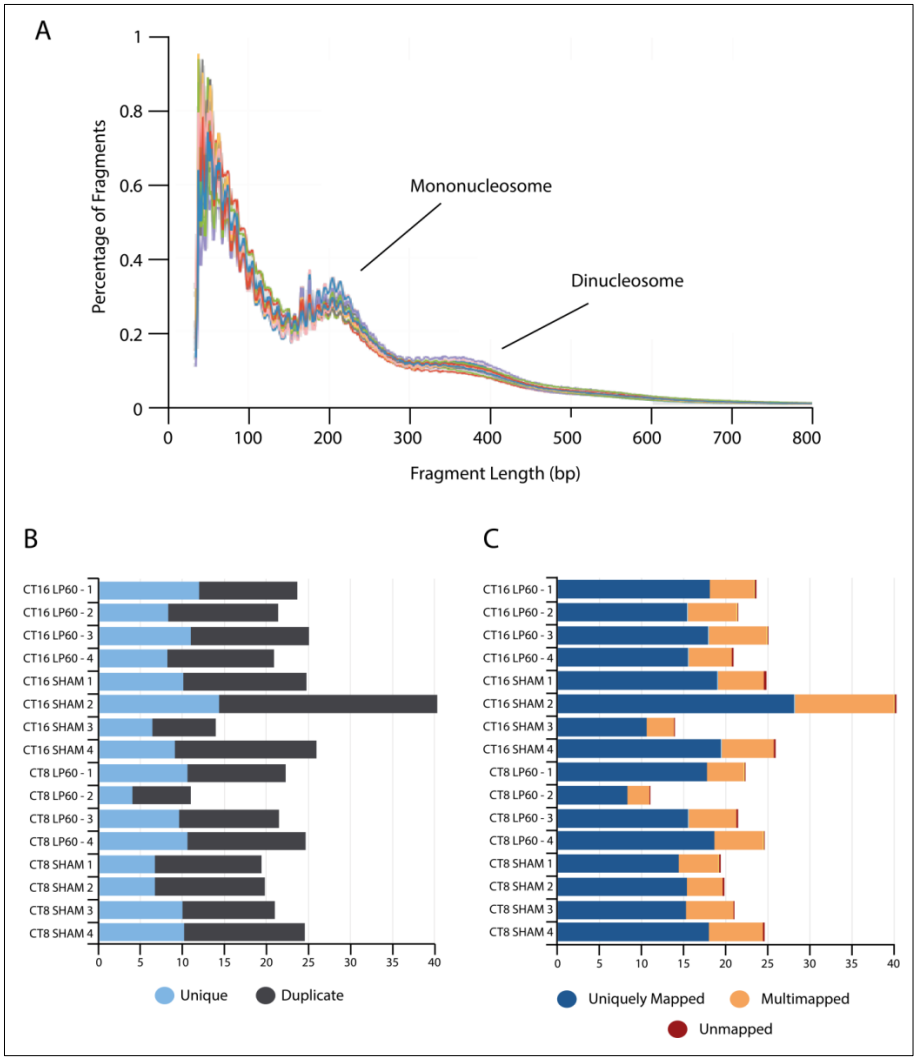


Figure 4.21: Further quality control light pulse experiment 2. A: Fragment size distribution shows both a mono and dinucleosome peak indicating good ATAC-seq library preparation; **B:** Number of duplicates are have improved over the first ATAC-seq experiment with most libraries now having at least 10M unique reads after filtering; **C:** Read mapping statistics to mm10 genome are excellent, with most reads mapping uniquely.

PCA was again carried out to determine if samples clustered into their respective sample conditions and whether any outliers are present. Although analysis of sequencing data quality appeared improved, PCA analysis demonstrated samples did not cluster into their respective treatment groups (Figure 4.22), with samples not clustering in the same manner that they did for the first ATAC-seq light pulse experiment. Several of the light pulsed samples did cluster together, but this included both CT16 and CT8 samples, and was in stark contrast to the first experiment where good separation was observed between the different time points. Therefore, this lack of clustering suggested there would be few common differences between the different conditions, potentially negatively impacting the number of statistically significant genes with differential chromatin discovered.

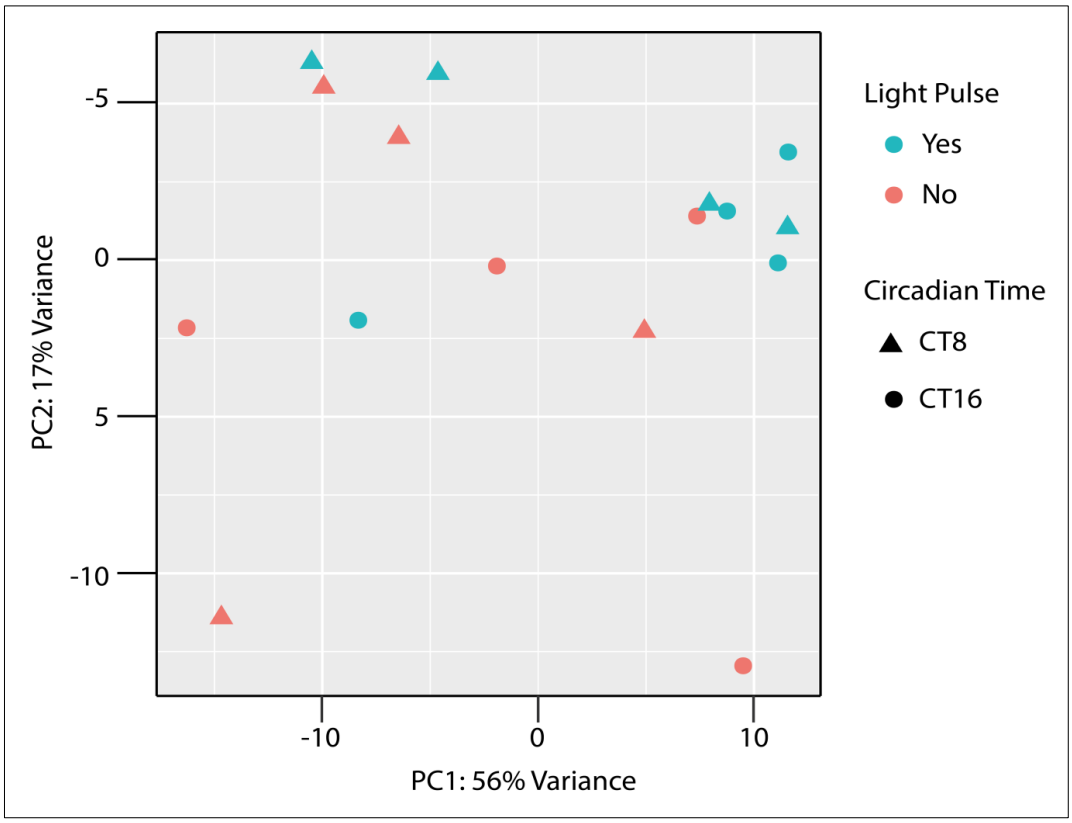


Figure 4.22 PCA plot for read counts from the second ATAC-seq light pulse experiment. Although some samples cluster by whether they were light pulsed or not, samples no longer cluster by the sample of day they were taken.

4.3.2: Differential analysis of aligned reads

Despite the PCA showing poor clustering, a full differential analysis was carried out using DESeq2 (as outlined in section 2.11), to identify areas of differential chromatin accessibility following light exposure. However, and as anticipated, few differential areas of chromatin were found across all sample conditions, and this is clearly illustrated by the MA plots (Figure 4.23. A small number of differential genes were identified for samples light pulsed at CT16 (Figure 4.23-B), providing some agreement with the first experiment.

Across the genome, reads aligned abundantly to the mm10 genome in greater numbers than the first experiment, for example far more reads aligned near to *Fos* for this experiment (Figure 4.24-A). Nevertheless, most of the genes identified as having differential chromatin accessibility in the first experiment now had low levels of aligned reads nearby, and when visualised using the IGB did not show convincing differential chromatin accessibility. In fact, none of the genes highlighted in the first experiment showed differential chromatin; indeed chromatin for almost all these genes now appeared tightly closely, including *Csnk1e* (Figure 4.24-B). For the remaining genes, chromatin which closed now remained open under all conditions, including *MafG* and *Nfil3* (Figure 4.25).

As a consequence of the low number of areas identified to contain differential chromatin, no further downstream analysis was carried out with data from the second ATAC-seq light pulse experiments.

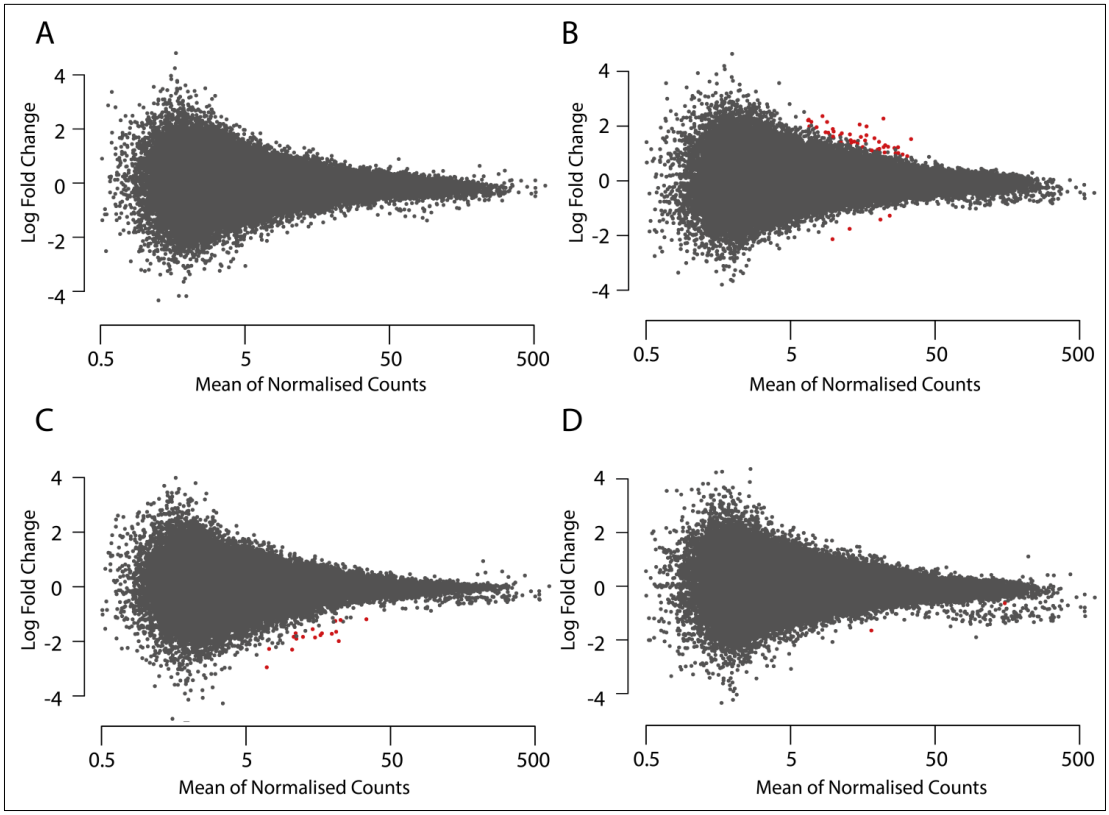


Figure 4.23: Bland–Altman plots for ATAC-seq light pulse experiment 2. Dots represent genes which have differential chromatin nearby; statistically significant genes with an adjusted P-value below 0.05 are highlighted in red **A:** CT8 SHAM vs CT8 LP60; **B:** CT16 SHAM vs CT16 LP60; **C:** CT8 SHAM vs CT16 SHAM; **D:** CT8 LP60 vs CT16 LP60

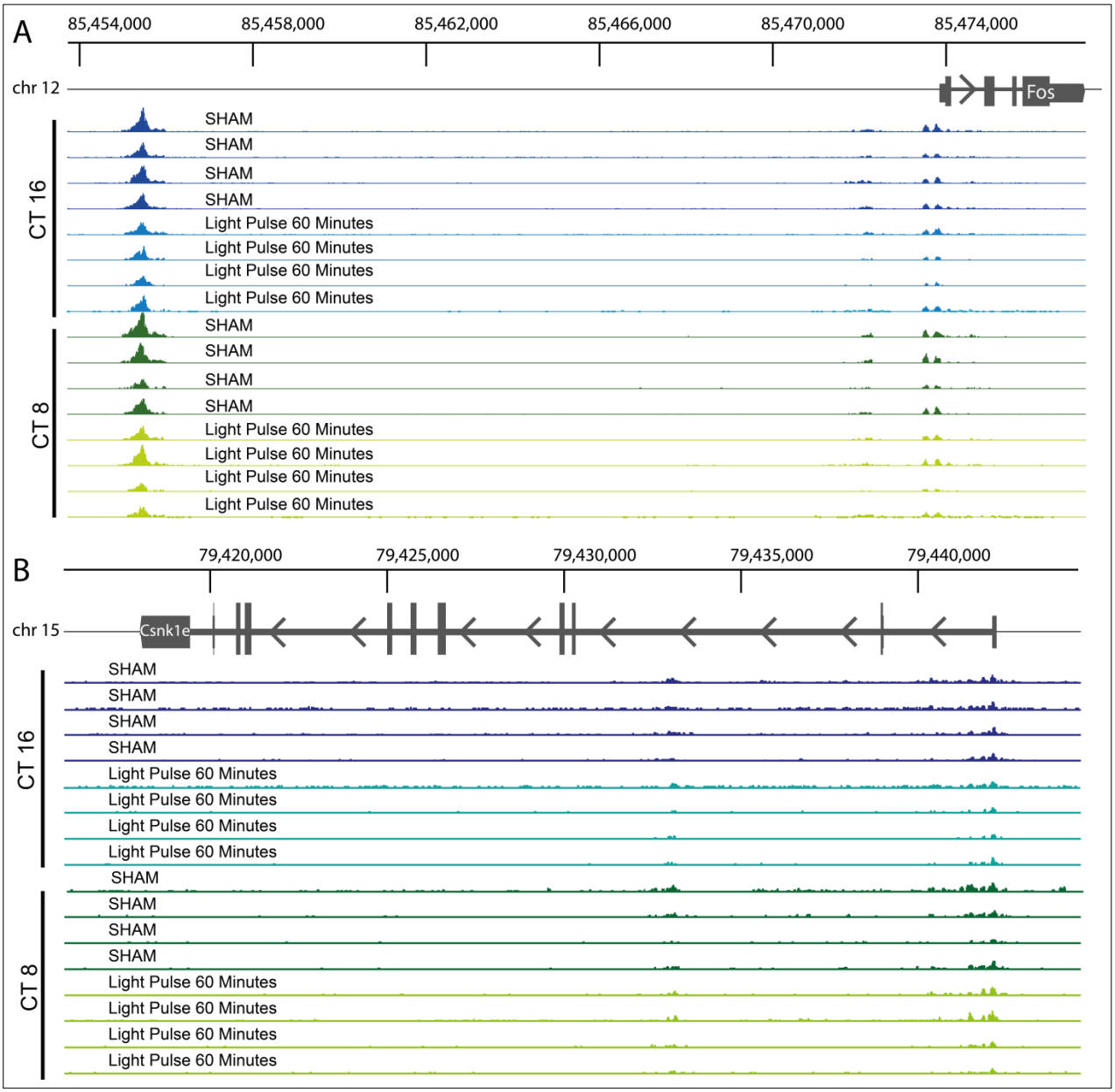


Figure 4.24: Read coverage tracks for *Fos* and *Csnk1e* from ATAC-seq light pulse experiment 2. **A:** Chromatin located upstream of *Fos* now shows a greater number of aligned reads, but remains open across all sample conditions; **B:** Chromatin near to *cnsk1e* in contrast to the first experiment is now closed across all sample conditions.

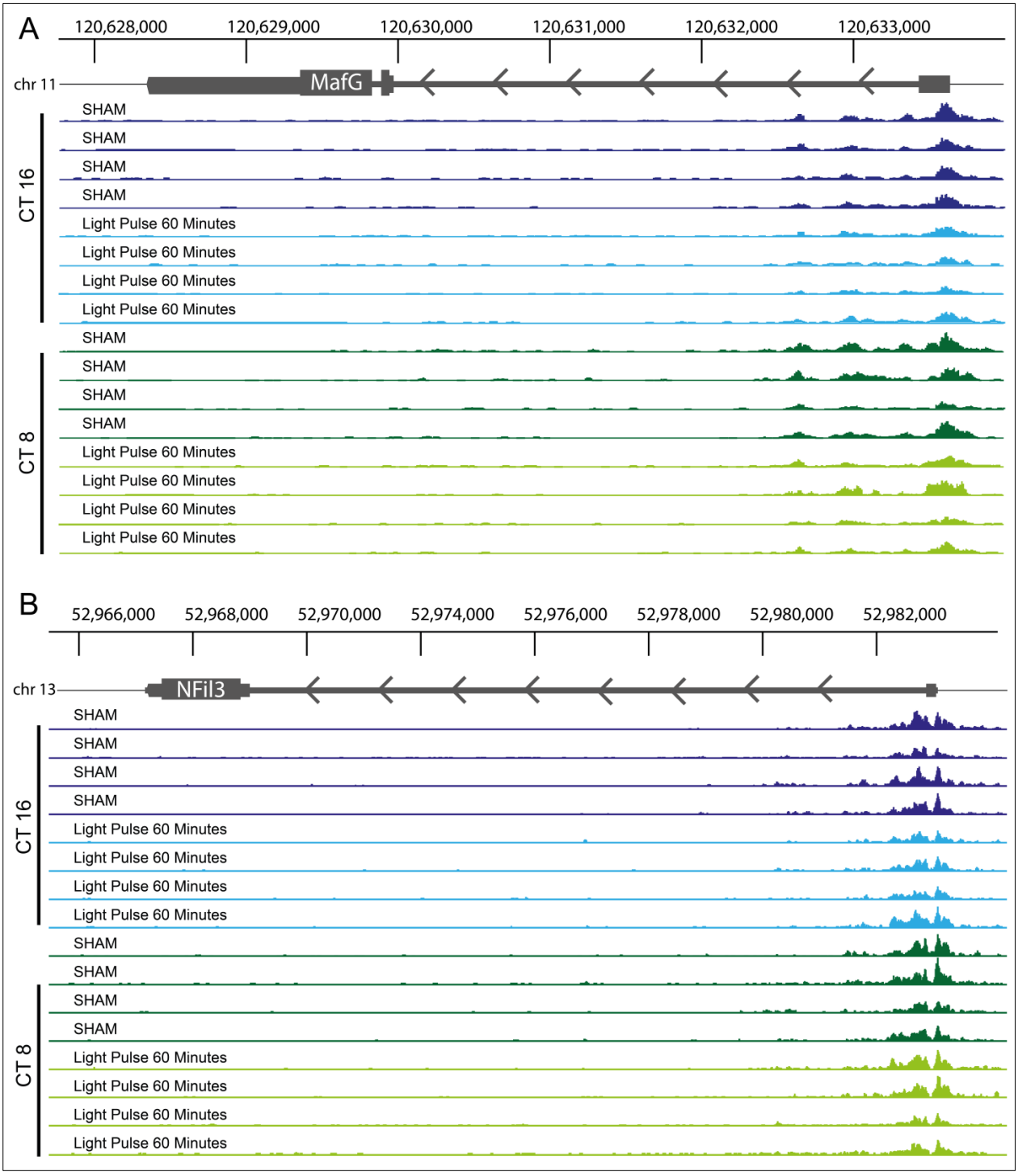


Figure 4.25: Read coverage tracks for *MafG* and *Nfil3* from ATAC-seq light pulse experiment 2; **A:** Chromatin located upstream of *MafG* is now open across all sample conditions; **B:** Chromatin near to *Nfil3* is now open across all open across all sample conditions.

4.4: Transfection of genomic sequences of interest into cultured cells

Following the lack of differential genes found following analysis of data from the second experiment, no further downstream analysis was carried out. Due to COVID-19 pandemic restrictions a third experiment could not be planned, owing to the complexity of the experiment and requirement for multiple persons to assist with tissue collection.

Instead, a series of in vitro investigations were carried out to validate the results obtained from the first experiment. Two genomic sequences of interest were chosen from the first light pulse experiment, in areas where chromatin closed following a light pulse. Chromatin associated with sequences from the promoter regions of *Csnk1e* and *MafG*, was identified as being variable, and had previously been shown to contain numerous epigenetic marks (Figure 4.18). These regions were scanned using Jaspar (Fornes et al., 2020), and were found to contain numerous TF binding sites, which were in addition to binding motifs already found to be enriched within differential chromatin. Due to the large number of TF binding sites present within these sequences, they were ideal for a luciferase reporter assay. The regions of interest were visualised using IGB and three smaller sequences containing multiple TF binding sites were identified within each region for *Csnk1e* (Figure 4.26) and *MafG* (Figure 4.27). The locations for AP-1, CREB and EG1 motifs which have previously shown to be enriched within areas of differential chromatin have been highlighted in these figures.

The genomic regions highlighted in Figures 4.23 and 4.24 were cloned into pGL4.27 plasmids, this plasmid encodes the luciferase reporter gene luc2P and contains multiple cloning regions for insertion of a response element of interest upstream of the promoter and the luc2P gene. The genomic regions were cloned into the pGL4.27 cloning region, and were expected to drive luciferase production due to the large number of TF binding sites they contain. If these regions drive luciferase expression, it can be inferred that they are also likely to drive expression of *Csnk1e* and *MafG*, and are therefore important for the regulation of these two important molecular clock genes.

Mouse NIH3T3 embryonic fibroblast, and human U2OS osteosarcoma cells were cultured as described in section 2.8, and pGL4.27 plasmids containing the sequences of interest were transfected

into cells as outlined in section 2.9. Mouse cells were chosen as mice were used throughout both ATAC-seq experiments, and therefore subsequent genomic analysis had been carried out within the mouse mm10 genome. Human cells were also selected, because the sequences of interest were highly conserved across both species, and if any response is detected in human cells, this is potentially very significant as it may provide insights into entrainment within humans.

To further investigate the importance of *Csnk1e* and *MafG* within the molecular clock and entrainment, following cloning, the cells were treated with Forskolin and horse serum as outlined in section 2.9. Forskolin was chosen to simulate a light pulse through its activation of CRE via the cAMP pathway (Yagita and Okamura, 2000), and AP-1 through induction of cFOS. Horse serum was selected because serum shock has been shown to induce a number of molecular clock genes (Balsalobre et al., 1998); therefore any response from the sequences of interest would confirm their importance within the molecular clock. Two controls were used; cells were transfected with a pGL4.27 plasmid containing only the luciferase reporter, and for cells treated with horse serum, plasmids containing the serum response motif in addition to the luciferase reporter were also transfected into cells.

Following treatment, luminescence through activation of the luciferase reporter was recorded, a baseline of 5 cycles was used, and output recorded for 48 hours following treatment of cells.

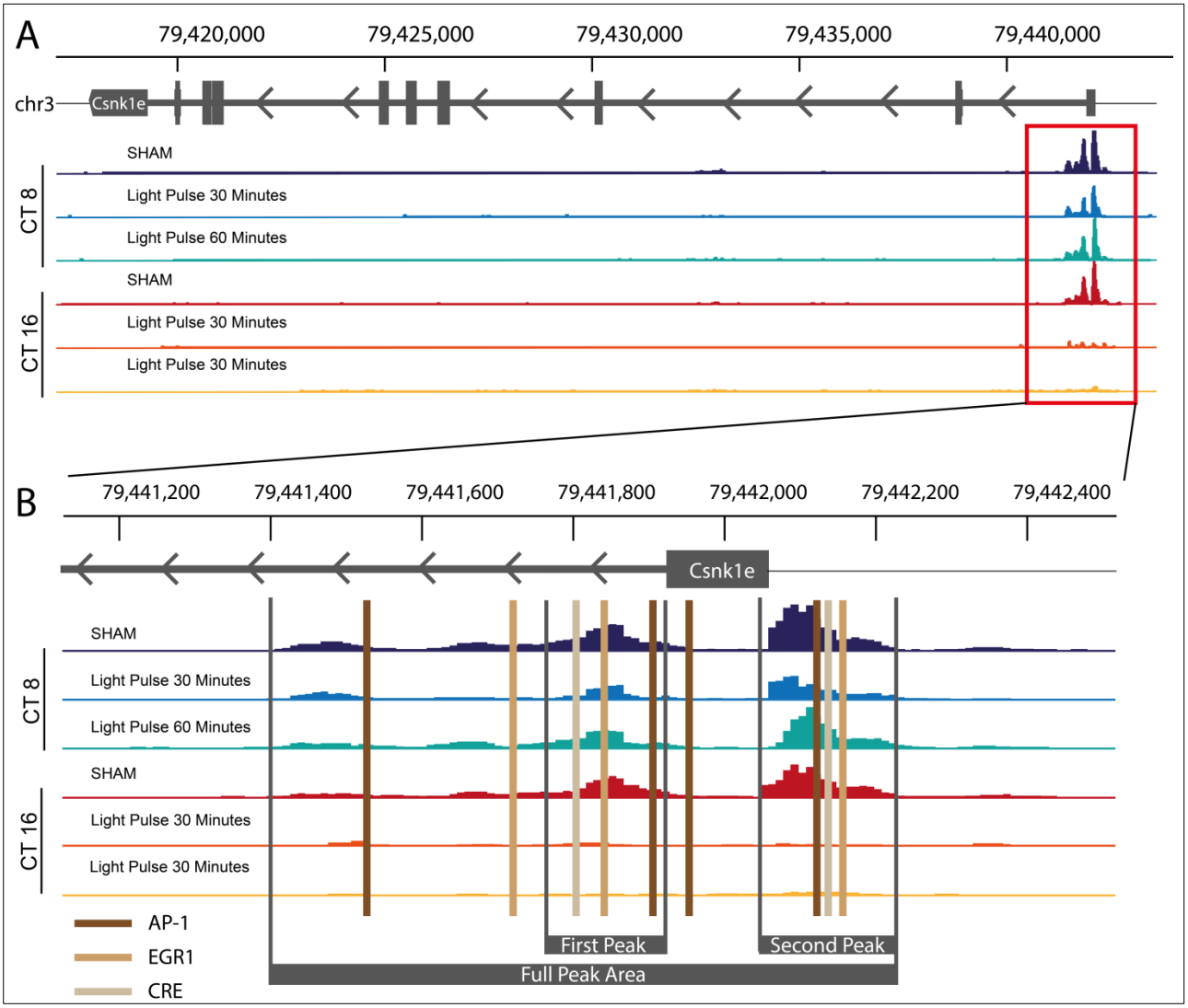


Figure 4.26: Read coverage tracks for *Csnk1e* experiment 1 to show motif locations; A: Read coverage across the whole of *Csnk1e*; **B:** Close-up of the peak region identified for *Csnk1e*; highlighted are the three genomic regions selected for cloning into cultured cells along with the location of several AP-1, CRE and EGR1 motifs.

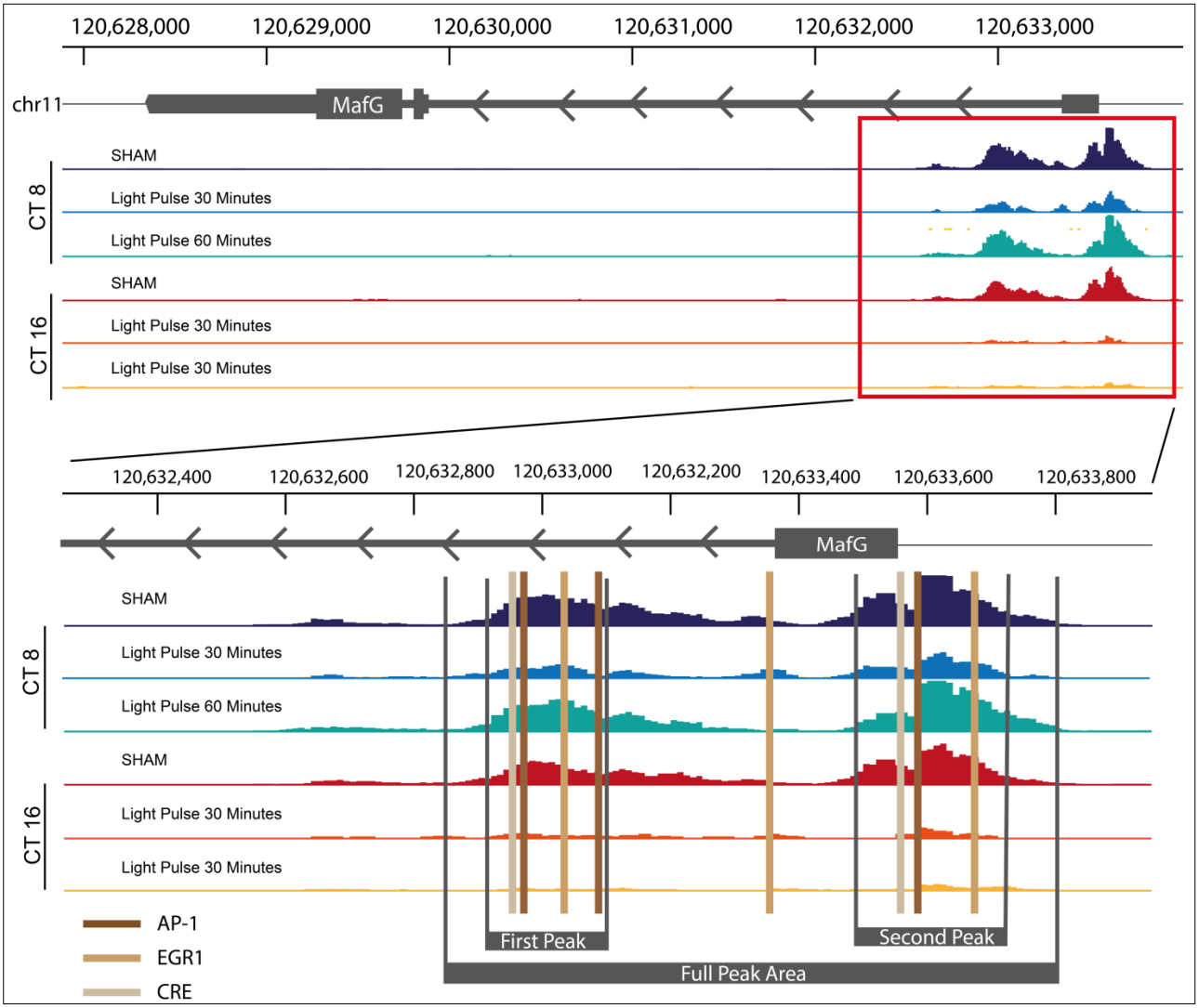


Figure 4.27: Read coverage tracks for *MafG* experiment 1 to show motif locations; **A:** Read coverage across the whole of *Csnk1e*; **B:** Close-up of the peak region identified for *MafG*; highlighted are the three genomic regions selected for cloning into cultured cells along with the location of several AP-1, CRE and EGR1 motifs.

Initial results proved to be highly positive; in both NIH3T3 and U2OS cells, all sequences of interest with the exception of the first *Csnk1e* peak region, without any treatment drove expression of luciferase via the reporter to a greater extent than either controls (Figures 4.28-A, 4.29-A, 4.30-A, 4.31-A). This therefore confirms that the sequences of interest are able to induce transcription, and are likely important for the initiation of transcription for the genes they are located near to.

In NIH3T3 cells, Forskolin further induced luciferase expression in cells transfected with the full and second *Csnk1e* peak regions (Figures 4.28-B/D), but had little effect on cells transfected with the first *Csnk1e* peak region (Figure 4.28-C). In cells transfected with the full and first *MafG* peak regions Forskolin strongly drove further luciferase expression (Figures 4.28-E/F), but had little effect on cells transfected with the second *MafG* peak region (Figure 4.28-G). Further analysis of luciferase expression following treatment with Forskolin did not fully support greater induction of luciferase following treatment (Figure 4.38-H). On the other hand horse serum very strongly induced further luciferase expression across all cells regardless of which peak region they had been transfected with (Figures 4.29-B/C/D/E/F/G). Further analysis gave strong support that treatment with horse serum further up regulates luciferase expression in cells transfected with the sequences of interest, to a greater extent than the serum response motif (Figure 4.29-H).

In U2OS cells Forskolin supressed luciferase expression for cells transfected with the full *Csnk1e* region (Figure 4.30-B), and had little effect on luciferase expressions for cells transfected with the first or second *Csnk1e* peak regions (Figures 4.30-C/D). For cells transfected with the full and first *MafG* peak regions, Forskolin somewhat further induced luciferase expression (Figures 4.30-E/F). While for cells transfected with the second *MafG* peak region Forskolin again supressed luciferase expression (Figure 4.30-G). Further analysis only gave weak support that Forskolin is able to induce or suppress further luciferase expression in U2OS cells transfected with the sequences of interest (Figure 4.30-H). In contrast to NIH3T3 cells, horse serum showed little effect in U2OS cells transfected with *Csnk1e* sequences of interest (Figures 4.31-B/C/D). Horse serum showed some effect in NIH3T3 cells transfected with *MafG* sequences of interest, in cells transfected with the full *MafG* peak region horse serum blunted luciferase expression (Figure 4.31-E), in those transfected with the

first *MafG* peak region horse serum increased luciferase expression (Figure 4.31-F), and cells transfected with the second *MafG* peak region horse serum had little effect (Figure 4.31-G). Further analysis supported the effects of horse serum on cells transfected with the full and first *MafG* peak regions (Figure 4.31-G).

Overall, all sequences of interest were shown to strongly drive luciferase expression without treatment in both cell lines, therefore indicating their importance in the initiation of transcription in both mice and humans, demonstrating a number of transcription factors must bind these sequences in both species. Following treatment the biggest response was observed when NIH3T3 cells were treated with horse serum, where treatment strongly drove greater luciferase expression across cells transfected with any of the sequences of interest. This potentially demonstrates the importance of the sequences of interest in relation to entrainment, due to serum inducing a number of molecular clock genes (Balsalobre et al., 1998). In U2OS cells, this was only observed for cells transfected with the first *MafG* sequence of interest, indicating species specific response to the treatment. While Forskolin showed some effect in NIH3T3 cells transfected with sequences of interest, but showed little effect in U2OS cells.

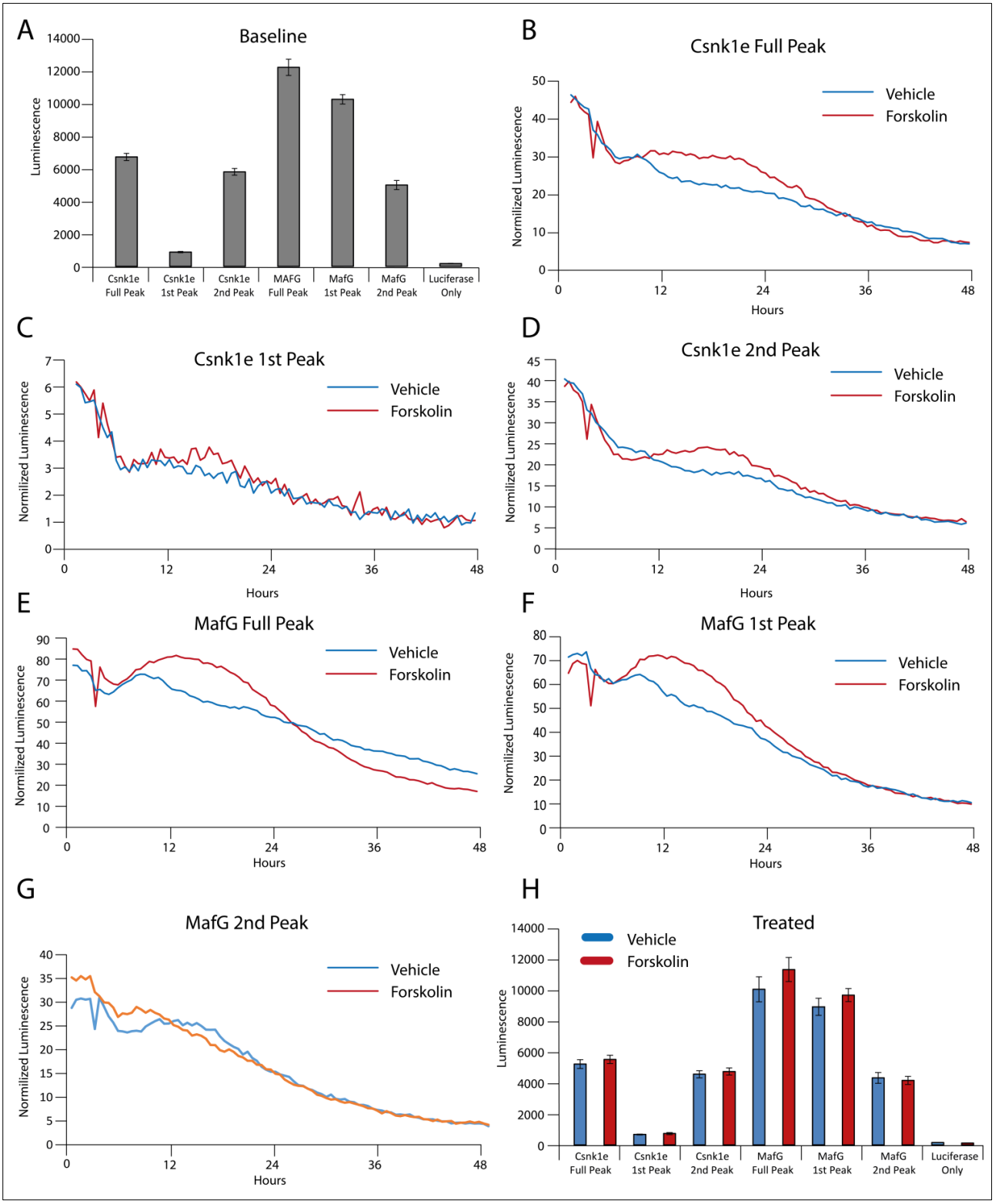


Figure 4.28: NIH3T3 cells transfected with the sequences of interest and treated with Forskolin or DMSO vehicle; A: Baseline reponse prior to treatment **B**; *Csnk1e* Full peak area **C:** *Csnk1e* first peak area; **D:** *Csnk1e* second peak area; **E:** *MafG* full peak area; **F:** *MafG* First peak area; **G:** *MafG* second peak area; **F:** Further analysis following treatment.

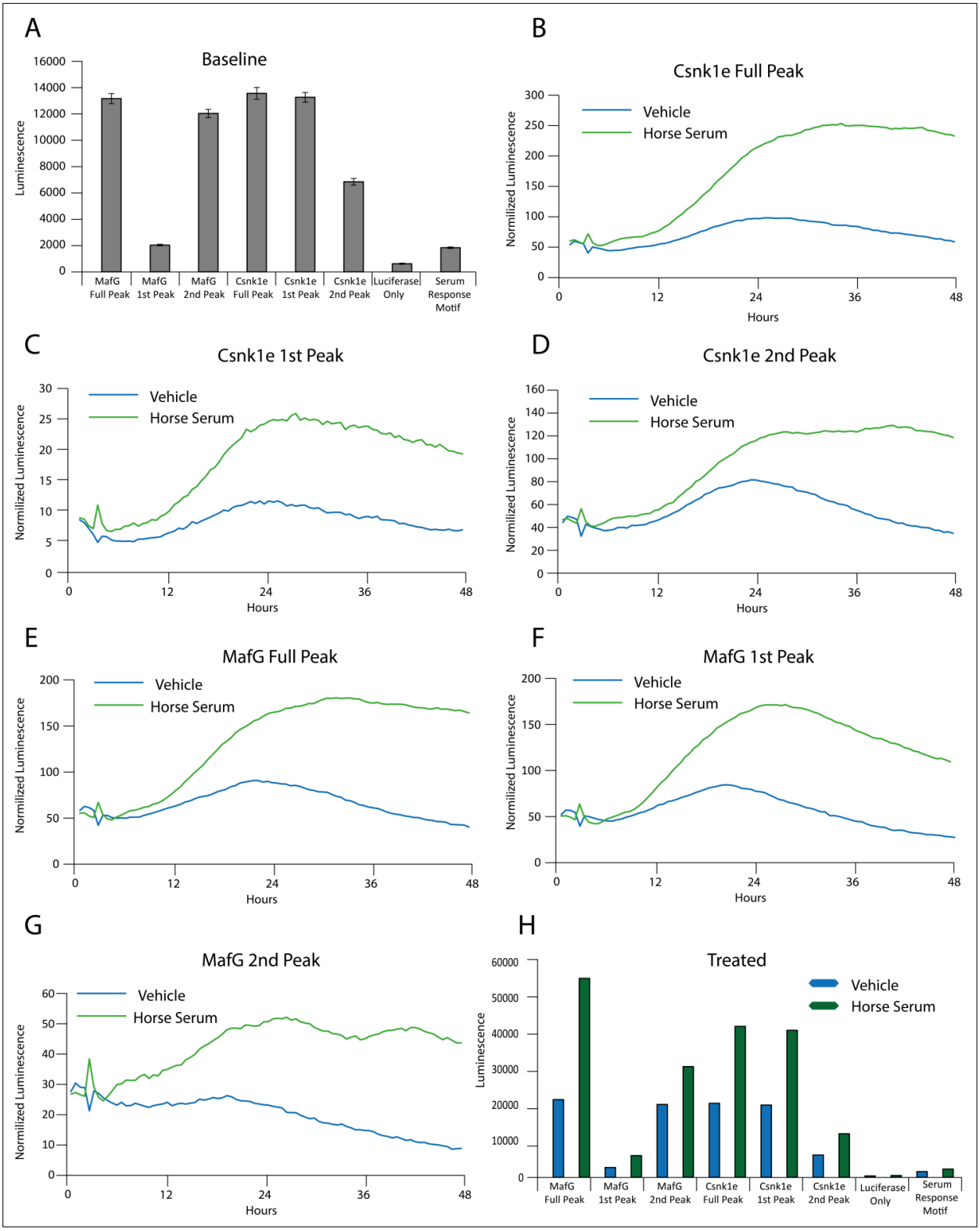


Figure 4.29: NIH3T3 cells transfected with the sequences of interest and treated with horse serum or DMSO vehicle; A: Baseline reponse prior to treatment **B**; *Csnk1e* Full peak area **C:** *Csnk1e* first peak area; **D:** *Csnk1e* second peak area; **E:** *MafG* full peak area; **F:** *MafG* First peak area; **G:** *MafG* second peak area; **F:** Further analysis following treatment.

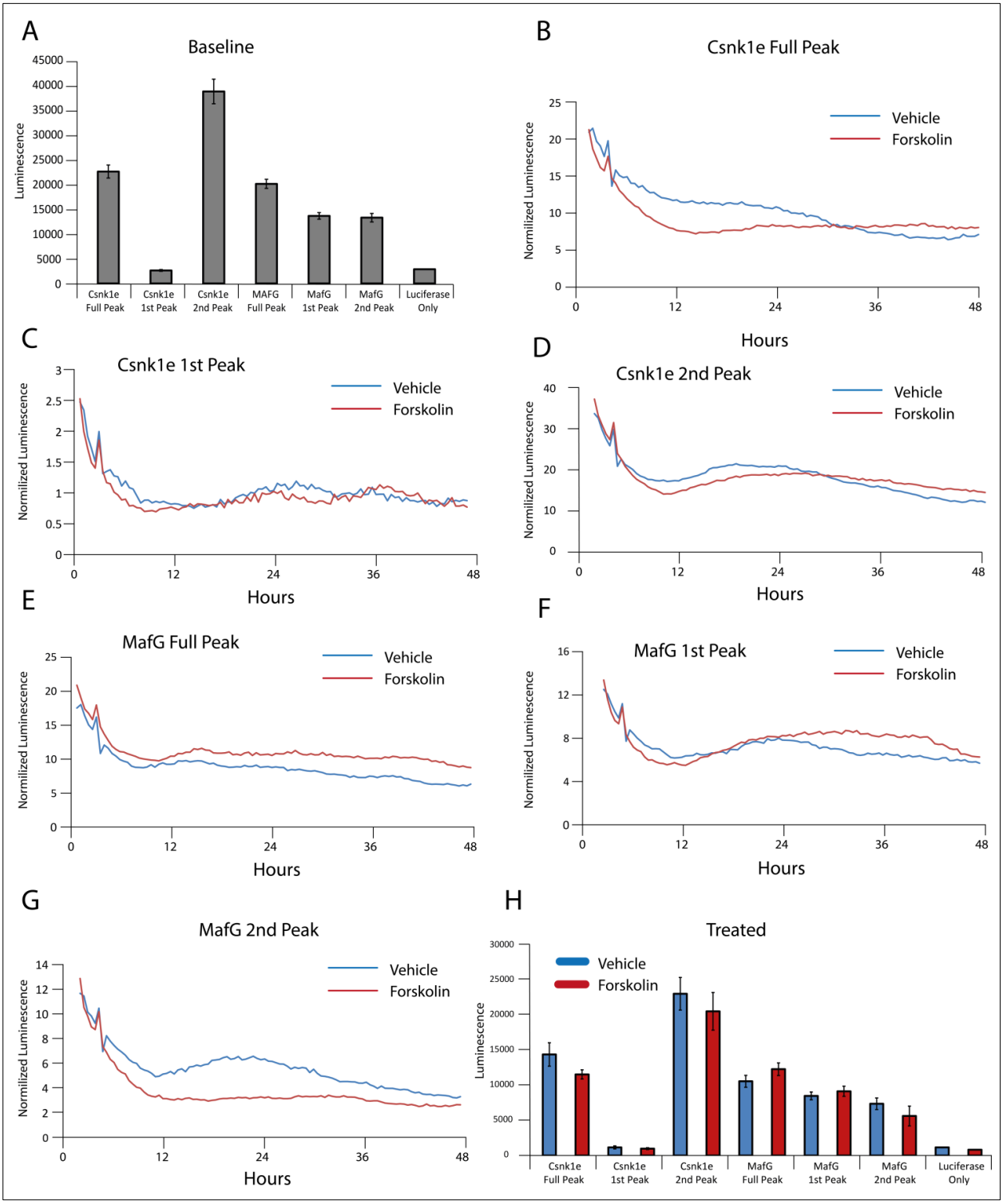


Figure 4.30: U2OS cells transfected with the sequences of interest and treated with Forskolin or DMSO vehicle; A: Baseline reponse prior to treatment **B**; *Csnk1e* Full peak area **C:** *Csnk1e* first peak area; **D:** *Csnk1e* second peak area; **E:** *MafG* full peak area; **F:** *MafG* First peak area; **G:** *MafG* second peak area; **F:** Further analysis following treatment.

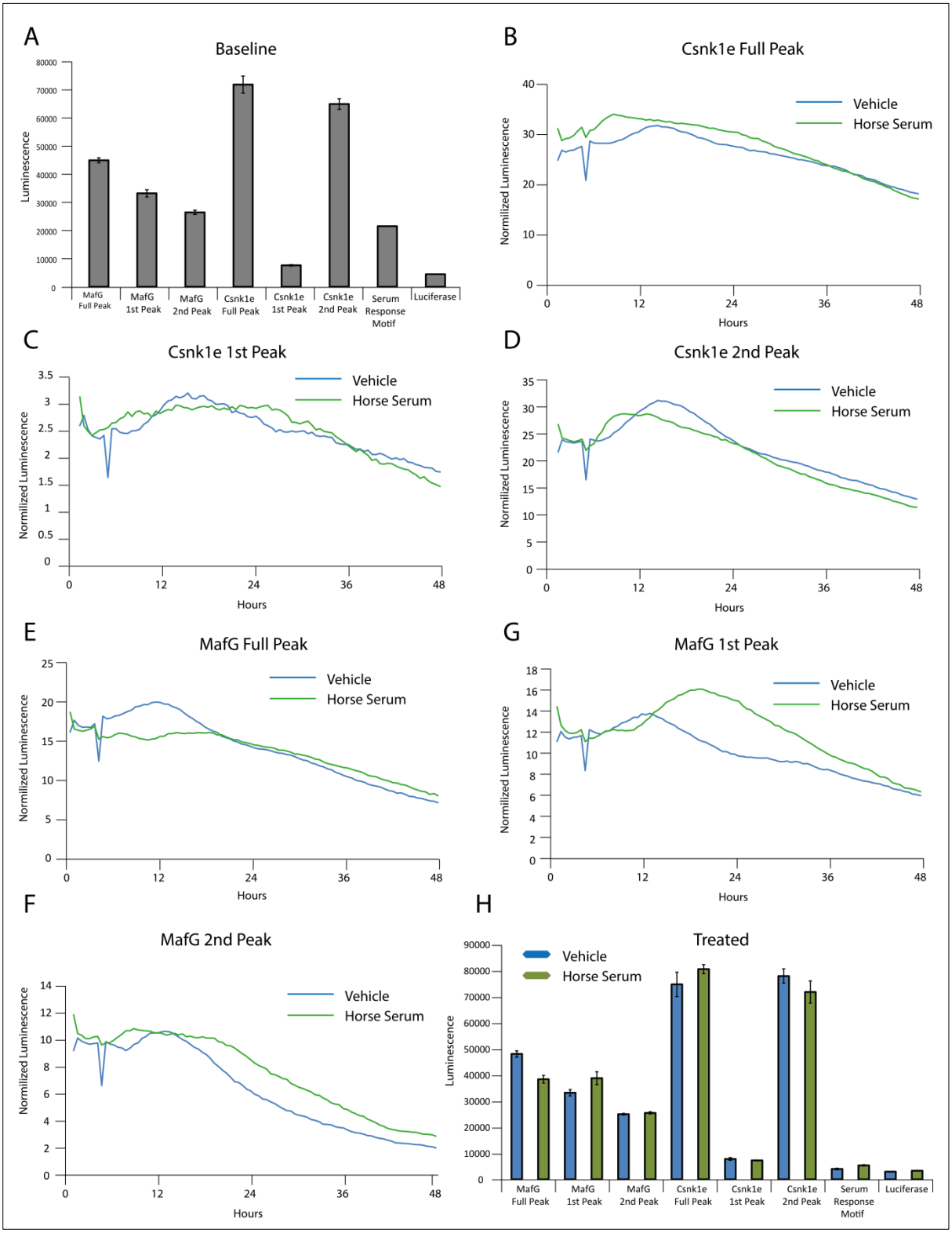


Figure 4.31: Figure 4.30: U2OS cells transfected with the sequences of interest and treated with horse serum or DMSO vehicle; A: Baseline reponse prior to treatment **B**; *Csnk1e* Full peak area **C:** *Csnk1e* first peak area; **D:** *Csnk1e* second peak area; **E:** *MafG* full peak area; **F:** *MafG* First peak area; **G:** *MafG* second peak area; **F:** Further analysis following treatment.

4.5: Chapter 4 discussion

This chapter aimed to expand the successful results obtained from the second pilot experiment with a much larger ATAC-seq light pulse experiment, and verify that despite the small number of cells contained within the SCN a method to successfully apply ATAC-seq to mouse SCN tissue had been accomplished. It has no doubt been demonstrated that ATAC-seq although a difficult technique to master has been used successfully used with mouse SCN tissue, with two datasets having been attained. Undesirably, the data from each experiment are not in full agreement, and due to Covid-19 restrictions a further repeat to resolve the discrepancy cannot be completed at present.

The results obtained from the first light pulse experiment support what is currently known about circadian biology and entrainment. The SCN is only responsive at certain times of the day (Khalsa et al., 2003), and here light pulsing only altered chromatin accessibility following a light pulse at CT16, this being the most light responsive time of day. The first light pulse experiment appeared to answer all questions proposed by the original hypothesis; firstly chromatin remodelling was clearly shown to be instigated by a light pulse, as undoubtedly shown in Figure 4.6 for *Csnk1e*, and these changes are only occurring at specific times of the day. When gene lists for those with differential chromatin located nearby following a light pulse were compared to those whose transcription changes following a light pulse, 22 genes were identified in both lists, this therefore demonstrates that chromatin accessibility can have an effect on transcription.

Notably, several genes already known to be associated with the molecular clock were identified to have dynamic chromatin organisation. Indeed, chromatin closed near *Nfil3* and *Csnk1e* following a light pulse at CT16 and chromatin within *Ror* was remodelled over the course of a day demonstrating diurnal variation in chromatin accessibility. The role of the protein product of *Csnk1e,* CSNK1E is already well understood within the molecular clock (Harms et al., 2004, Narasimamurthy et al., 2018), it is a protein kinase responsible for marking PER1 for degradation, helping prevent its re-entry into the nucleus. Although not investigated here, if the promoter region of CSNK1E are indeed closing at CT16 following a light pulse, this will likely switch off production of CSNK1E, and may therefore form an important part of re-setting the molecular clock following light exposure and subsequent

phase shifting. By simply closing the chromatin, the production of CSNK1E is hypothesised to be prevented at the beginning of the chain of transcription; therefore it may take some time for it to actually have any effect on the activity of CSNK1E. This may be part of a fine balance within the molecular clock where a light pulse initially increases PER1 production, but when this induction is switched off by SIK1, under certain conditions switching off of CSNK1E, may ensure the initial up–regulation of PER1 persists by slowing its degradation.

Both *Ror* and *Nfil3* are integral parts of the ancillary molecular clock, as shown in Figure 1.8, and therefore modifications to chromatin near these genes will undoubtedly have an impact on the expression of a number of molecular clock genes. Both ROR and NFIL3 bind to the D-box within promoters of genes within the ancillary molecular clock and core molecular clock (Yoshitane et al., 2019), therefore their up or down regulation will have implications for both the ancillary and core molecular clock. Indeed, differing expression of *Ror* or *Nfil3* will have an effect on the expression of core clock genes *Per1* and *Cry*, therefore other clock controlled genes within the SCN may also be affected by differing expression of *Ror* and *Nfil3*.

Downstream analysis of data from the first light pulse experiment continued to support the initial hypothesis; both motif searching and gene enrichment analysis of differential gene lists gave sensible results in line with current literature. Motif searching found enrichment for AP-1, ZFX and ROR motifs, already known to be integral for circadian entrainment (Preitner et al., 2002, Takeuchi et al., 1993, Parsons et al., 2015) , and gene enrichment analysis showed biological and molecular functions such as regulation, protein binding and kinase activity critical for proper circadian function (Obrietan et al., 1998). This therefore indicates other motifs also found to be enriched, are worth further investigation, and their activity as cis regulatory elements provide potential candidates for additional experiments. The sequences of the motifs could be knocked out in cultured cells using techniques such as CRISPR/Cas9 (Ceasar et al., 2016), and the effect this has on expression of core clock genes such as *Bmal1* could be monitored using a luciferase reporter.

Indeed, further experiments already carried out using a luciferase reporter assay to investigate regions of differential chromatin within the promoters of *Csnk1e* and *MafG,* have demonstrated their

importance in transcriptional regulation. This gives further support that areas of differential chromatin identified following light pulsing at CT16 in the first light pulse experiment, are important for the regulation of transcription. The sequences of interest were shown to induce luciferase expression when transfected into cells, and responded to Forskolin and serum treatment in both human and mouse cells. This demonstrates these areas not only induce gene expression, but respond to further stimuli, with serum activating a broad spectrum of molecular clock genes (Balsalobre et al., 1998), and Forskolin activates several clock genes via the CREB pathway (Yagita and Okamura, 2000), their response to the treatment shows the importance of *Csnk1e* and *MafG* in circadian entrainment.

Owing to the fact results from experiment 1 are in full agreement with the original hypothesis, and fall in line with what is already understood about circadian biology and entrainment, it would seem likely a technical error with the second light pulse experiment may have be responsible for the discrepancy. However, initial QC of data from experiment 2 indicated the data obtained represented a better dataset; this would therefore indicate a technical error to be unlikely. Nevertheless, ATAC-seq is still a relatively new technique, and there still remain numerous gaps in our understanding of DNA accessibility and chromatin structure. As the second experiment gave a greater data quality, it is likely that more nuclei were successfully isolated, and subsequently incubated with the Tn5 enzyme. The greater abundance of nuclei could therefore have presented a problem for the Tn5 enzyme, and perhaps only the most accessible areas of chromatin were digested. Whereas when fewer nuclei are present, as was likely the case in the first experiment, the Tn5 enzyme may act more aggressively and cut out areas that are less accessible but still open. The published protocol indicates a ATAC-seq can be applied to as few as 500 nuclei, or as many as 50,000 nuclei, prior nuclei counting confirms the upper threshold is unlikely to have been passed, but the protocol gives no indication how the number of nuclei assayed at one time may affect data obtained. This consequently presents a further experiment which must be carried out both to resolve the discrepancy here, and for the wider scientific field, because ATAC-seq is no doubt being used throughout numerous laboratories. There are additionally other factors which could have had an effect on the nuclei extracted, which could include for example how long the SCN were left on ice before incubation with Tn5.

Moving forward, a full repeat of the ATAC-seq light pulse experiment is required to resolve the inconsistencies between the two experiments. I hypothesise that the results will concur with the first experiment, as this undoubtedly was most consistent with current knowledge about circadian and epigenetic biology; it seems highly unlikely that chromatin structure remains static, and does not change in line with external stimuli. In addition to repeating the light pulse experiment, further ATAC-seq experiments should also be carried to define other factors such as whether varying levels of starting material or freezing tissue have an effect on data obtained. Assuming the results from the first experiment are confirmed, a plethora of potential hypothesis will be available for further downstream investigation. A list of over 400 genes were shown to have differential chromatin, many other these genes are likely to play an important role in circadian entrainment and should therefore be investigated further. Additional luciferase reporter assays could be carried out to determine of other areas of differential chromatin accessibility are also able to drive transcription. Promising genes with differential chromatin nearby could also be knocked out in whole animals, this will help determine their importance and allude to what effect their loss has across the entire organism.

Chapter 5

The effect of sleep deprivation on the chromatin structure and transcriptional landscape of specific brain regions

5.1: Introduction

The need for sleep has persisted throughout evolution (Miyazaki et al., 2017); this in itself demonstrates its importance, and lack of regular sleep leads to a variety of health consequences throughout the body (Buxton and Marcelli, 2010). Chronic SD is an endemic problem of modern society, in the UK one in every eight people work night shifts (Gill, 2016), the result of this is a combination of both a disrupted circadian rhythm and continuous SD (Ohayon et al., 2002). Therefore a greater understanding of the molecular pathways and processes affected by SD will help to elucidate, exactly why so many health complications result from a continuous lack of sleep.

Although SD has consequences throughout the body, it is frequently in the brain where SD has the largest impact; indeed a lack of sleep often results in an immediate decline of cognitive abilities, including learning and memory (Alhola and Polo-Kantola, 2007). However, it is often over longer periods of time that the negative impacts of continuous SD result in the most harm to an individual, and this can manifest in a number of different ways within the brain. Moreover, among retirees from occupations involving night and rotational shift work, there is often an increased likelihood of developing many long term debilitating neurodegenerative disorders, for example a greater occurrence of dementia is frequently observed (Thomas et al., 2019). These conditions usually worsen over time, often leading to more severe forms of dementias such as Alzheimer's Disease (Wu et al., 2019), and therefore this often leads to a poor quality of life for shift workers following retirement.

Although SD will almost always have some negative consequence on an individual, there are frequently great variances on the severity of both short term (Kuna et al., 2012) and longer term health implications (Lourida et al., 2019). Indeed, it is not uncommon for some individuals to work night

shifts throughout their life and never develop a neurological disorder such as dementia; this therefore indicates that there is a fundamental genetic basis to an individual's susceptibility to SD. This project aims to gain a greater understanding of the underlying mechanisms at the molecular level of SD. This will help to reveal why some individuals are more susceptible to the harmful consequences of SD, and in turn begin to form the basis of future treatments for individuals suffering from chronic SD, such as night shift workers.

At the molecular level, SD is well documented to lead to a variety of changes in gene expression (Cirelli and Tononi, 1999), and is known to influence metabolism, synaptic plasticity, stress response, and neuroprotection (Husse et al., 2012, Terao et al., 2003, Chen and Bazan, 2005). In the rat brain, stress response genes such as *Mrp14* have been shown to increase following SD, and levels of a number of mitochondrial genes encoded by the mitochondrial genome are increased following SD in several cortical and subcortical areas (Terao et al., 2006, Cirelli and Tononi, 1998). In the mouse cerebral cortex, *BiP* expression is increased following 6 hours of SD; an increase in BIP is related to triggering the unfolded protein response, which includes the slowing down of protein synthesis, a further indication of cellular stress (Naidoo et al., 2005).

Several genes have already been shown to be important in the molecular response to sleep homeostasis following SD. Activity-regulated cytoskeleton-associated protein (*Arc*) is strongly induced following SD, especially in the cortex and hippocampus (Thompson et al., 2010). *Arc* appears to play an important role in recovery from SD, with *Arc* knockout (KO) mice displaying increased time spent in rapid eye movement (REM) sleep and attenuation of sleep rebound following SD (Suzuki et al., 2020). Furthermore, up regulation of *Arc* influences the induction of a further subset of SD-response genes, including *c-fos*, *Egr-1*, *Homer1a*, *Nur77*, and *Bip/GRP7/8* (Suzuki et al., 2020), with up regulation of *homer1a* being frequently postulated as a notable marker of SD (Maret et al., 2007).

Due to the large number of gene expression changes following SD, it is reasonable to hypothesise that changes must be occurring at the epigenetic level to facilitate such wide ranging changes in expression. Indeed SD has already been shown to lead to both DNA methylation (Lahtinen et al.,

2019) and histone modification (Duan et al., 2016), as well as changes in levels of small RNAs (Davis et al., 2007). Following SD the landscape of DNA methylation has been shown to radically change, with synaptic plasticity being particularly affected (Massart et al., 2014). Methylation occurs in both a dynamic manner changing on a daily basis (Xia et al., 2015), and has been observed to accumulate to a greater extent as we age (Horvath and Raj, 2018). For SD this has two implications, by changing DNA methylation on a daily basis, this may have an immediate and direct effect on gene expression, possibly explaining the immediate consequences of SD on cognition and memory. In addition, if SD is indeed increasing DNA methylation, and this change is cumulative, SD may actually be increasing the rate of ageing within individuals working night shifts (White et al., 2019). Furthermore, modifications to histone proteins following SD have been shown to cause spatial memory impairment in rats (Ruifeng et al., 2016), with the inhibition of HKMTs shown to effect both the sleep and modulation of neurochemicals in rats (Murillo-Rodríguez et al., 2019). Therefore HKMTs likely play an important role in the recovery from SD, and gaining a greater understanding of their function may provide clues about how the brain responds to SD.

These extensive epigenetic modifications will influence chromatin accessibility; this project aims to investigate if these changes are occurring in areas where chromatin accessibility is being altered. It is probable that changes to chromatin accessibility will occur within regulatory regions of DNA such as promoters, and this will directly influence which transcription factors are able to bind promoter regions. Meta-analysis of short-term SD data in mouse brain, has already demonstrated extensive differences in transcription factor binding following SD, with the CRE motif being one of the key cis-regulatory elements controlling SD-affected genes (Wang et al., 2010). In a similar manner to Hor et al. (2019), this project aims to use ATAC-seq to determine if chromatin accessibility at cis-regulatory elements within DNA is changing following SD, and RNA-seq to investigate the impact of any chromatin accessibility changes discovered on the expression of genes located nearby. However, this project will also focus on the immediate effects of SD, and investigate chromatin structure in the SCN and DG.

Furthermore, sleep is well recognized as essential for recovery from wakefulness, carrying out numerous functions at both the neural and molecular level (Cirelli and Tononi, 2008). Oxidative stress has frequently been shown to increase during periods of wakefulness, resulting a greater number of DNA double-strand breaks (DSBs) occurring (Lieber, 2010). To facilitate the effective restoration of DNA following prolonged SD, chromatin must be relaxed to enable effect repair (Hauer and Gasser, 2017). An increase in chromosome dynamics to facilitate repair of DSBs has been shown in zebrafish larvae following SD (Zada et al., 2019), ATAC-seq will be employed to further investigate if areas of the genome become more accessible to enable repair of DSBs following SD.

This project aims to answer the following questions:

- Does SD cause changes in chromatin structure?
- Are SD-induced changes in chromatin structure leading to changes in gene expression?
- Do changes in epigenetic marks take place in areas of the genome where chromatin accessibility is changing?
- Does SD lead to changes in the chromatin accessibility of important cis-acting TF binding motifs?

5.2: Chromatin accessibility dynamics in the mouse brain after acute SD

To assess the impact of acute SD on chromatin structure throughout the brain, ATAC-seq was conducted on the SCN, DG and cortex. The SCN was chosen due to its importance in entrainment, to investigate if any genes linked to entrainment are also affected by SD. In mice, the SCN is less responsive to light in the evening (Phillips et al., 2019), and the mechanism for this is not fully understood, but is likely to be due to either a circadian, or sleep/tiredness affect. The DG was selected because it is responsible for memory formation that occurs during the night; and SD has already been shown to effect gene expression in the hippocampus (Delorme et al., 2019). Finally the cortex was assayed because the chromatin state within the murine cortex has already been shown to change following SD (Hor et al., 2019), enabling comparisons to be made with published data already available.

Mice were sleep deprived for 6 hours at CT0 (control animals (sham) were not sleep deprived), and then brain punches taken from the different brain regions. Two replicates were collected for each sample condition, with each replicate containing tissue from four mice. Table 5.1 summarises the experimental groups. Nuclei were isolated as described in section 2.1; ATAC-seq libraries were prepared from nuclei as outlined in section 2.3, libraries were then quantified as outlined in section 2.4, sequenced as described in section 2.5, and the resulting data analysed as set out in section 2.10.

Group Name	Tissue	Sleep Deprived
SCN-SD	SCN	Yes
SCN-SHAM	SCN	No
DG-SD	DENTATE GYRUS	Yes
DG-SHAM	DENTATE GYRUS	No
Cortex-SD	CORTEX	Yes
Cortex-SHAM	CORTEX	No

Table 5.1: The experimental groups for the ATAC-seq SD experiment. Each sample condition comprised 2 replicates.

5.2.1: Quality Control

Following sequencing of the ATAC-seq libraries, quality control was conducted to assess overall sample quality, and the samples passed numerous quality control tests. Initial read count analysis demonstrated that the reads were evenly distributed across the different sample groups, indicating the libraries were of good quality and had been successfully diluted prior to pooling (Figure 5.1).

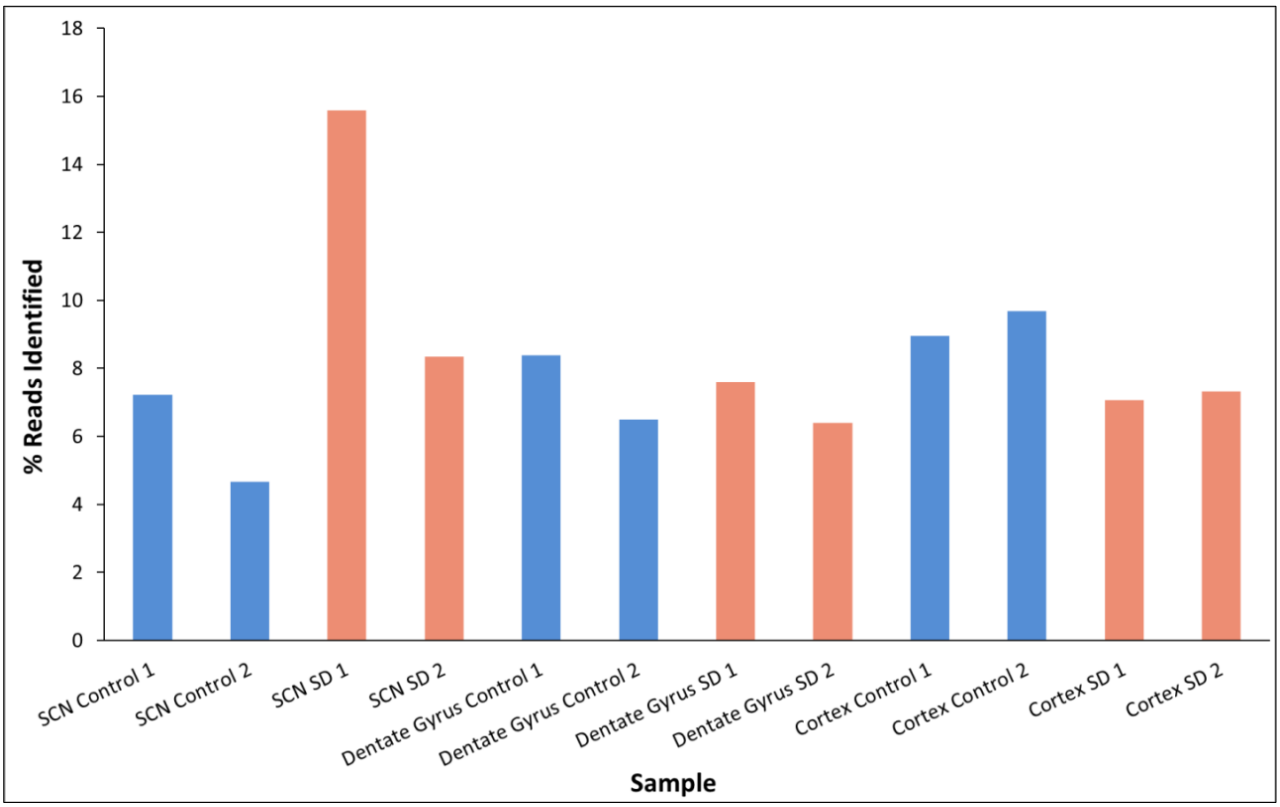


Figure 5.1: Read distribution across the different ATAC-seq samples. An even distribution can be seen, indicating good library preparation, and even dilution of samples prior to sequencing.

All tests carried out by FastQc to determine sequencing quality indicated data to be of a reasonable quality. Overall read quality was high, with all samples displaying good per sequence (Figure 5.2A) and per base (Figure 5.2B) quality scores. Average GC content and sequence duplication was high, but in line with the light pulse experiments (Figure 5.2C/D). Per base N Content is excellent,

indicating the sequencer successfully discerned the base at each point in the sequence (Figure 5.2- E). Read length distribution shows most sequences are 37bp or 38bp as expected due to a 70 cycle cartridge being used for paired end sequencing (5.2-F).

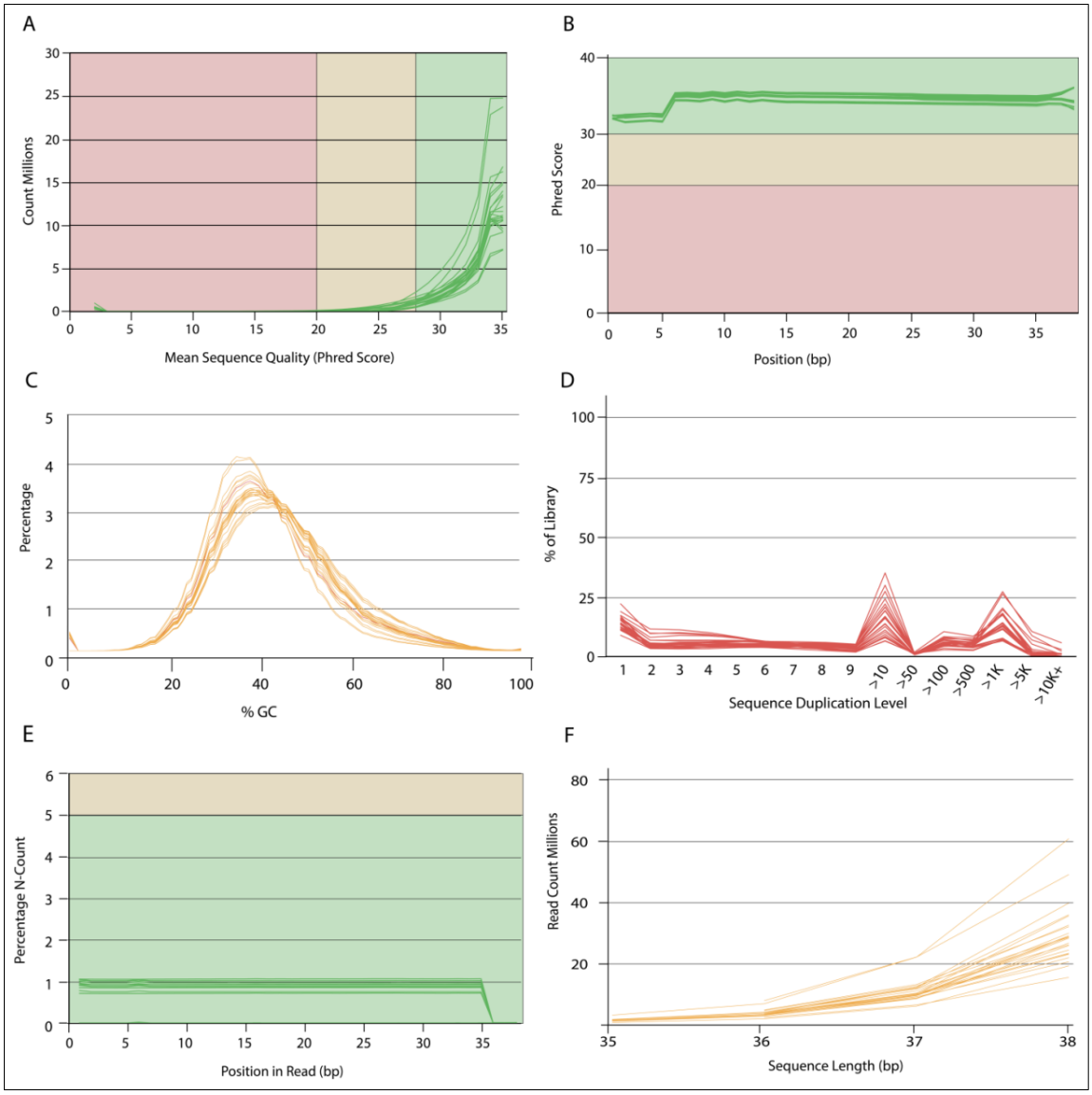


Figure 5.2: FastQc results for SD ATAC-seq. A: Per sequence quality is high; **B:** Mean quality score is high; **C:** GC content shows some bias; **D:** A high level of sequence duplication was present; **E:** Per base N Content is very low; **F** Most sequences are 37bp or 38bp long.

Data fragment size distribution was in line with what is predicted for successful ATAC-seq library generation (Figure 5.3A), however one sample (Cortex SD 2) had an unusual distribution, indicating that library preparation may not have been successful for this replicate. The number of duplicates was high (Figure 5.3B), most likely due to the small amount of starting material. However, most datasets contained at least 10M unique reads, with these evenly spread across the samples, and this was sufficient to carry out a full analysis on the data (Liu et al., 2014). The samples were subsequently mapped to the mouse genome, and read mapping analysis demonstrated that almost all reads mapped to the mm10 genome, with most samples displaying a high percentage of uniquely mapped reads (Figure 5.3C).

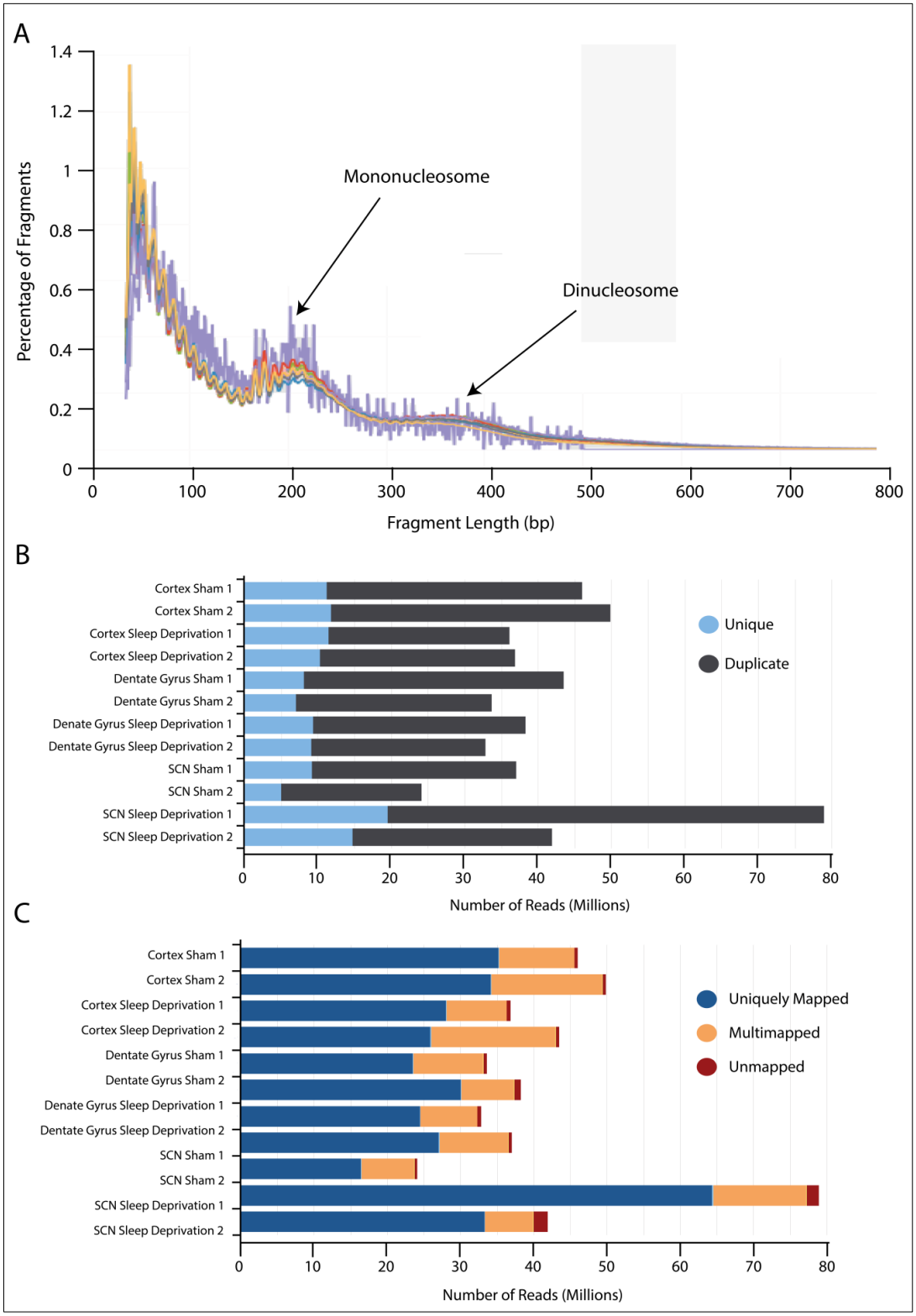


Figure 5.3: Quality control for fragment size and alignment. A: Fragment size distribution shows both a mononucleosome peak at 200bp and dinucleosome peak at 400bp; **B:** Number of duplicates is high, but most samples contain around 10M unique reads; **C:** Read mapping statistics to mm10 genome are high, with most reads mapping uniquely.

Principle component analysis was carried out to determine if samples cluster into their respective sample conditions and whether any outliers are present. For this experiment samples were expected to first cluster by brain region, and then by whether they had been sleep deprived or not.

PCA indicated samples clustered well by their tissue type, and showed further clustering by whether the mice were sleep deprived or not (Figure 5.4).

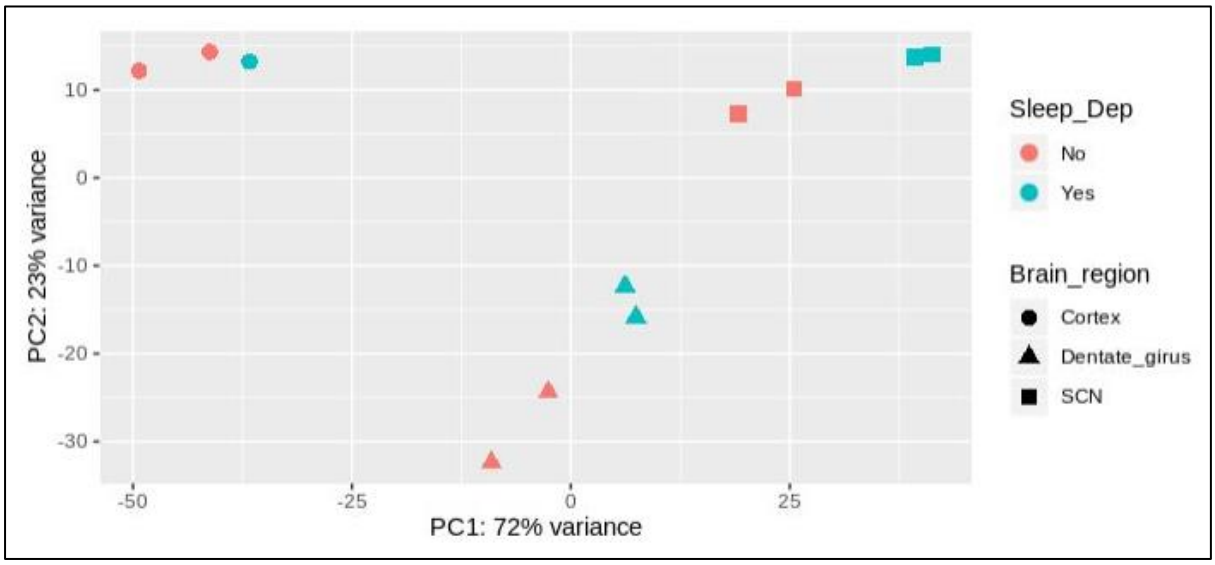


Figure 5.4: PCA plot for read counts from ATAC-seq SD experiment. Clustering by both tissue and experimental condition can be clearly seen, demonstrating successful dissection of different brain regions, and indicating SD had an effect on chromatin accessibility.

5.2.2: Differential analysis of ATAC-seq aligned reads

Following successful QC and read alignment to the mouse genome, to determine where chromatin areas had differential accessibility following SD, differential analysis of aligned reads was carried out using DESeq2. Analysis revealed both tissue specific chromatin states across the brain regions sampled, and statistically significant chromatin alterations in each region following SD (adjusted p-value < 0.05). This allowed the identification of several interesting candidate genes and molecular pathways for further downstream analysis.

The greatest variance in chromatin state was observed between the different brain regions (Figure 5.5), with numerous genes having nearby chromatin changes. This result is not unexpected; each tissue has a different chromatin landscape to control gene expression, and this allows each tissue to have a niche gene expression profile needed for that particular tissue/region to perform its specialised function (Rapkin et al., 2012). This result helps to demonstrate that ATAC-seq has successfully assayed chromatin structure, and provides a powerful insight into how genes can be regulated by their nearby cis-regulatory sequences being made inaccessible within chromatin, therefore switching off promoters and enhancers. For example, two genes *Chia* and *Jak2*, demonstrate substantial differences in chromatin accessibility across the different brain regions (Figure 5.8C and D).

The chromatin state of numerous genes changed following SD in the SCN, DG and cortex (Figure 5.6), with gene adjacent statistically significant differential chromatin (adjusted p-value < 0.05) primarily closing in the SCN and DG, and primarily opening in the cortex. Following differential analysis (Figure 5.7), chromatin near to 19 genes in the SCN, 28 genes in the DG, and 15 genes in the cortex showed a statistically significant change in accessibility (adjusted p-value < 0.05). Differential analysis prior to p-value correction revealed a greater number of genes (>200) across all brain regions with differential chromatin (unadjusted p-value < 0.05), and of these several genes already known to be affected by SD were identified. For example, chromatin near to the miR-17/92 cluster of microRNAs, which are thought to be important for the repair of DSBs (Xie et al., 2019) closed in both the SCN and DG (Figure 5.8A). Alongside, chromatin near to Cirbp, a known modulator of murine

molecular clock gene expression (Hoekstra et al., 2019) and Hspa1b, a heat shock gene induced by SD (Ackermann et al., 2013) appeared to close following SD in the DG.

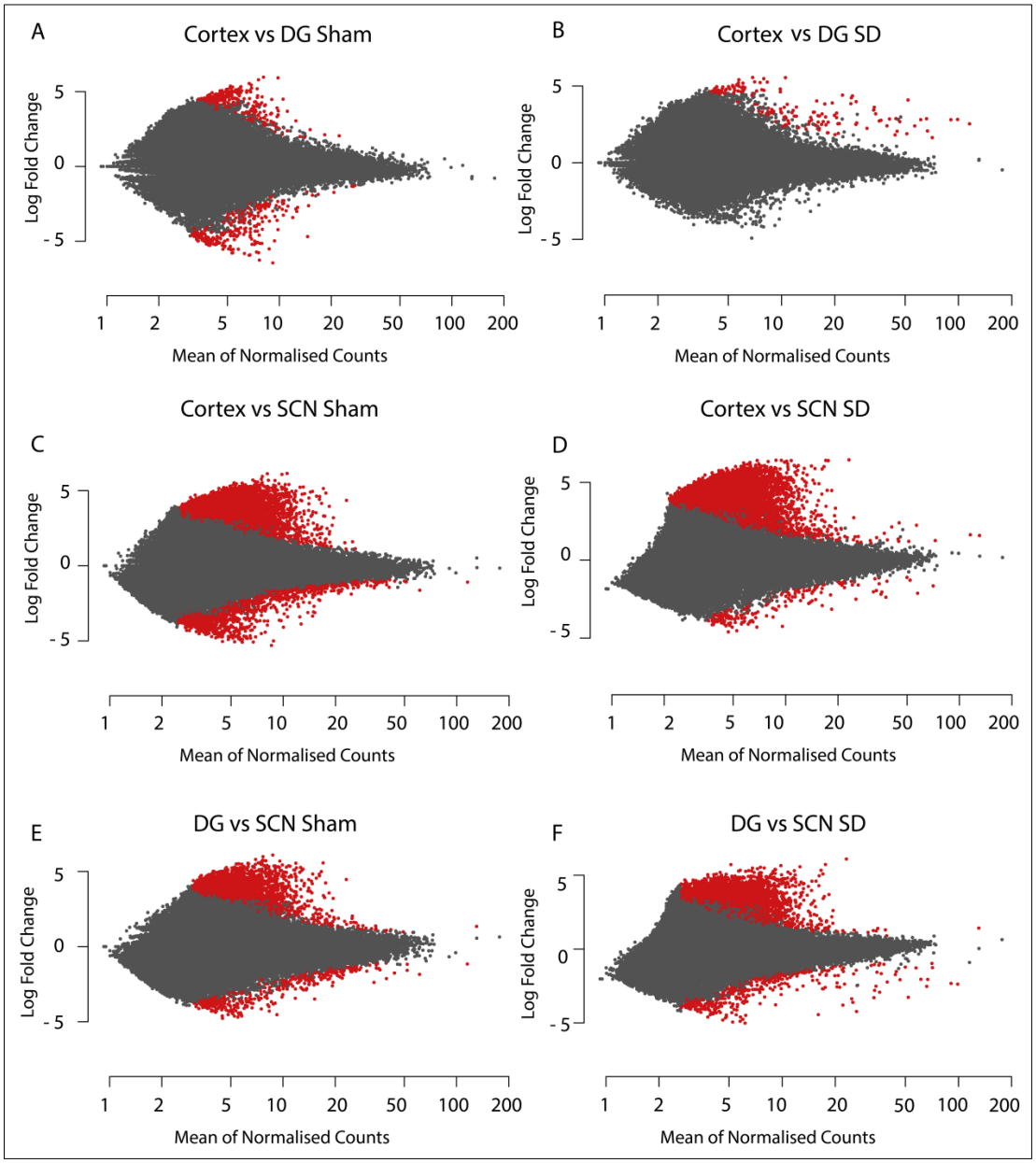


Figure 5.5: Bland–Altman plots across brain regions for SD ATAC-seq experiment. Dots represent genes which have differential chromatin nearby; statistically significant genes with an adjusted P-value below 0.05 are highlighted in red. **A:** Cortex vs. DG, sham; **B:** Cortex vs. DG, SD; **C:** Cortex vs. SCN, sham; **D:** Cortex vs. SCN, SD; **E:** DG vs. SCN, sham; **F:** DG vs. SCN, SD.

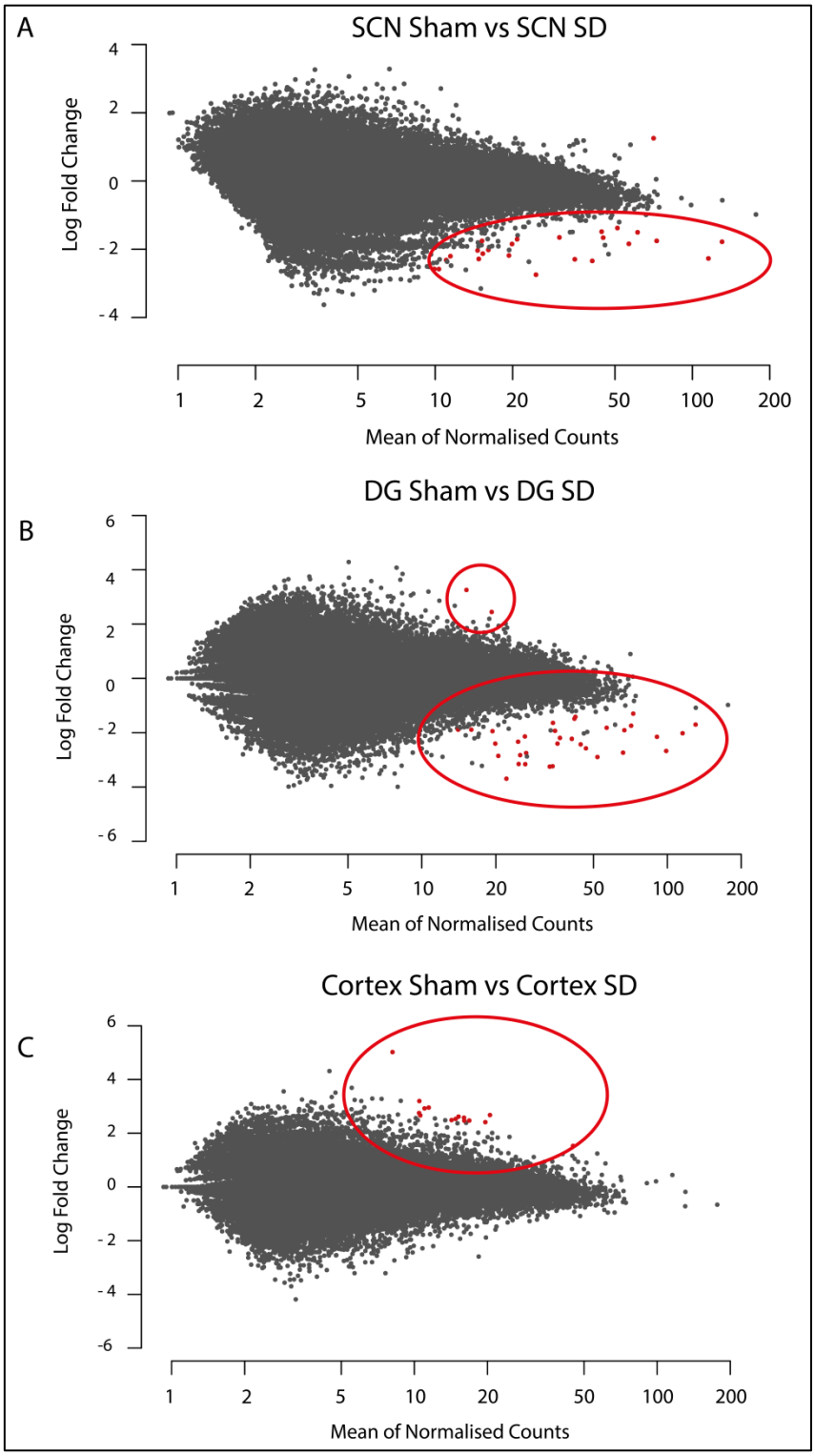


Figure 5.6: Bland–Altman plots for ATAC-seq SD differential analysis. Dots represent genes with differential chromatin accessibility nearby, statistically significant genes with an adjusted P-value below 0.05 are highlighted in red; **A:** The SCN showed some statistically significant genes adjacent to decreased chromatin accessibility; **B:** The DG has statistically significant genes with both increased and decreased chromatin accessibility nearby; **C:** The cortex showed some statistically genes with increased chromatin accessibility nearby.

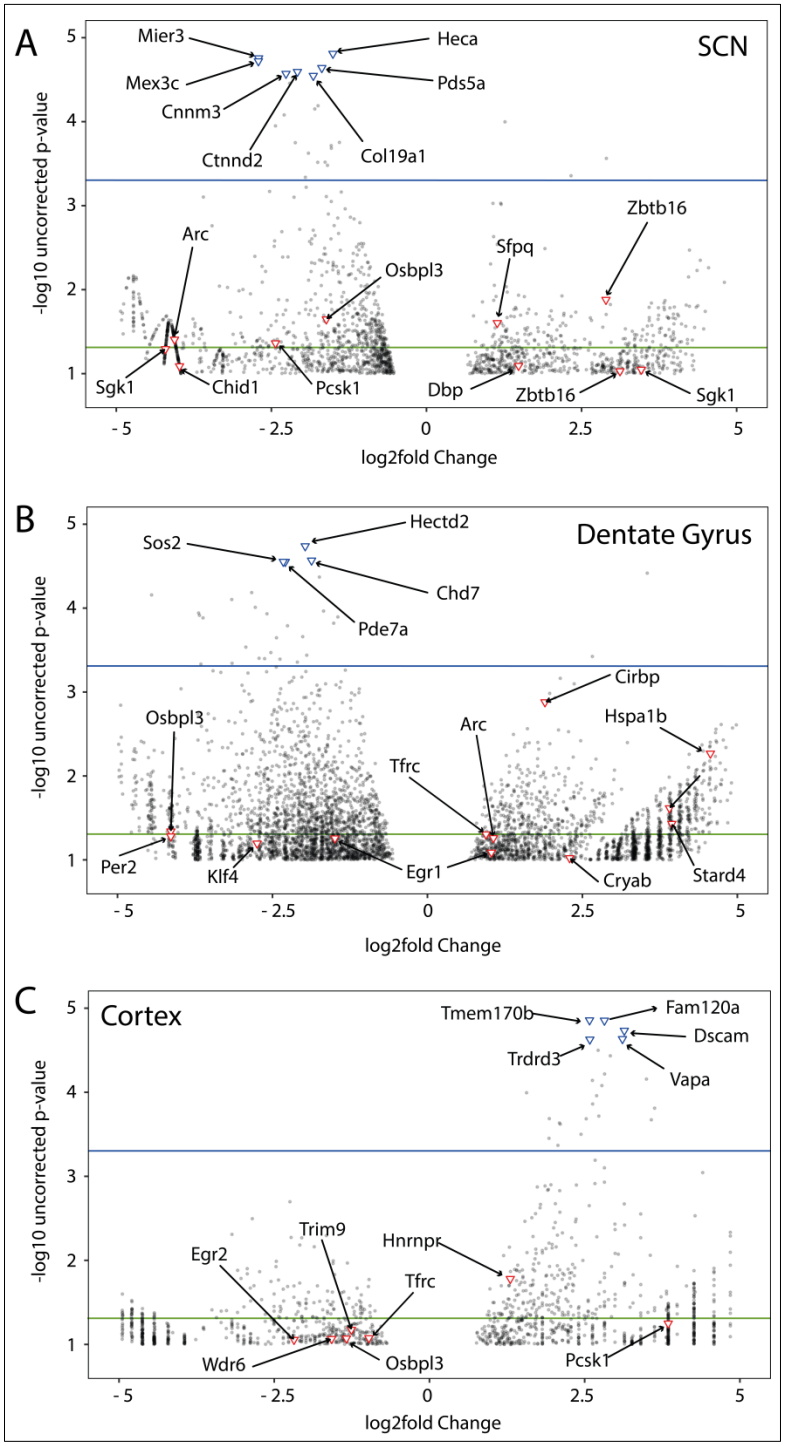


Figure 5.7: Volcano plots for areas of differential chromatin accessibility following SD. Points signify an area of differential chromatin, blue triangles indicate the most statistically significant areas of differential chromatin, red triangles show areas of differential chromatin near to genes known to be effected by SD, labels indicate the nearest gene to the differential chromatin, the green line is the uncorrected P-value threshold of 0.05, and the blue line is the adjusted P-value threshold of 0.05. **A:** SCN; 4 areas of differential chromatin above the unadjusted p-value threshold are near genes associated with SD. **B:** DG; 4 areas of differential chromatin above the unadjusted p-value threshold are near genes associated with SD, including *Cirbp*. **C:** Cortex; 1 area of differential chromatin above the unadjusted p-value threshold has nearby gene associated with SD.

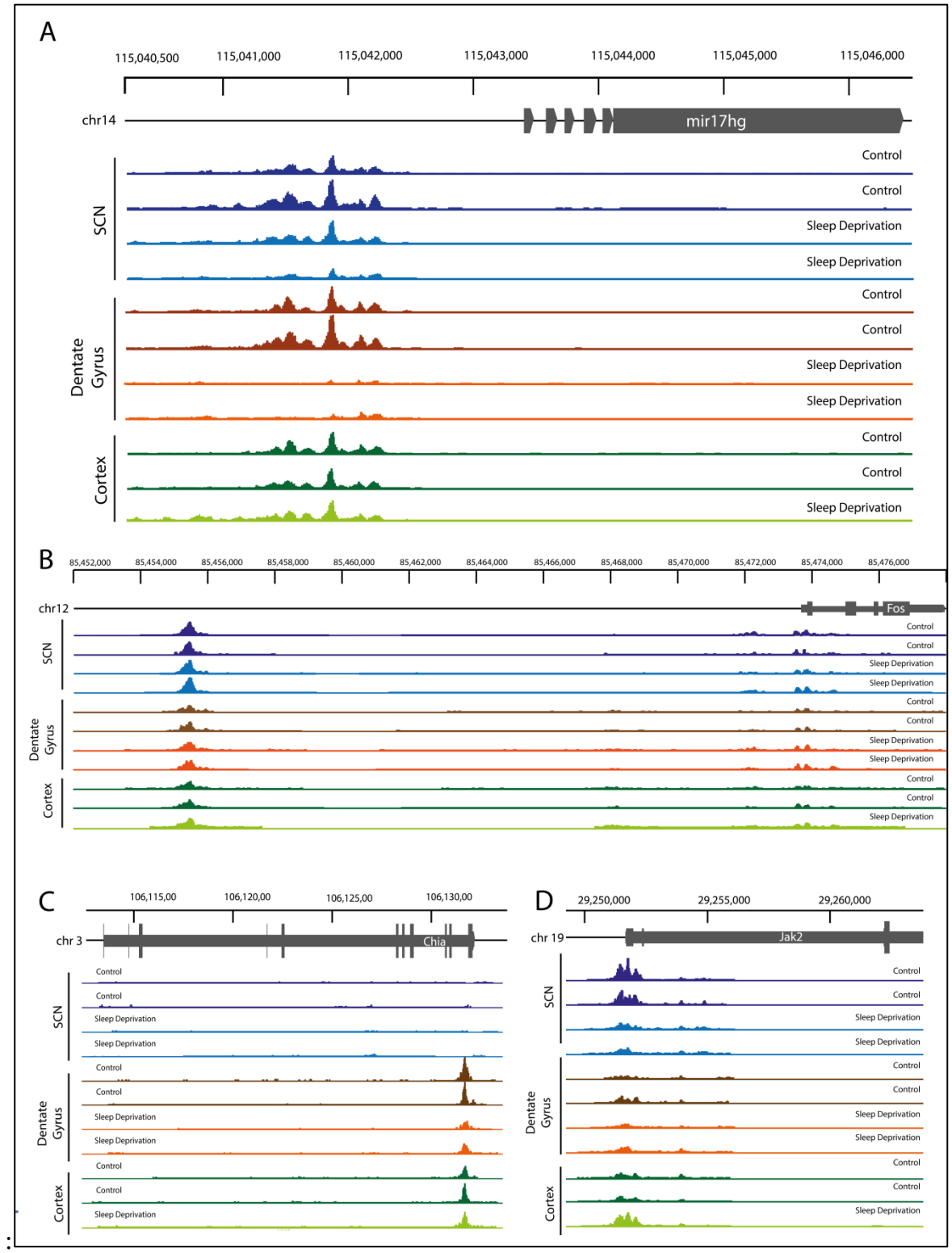


Figure 5.8: Read coverage tracks for ATAC-seq; A: Chromatin located upstream of *MiR17hg* appeared to close following SD in the DG and partially in the SCN, but remained open in the Cortex; **B:** Chromatin located near to *Fos* did not change following SD in all brain regions; **C:** Chromatin near to *Chia* was completely closed in the SCN under all conditions, but open in the DG and Cortex; **D:** Chromatin near to *Jak2* closed after SD in the SCN but opened in the Cortex, and was always closed in the DG.

5.2.3: Gene set enrichment analysis

To determine if genes with differential chromatin nearby following SD shared common functions, or were involved in similar biological pathways, GO biological process (GO BP) enrichment analysis was carried out for all genes having differential chromatin accessibility nearby (adjusted p-value < 0.2) for each brain region (Figure 5.9). A less stringent statistical threshold was used because this maximised the number of genes available for analysis, therefore covering a wider range of biology. This risked potentially including some false positive genes, but as GSEA uses groups of genes, this typically overcomes spurious genes that are included as a result of statistical noise.

Analysis revealed genes clustered into 3 groups for the SCN, and 5 for the DG and Cortex (Figure 5.10), with enriched GO terms including many that would be anticipated based on current literature regarding SD. GO terms clustering within the SCN notably included phosphorylation and gene regulation; previous studies have shown many proteins are known to be phosphorylated in SCN following SD (Ode and Ueda, 2020). GO terms clustering within the DG included metabolism and cell cycle; both of these have been demonstrated to be altered by SD (Guzmán-Marín et al., 2003), and indicate chromatin accessibility may play a role explaining these changes. Interestingly within the cortex GO terms associated with cellular transportation clustered, this may be neurons attempting to remove excess toxins and other metabolites following an extended period of SD (Eugene and Masiak, 2015).

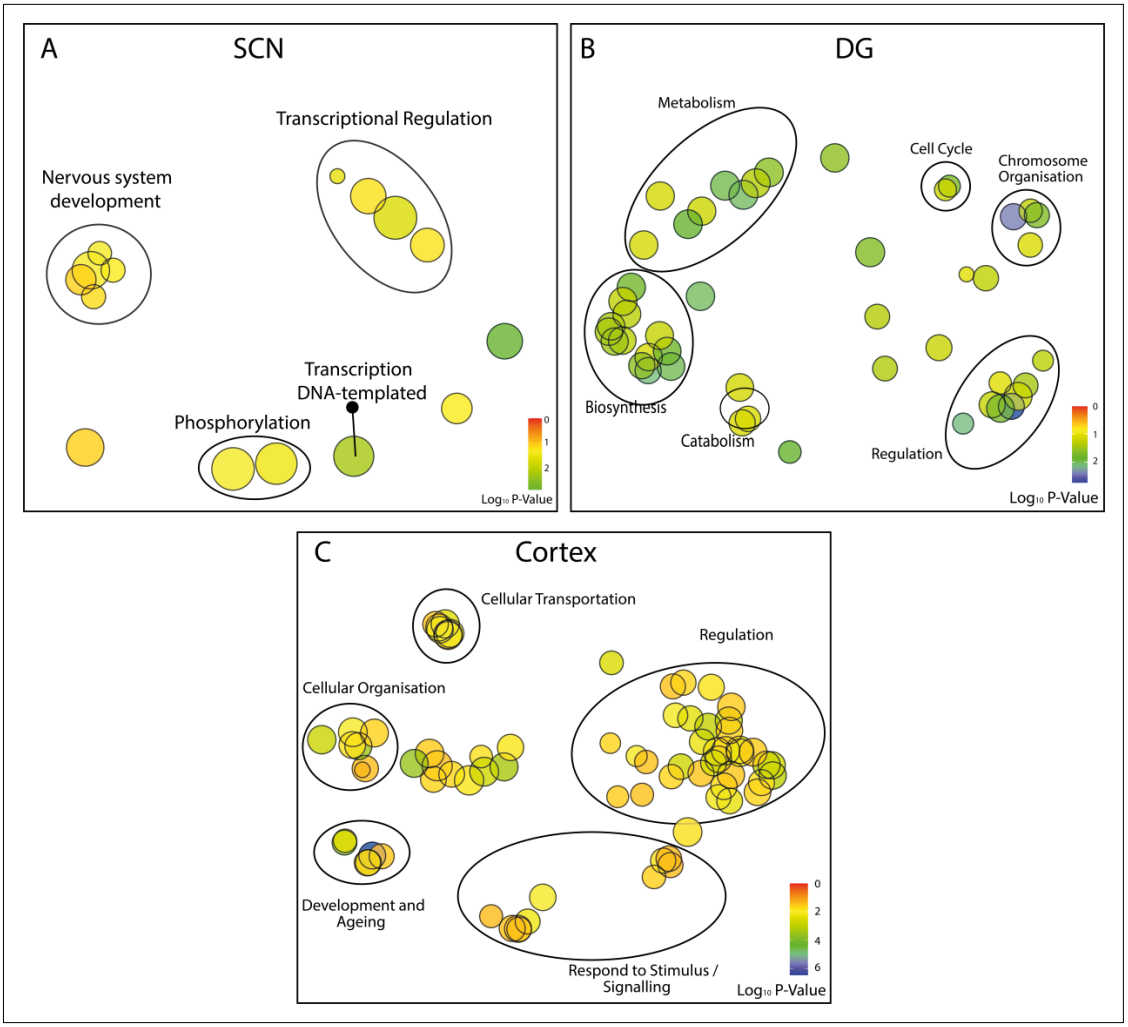


Figure 5.9: Clustering of GO BP terms for ATAC-seq experiment 1 visualised using Revigo. GO BP terms associated with genes with differential chromatin accessibility nearby can be seen clustering into functional groups, data visualised using Revigo; **A:** Three small clusters GO BP terms cluster for the SCN; **B:** A larger number GO terms clustered for the DG into 5 clusters; **C:** Many GO BP terms cluster for the Cortex into 5 clusters.

5.2.4: Motif Analysis

Analysis of the location of differential chromatin following SD revealed that differential chromatin was primarily located within 1kb of gene promoter regions across all tissues (Figure 5.9), with over 50% of differential chromatin being found in promoter regions for the SCN and cortex. This demonstrates that chromatin accessibility is changing dynamically at regions of the genome with the highest abundance of transcription factor motif binding sites, which will likely have a significant effect on transcription factor binding, and therefore downstream gene expression.

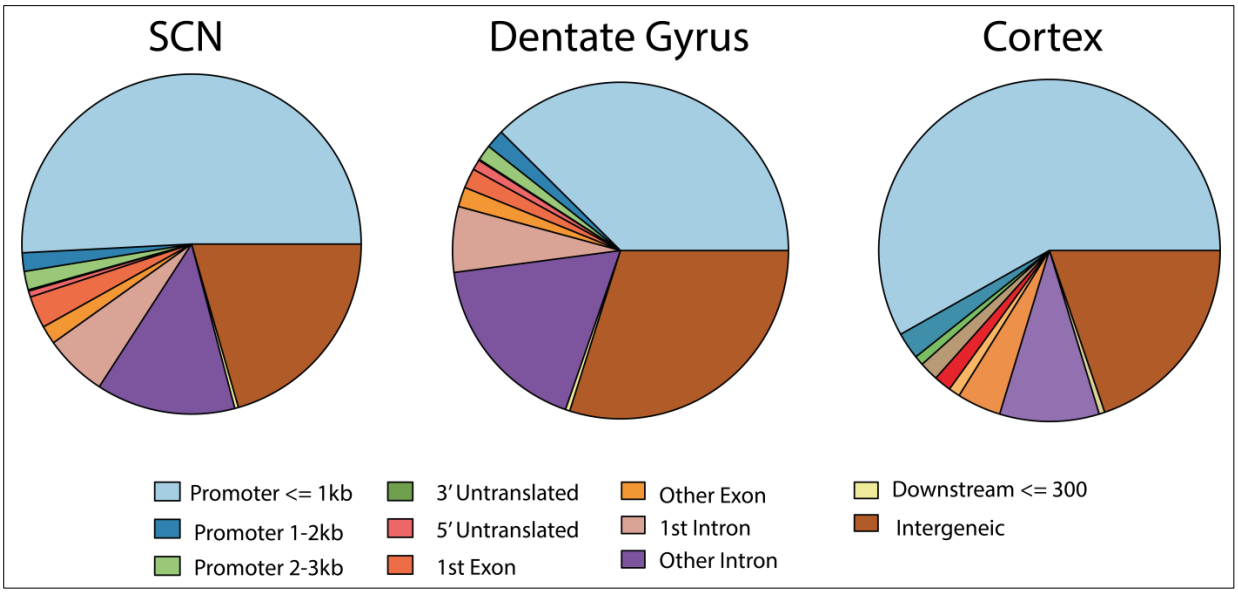


Figure 5.10: Location of differential chromatin within the genome following SD. Across all brain regions the majority of differential chromatin identified was found within 1kb of gene promoter regions.

To further investigate if SD was having an effect on transcription factor binding sites, motif analysis was carried out for all genomic areas containing differential chromatin (adjusted p-value < 0.2). A less stringent statistical threshold was again used to increase the number of genomic areas available for motif analysis. As this analysis carries out statistical testing for any enriched motifs, including less statistically significant areas of differential chromatin should not affect the validity of results obtained.

For the SCN 11 enriched motifs were identified (Table 5.2), containing 8 statistically significant motifs (p-value < 0.05). E2F7 and E2F3 were identified to be the most statistically significant motifs (p=1e-8 and 1e-7 respectively); these act as binding sites for E2F transcription factors including E2F7, a protein playing an important role in DNA repair (Mitxelena et al., 2018). Although less statistically significant (p-value 0.01), the binding motif for Myc was also identified as enriched in the SCN; Myc has been shown to disrupt the molecular clock in cancer cells (Altman et al., 2015), and despite not being statistically significant, its prior association with the molecular clock supports it playing a role in the SCN following SD.

Numerous enriched motifs (p-value < 0.05) were identified within differential chromatin following SD in the DG and cortex (Tables 5.3 and 5.4, respectively). E2F3 was again identified as an enriched motif in both these brain regions, further demonstrating the importance of DNA repair following extended periods of SD. Other enriched motifs in the DG included C/EBP whose TFs regulate cell growth and has been shown to up-regulate Per2 and NFATC2 whose TF contains binding motifs for molecular clock proteins (Gery et al., 2005) (Poulsen et al., 2021). Within the cortex enriched motifs included ARNT, whose consensus sequence is very similar to the core molecular clock binding motif ARNTL, but lacks the first two and last bases (Figure 5.11). Together, these findings may provide a link between SD and entrainment of the molecular clock within the cortex.

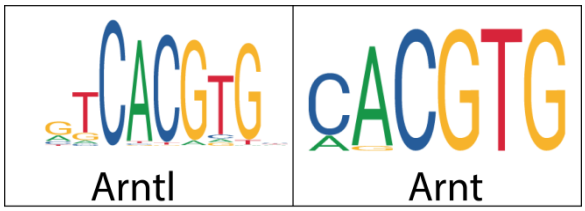


Figure 5.11: A comparison of the motif ARNT with ARNTL. ARNT was found to be an enriched motif in the Cortex following SD, its consensus sequence shows a high level of similarity to the core molecular clock binding motif ARNTL.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-8	-1.940e+01	15.15%	0.18%	E2F7(E2F)
	1e-7	-1.807e+01	36.36%	4.64%	E2F3_2
	1e-6	-1.509e+01	15.15%	0.42%	KLF7
	1e-5	-1.311e+01	39.39%	8.81%	Myb
	1e-5	-1.297e+01	18.18%	1.18%	ISL2
	1e-4	-1.076e+01	18.18%	1.75%	Gem_1
	1e-4	-9.674e+00	18.18%	2.13%	NFIX
	1e-3	-6.995e+00	3.03%	0.00%	SMAD
	1e-2	-6.302e+00	3.03%	0.01%	BHLHE22
	1e-2	-6.302e+00	3.03%	0.01%	Sox11
	1e-2	-4.518e+00	12.12%	2.65%	Myc

Table 5.2: Motifs enriched within differential chromatin identified in the SCN following SD. Enriched motifs include E2F7 and E2F3 which bind TFs important for DNA repair; MYC whose TFs include important oncogenes under circadian control; and KLF7 whose TFs play a critical role in neuronal morphogenesis and survival of sensory neurons.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-14	-3.307e+01	8.33%	0.02%	Hoxd13
	1e-10	-2.392e+01	8.33%	0.08%	Six3
	1e-10	-2.372e+01	9.72%	0.17%	E2F3
	1e-10	-2.370e+01	6.94%	0.03%	Arid5a
	1e-9	-2.181e+01	23.61%	3.47%	PRDM14 (Zf)
	1e-8	-1.878e+01	11.11%	0.55%	ZNF711 (Zf)
	1e-7	-1.766e+01	26.39%	5.85%	BARHL2
	1e-7	-1.764e+01	8.33%	0.23%	Oct6
	1e-7	-1.732e+01	8.33%	0.25%	NFATC2
	1e-7	-1.716e+01	6.94%	0.12%	Sox12
	1e-7	-1.694e+01	5.56%	0.05%	HOXC12
	1e-7	-1.641e+01	5.56%	0.05%	CEBP
	1e-6	-1.483e+01	6.94%	0.20%	SPIB
	1e-5	-1.317e+01	13.89%	2.02%	Gm397
	1e-4	-1.095e+01	9.72%	1.13%	Zfp161

Table 5.3: Motifs enriched within differential chromatin identified in the DG following SD. Enriched motifs again include E2EF; C/EBP whose TFs regulate cell growth and has been shown to up-regulate PER2; and NFATC2 whose TF contains binding motifs for molecular clock proteins.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-13	-3.065e+01	8.33%	0.03%	Hoxd13
	1e-10	-2.442e+01	6.94%	0.03%	Arid5s
	1e-10	-2.423e+01	18.06%	1.42%	MyoG
	1e-9	-2.273e+01	9.72%	0.20%	Zfx
	1e-9	-2.251e+01	8.33%	0.10%	Six3
	1e-9	-2.243e+01	9.72%	0.20%	E2F3
	1e-7	-1.715e+01	11.11%	0.68%	Oct6
	1e-7	-1.699e+01	19.44%	3.10%	RXRG
	1e-7	-1.625e+01	5.56%	0.05%	NFIA
	1e-6	-1.610e+01	5.56%	0.06%	CEBP
	1e-6	-1.595e+01	5.56%	0.06%	HOXC12
	1e-6	-1.518e+01	18.06%	3.05%	POLO10
	1e-6	-1.482e+01	12.50%	1.27%	Zfp161
	1e-5	-1.331e+01	4.17%	0.03%	Arnt
	1e-5	-1.314e+01	25.00%	7.08%	TBR1

Table 5.4: Motifs enriched within differential chromatin identified in the cortex following SD. Enriched motifs include E2F, and ARNT whose consensus sequence is highly similar to ARNTL, a core molecular clock TF binding site.

ATAC-Sequencing Sleep Deprivation Repeat and Covid-19

Due to the low number of replicates within this experiment, and as a consequence the low number of statistically significant (adjusted P-value < 0.05) differential genes identified, a follow up repeat of the experiment was carried in an attempt to further validate the results obtained, and identify additional genes with differential chromatin following SD. However, working under the current Covid-19 social distancing restrictions, coupled with the complexity of the experiment proved to be far too difficult to repeat. As a result, the data quality generated was far inferior to the first experiment, and proved to be unsuitable for differential analysis. Viewing read coverage across the genome following alignment showed only a small number of aligned reads, even areas where chromatin had previously been shown to be open such as near to *Fos*, and *MiR17hg* (Figure 5.11). The pattern of low reads was in stark contrast to the first ATAC-seq experiment, when compared to Figure 6.8, it clearly suggests that ATAC-seq library preparation was unsuccessful this time.

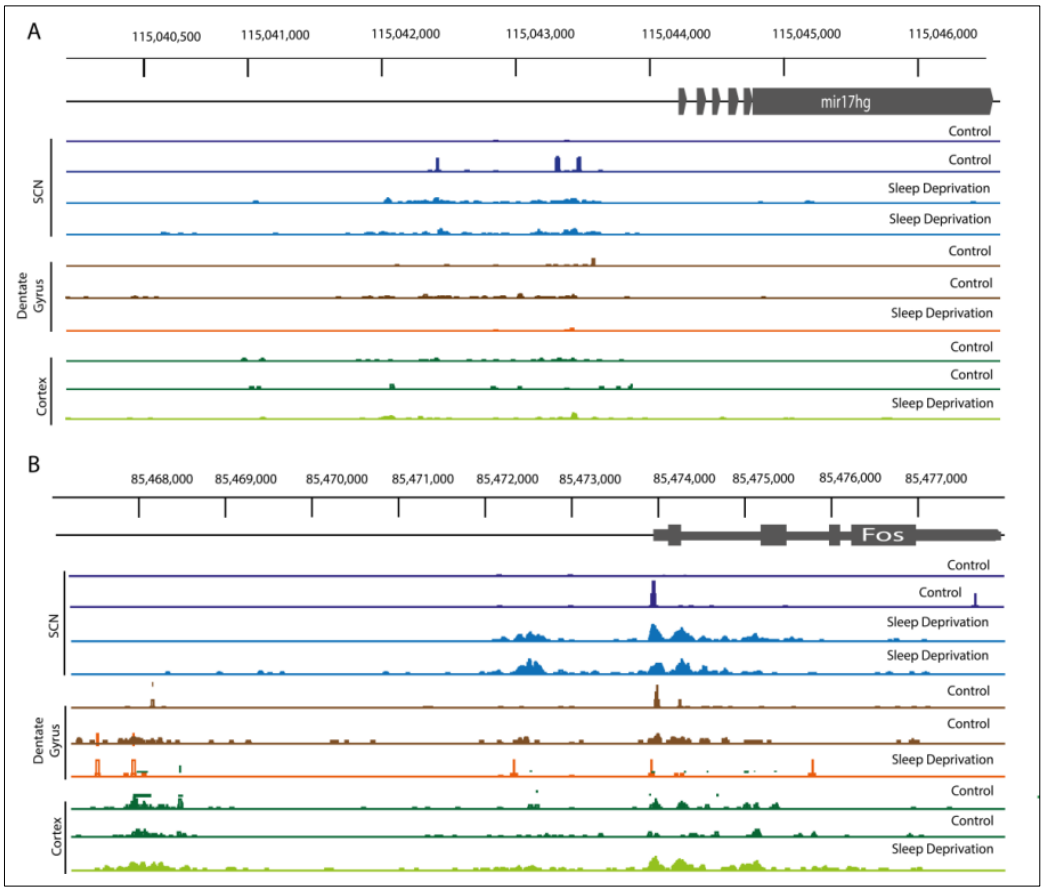


Figure 5.11: Read coverage tracks for the second ATAC-seq following SD. **A:** Chromatin accessibility appears very low close to *MiR17hg* with very few reads being identified across all sample conditions and tissues, and **B:** *Fos*, a far smaller number of reads are identified likely indicating poor ATAC-seq library preparation because chromatin accessibility in regions near *Fos* has always previously been open.

5.3: Re-analysis of published cortex ATAC-seq data

Following closure of all laboratory space due to the coronavirus pandemic; a thorough analysis of published cortex ATAC-seq data from Hor et al. (2019) was carried out. In their experiment the effects of 6 hours of SD on chromatin accessibility within the murine cortex was investigated over a 3 day period, with SD being carried out between ZT0 and ZT6 following a baseline of 24 hours. Although my ATAC-seq experiment focussed primarily on the more immediate effects of SD and was conducted in CT, by focussing on the time points within their data closest to mine, comparison with this dataset assists in the validation of my findings for ATAC-seq carried out in the cortex following SD.

The initial SD ATAC-seq experiment generated some useful data, but it lacked replicates, and only a small number of genes with differential chromatin were identified; this is in contrast to the published data, where a larger number of genes with differential chromatin close by were identified in the murine cortex following SD, and presents a small discrepancy with their data. To further investigate this, cortex data from the published dataset for mice culled at ZT6 immediately following SD, and data for mice culled at ZT6 without SD was analysed using the full in depth bioinformatic analysis described in methods section 2.11.

Following initial re-analysis of data, results appeared consistent with those which had been published, as PCA analysis following differential analysis showed similar clustering to the published PCA (Figure 5.12). Samples generally clustered by the ZT samples were collected, and by whether samples were from SD or control animals.

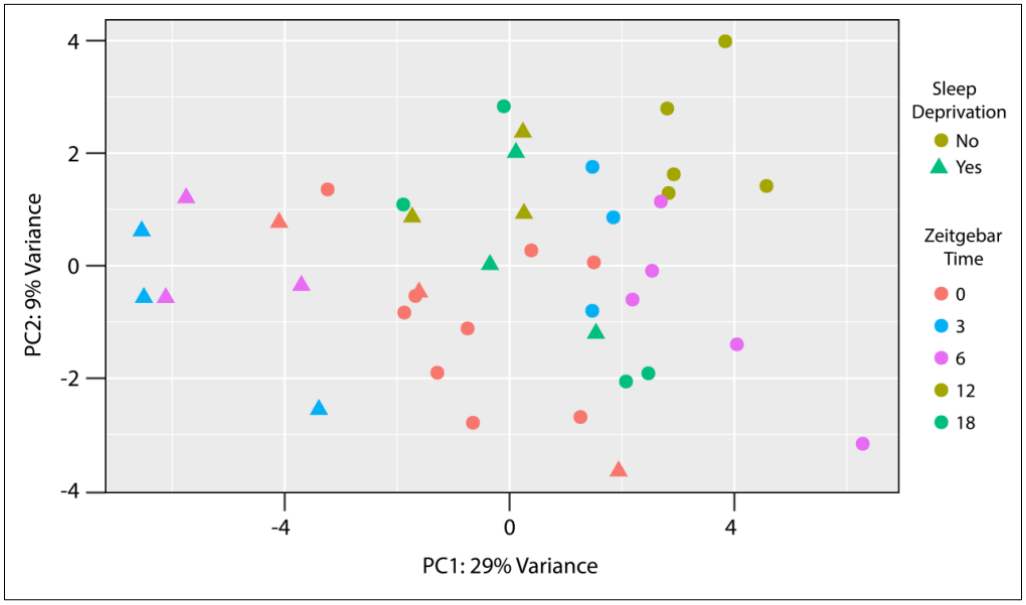


Figure 5.12: PCA plot for ATAC-seq data from Hor et al. (2019). Samples show some clustering by the ZT samples were collected and by whether samples were SD or not.

Differential analysis was carried out on the published data in an attempt to ascertain why their experiment found many more areas of differential chromatin within the cortex following SD than my own experiment. Re-analysis of the published data was in line with their published results, and therefore in the cortex a large number of genes with differential chromatin nearby were identified following SD. When compared to my own data the published data did indeed identify more areas of statistically significant differential chromatin following SD in the cortex, but in agreement with my own data, more areas of gained open chromatin were identified in both datasets (Figure 5.13).

Curiously when genes located near to statistically significant areas of chromatin were compared across both datasets, none were found in both sets. Yet, when the locations of aligned reads from the two datasets was viewed using the IGB, the broad pattern of read alignment appeared very similar, indicating chromatin accessibility was similar across the two datasets. Despite this, certain peak regions which appeared in the same location across both datasets showed a different alignment pattern within the peak itself. For example chromatin near to *MiR17hg* remained open in both datasets whether mice were SD or not, but the reads themselves align slightly differently, demonstrating differential accessibility within the peak (Figure 5.14).

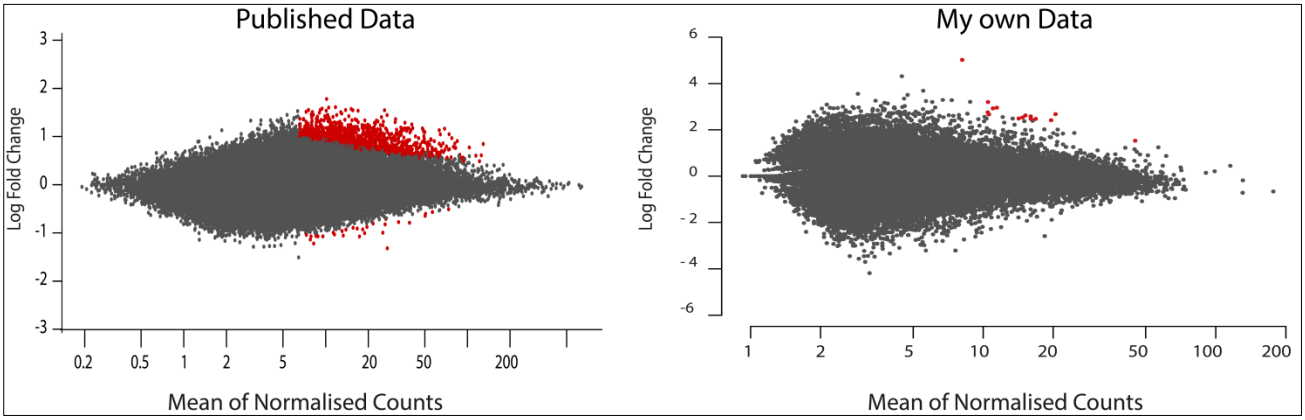


Figure 5.13: Bland–Altman plot for differential genes comparing published data with my own. Dots show genes with differential chromatin nearby following SD, those with an adjusted P-value below 0.05 are highlighted in red. In the published data a large number of statistically significant genes can be seen with chromatin showing a gain of accessibility, while in my own only a smaller number were identified.

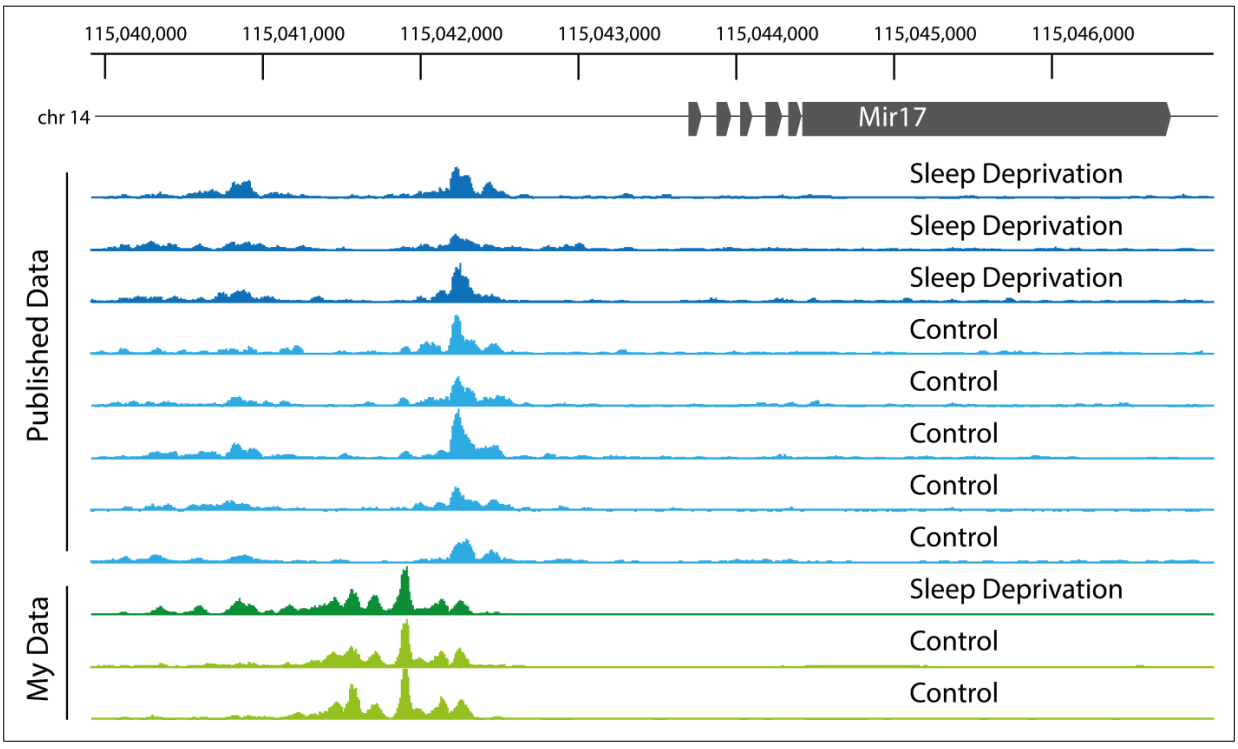


Figure 5.14: Read coverage tracks near to *MiR17hg*. Data from Hor et al. (2019) at ZT6 and data from ATAC-seq experiment 1, for the cortex at CT6, open chromatin can be seen near to *MiR17hg* across both samples regardless of whether mice were SD or not, but the reads themselves align slightly differently.

To determine if the location of differential chromatin is near to genes that share similar functions to those identified close to differential chromatin in my own data, GSEA was carried out on the differential gene lists generated from the published dataset. As a result of the far greater number of genes identified close to differential chromatin in the published dataset, gene enrichment analysis yielded far more enriched gene ontologies, therefore giving a further insight into the effects of SD chromatin structure in the cortex. Notwithstanding the large number of GO terms which clustered, the GO terms clustered into similar groups to those identified from my own dataset, and despite the large number of GO terms identified, no further groups of clusters were obtained. For example, regulation, response to stimulus and transport remained the largest clusters of GO BP terms, and transmembrane activity and receptor activity were the largest clusters of GO MF terms (Figure 5.15). This analysis provides good support for the validity of my own dataset, and continues to sustain the premise that regulation of gene expression changes following SD, and neurons attempt to remove excess toxins and other metabolites following extended periods of SD.

Overall analysis of published data from Hor et al. (2019) supports conclusions drawn from my own dataset, with the majority of statistically significant chromatin showing a gain of accessibility. Downstream GSEA of genes located close to differential chromatin within the published data revealed clustering of GO terms similar to those identified from my own dataset, providing further support for the validity of my data. Despite identifying far fewer areas of differential chromatin from my own data, this analysis demonstrates this discrepancy is likely due to the limitations of differential analysis imposed by only having two replicates, rather than the experiment itself being flawed.

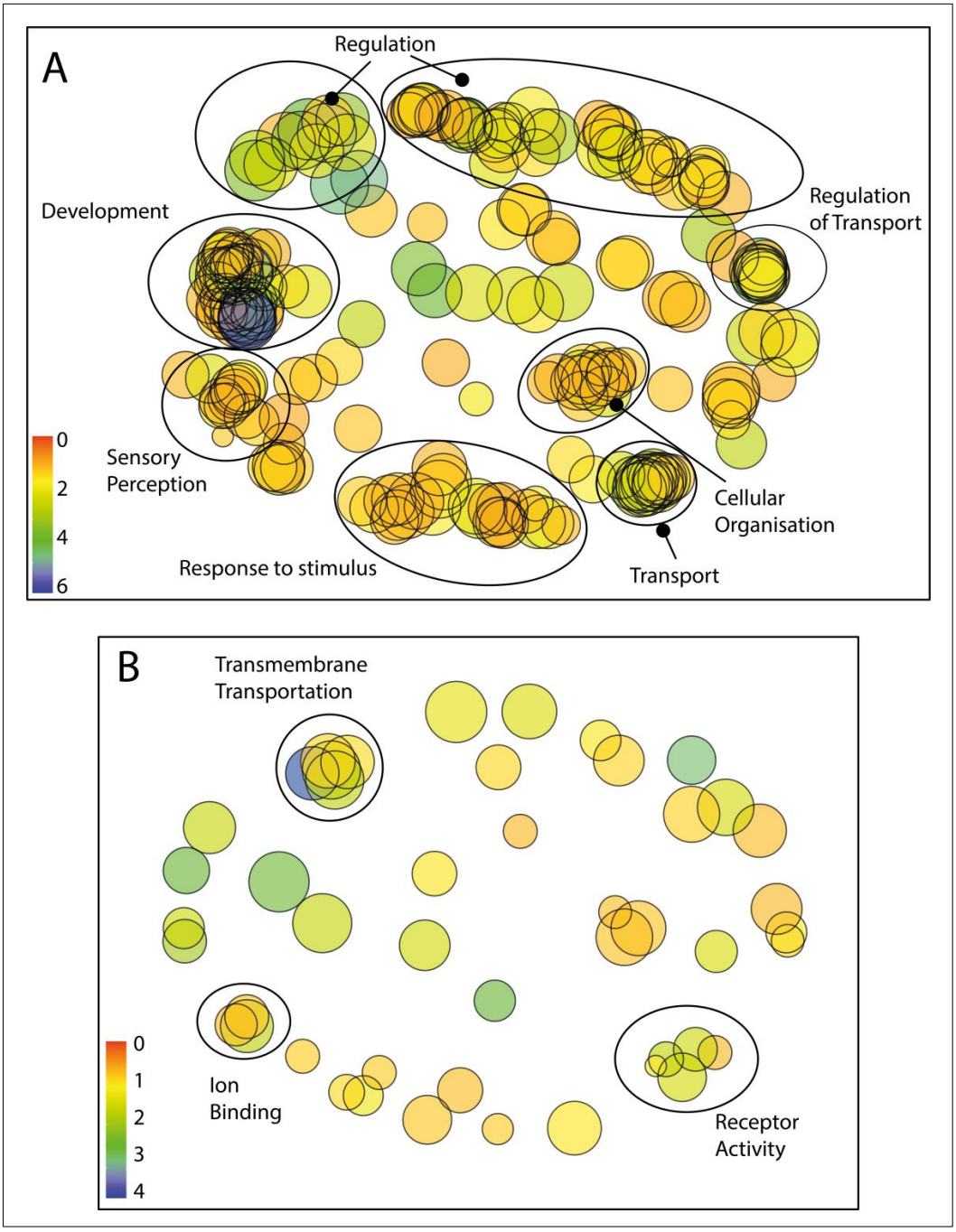


Figure 5.15: Clustering of GO terms for Hor et al. (2019) visualised using Revigo. GO terms associated with genes with differential chromatin accessibility nearby can be seen clustering into functional groups, data visualised using Revigo; **A:** GO BF, a large number of GO terms cluster into 7 groups; **B:** GO Molecular Function, fewer GO terms cluster in 3 groups.

5.4: The effect of acute SD on the transcriptome of the SCN, DG and cortex

Following profiling of the chromatin landscape following acute SD, RNA-seq was employed to investigate the transcriptome within the SCN, DG, and Cortex following SD, to investigate if alterations in chromatin state could be connected to gene expression changes. Using similar animal protocols as for ATAC-seq, mice were sleep deprived for 6 hours (control animals (sham) were not sleep deprived), and then brain punches taken from the different brain regions. Four replicates were collected for each sample condition, with each replicate containing tissue from two mice. Table 5.8 summarises the experimental groups. Total RNA was then isolated and RNA-seq libraries prepared as described in section 2.3, quantified as outlined in section 2.4, sequenced as described in section 2.5, and the resulting data analysed as set out in section 2.10.

Group Name	Tissue	Sleep Deprived
SCN-SD	SCN	Yes
SCN-SHAM	SCN	No
DG-SD	DENTATE GYRUS	Yes
DG-SHAM	DENTATE GYRUS	No
Cortex-SD	CORTEX	Yes
Cortex-SHAM	CORTEX	No

Table 5.5: The experimental groups for the RNA-seq SD experiment. Each sample condition comprised 4 replicates.

5.4.1: Quality Control

The samples passed numerous quality control testing; duplication levels were greatly improved for RNA-seq, and initial read count analysis across different sample groups appeared fairly even, indicating good quality libraries and a successful dilution before pooling (Figure 5.17). One DG control and one DG SD sample did not sequence.

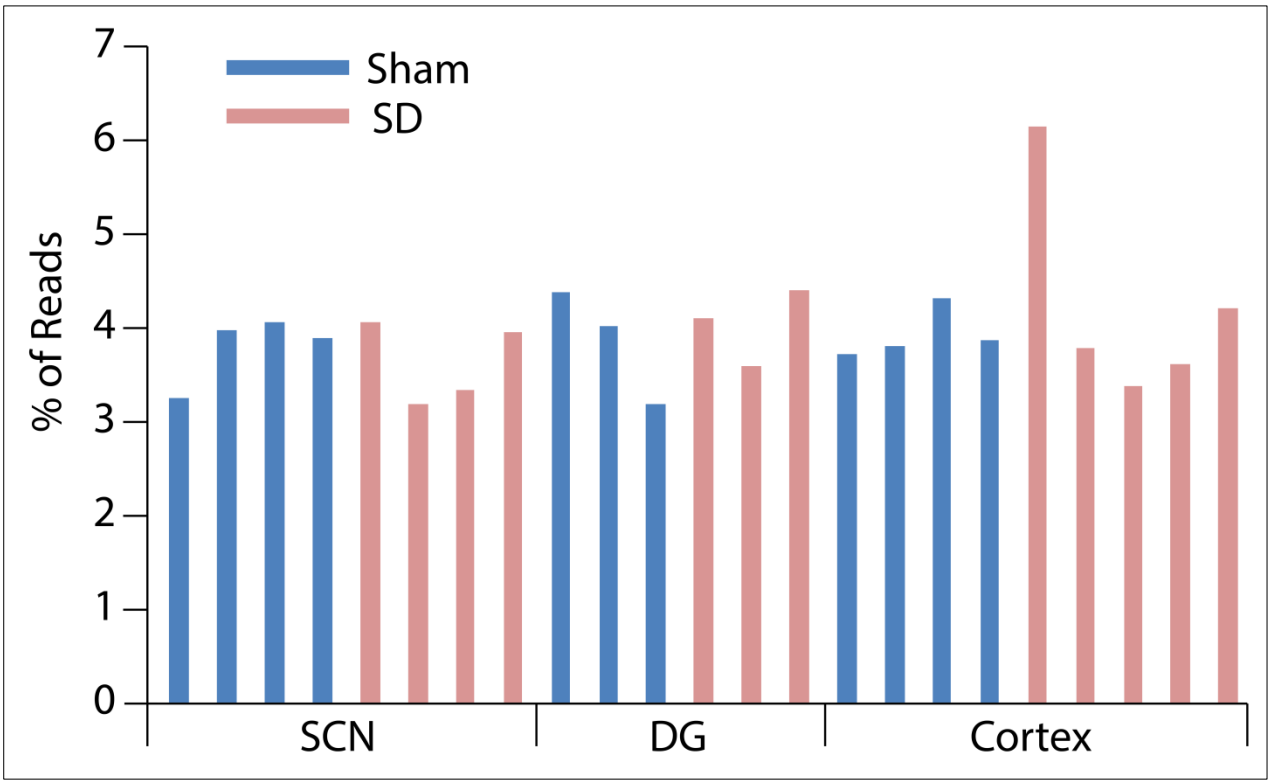


Figure 5.16: Read distribution across the different RNA-seq samples. An even distribution can be seen, indicating good library preparation, and consistent dilution of samples prior to sequencing.

Sequencing quality was assessed using FastQc, all tests indicated data to be of high quality. Overall read quality was high, with all samples displaying good per sequence (Figure 17.2A) and per base (Figure 17.2B) quality scores. Both GC content and levels of duplication appeared excellent (Figure 5.17–C/D). Per base N Content is excellent, indicating the sequencer successfully discerned the base at each point in the sequence (Figure 5.17-E). Read length distribution shows most sequences are above 70bp as expected due to a 150 cycle cartridge being used for paired end sequencing (5.17-F).

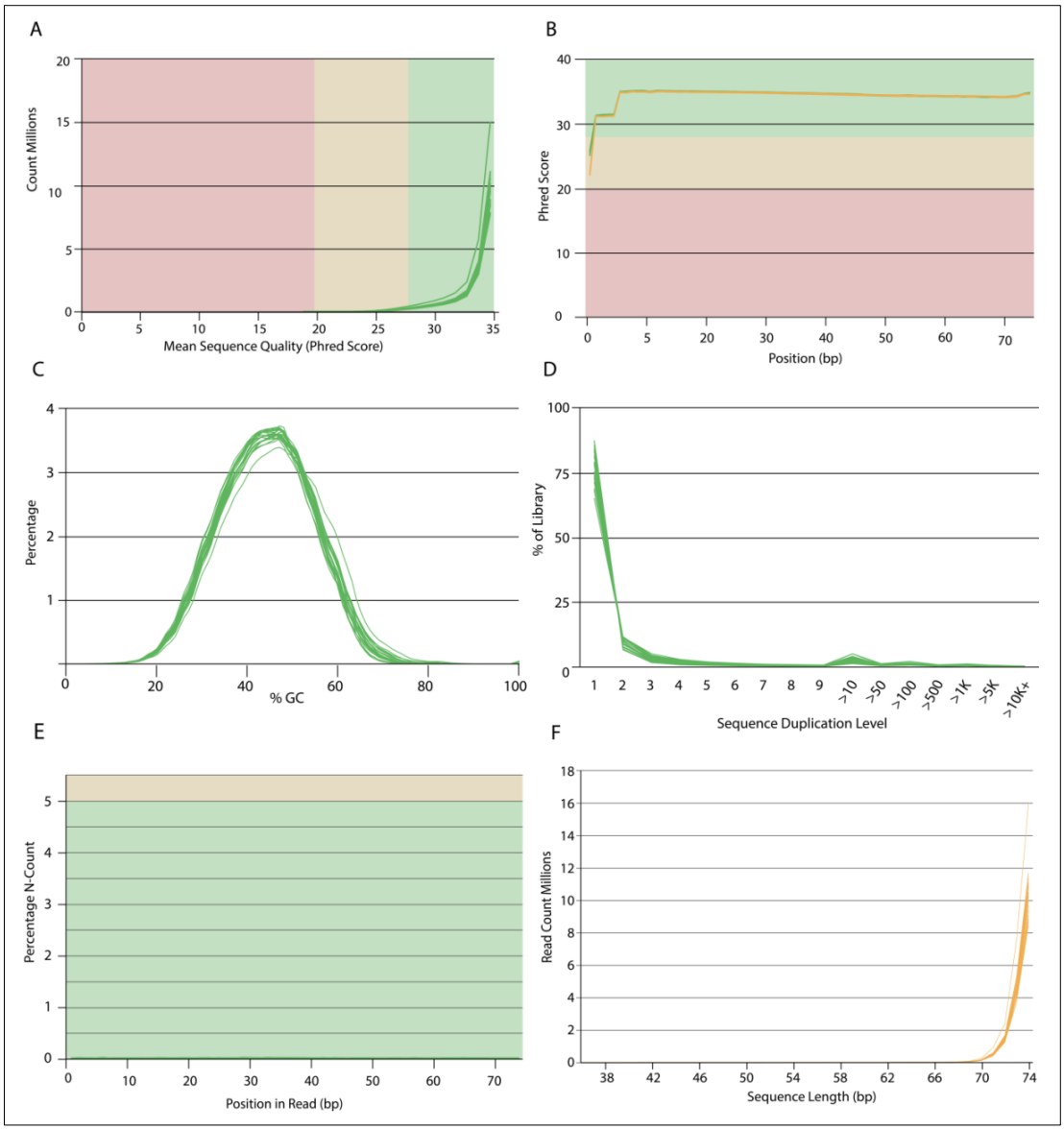


Figure 5.17: FastQc results for SD ATAC-seq. A: Per sequence quality is high; **B:** Mean quality score is fairly high; **C:** There is little GC bias; **D:** Sequence duplication levels are low; **E:** Per base N Content is very low; **F** Most sequences are greater than 70bp long.

Read counts across all samples gave a high number of reads per sample (Figure 5.18-A), and despite multiplexing 23 different libraries, the library with the lowest depth still contained 12 million non-duplicated reads. Following genome alignment, all samples aligned at over 95% to the mouse mm10 genome, with almost all reads aligning uniquely (Figure 5.18-B).

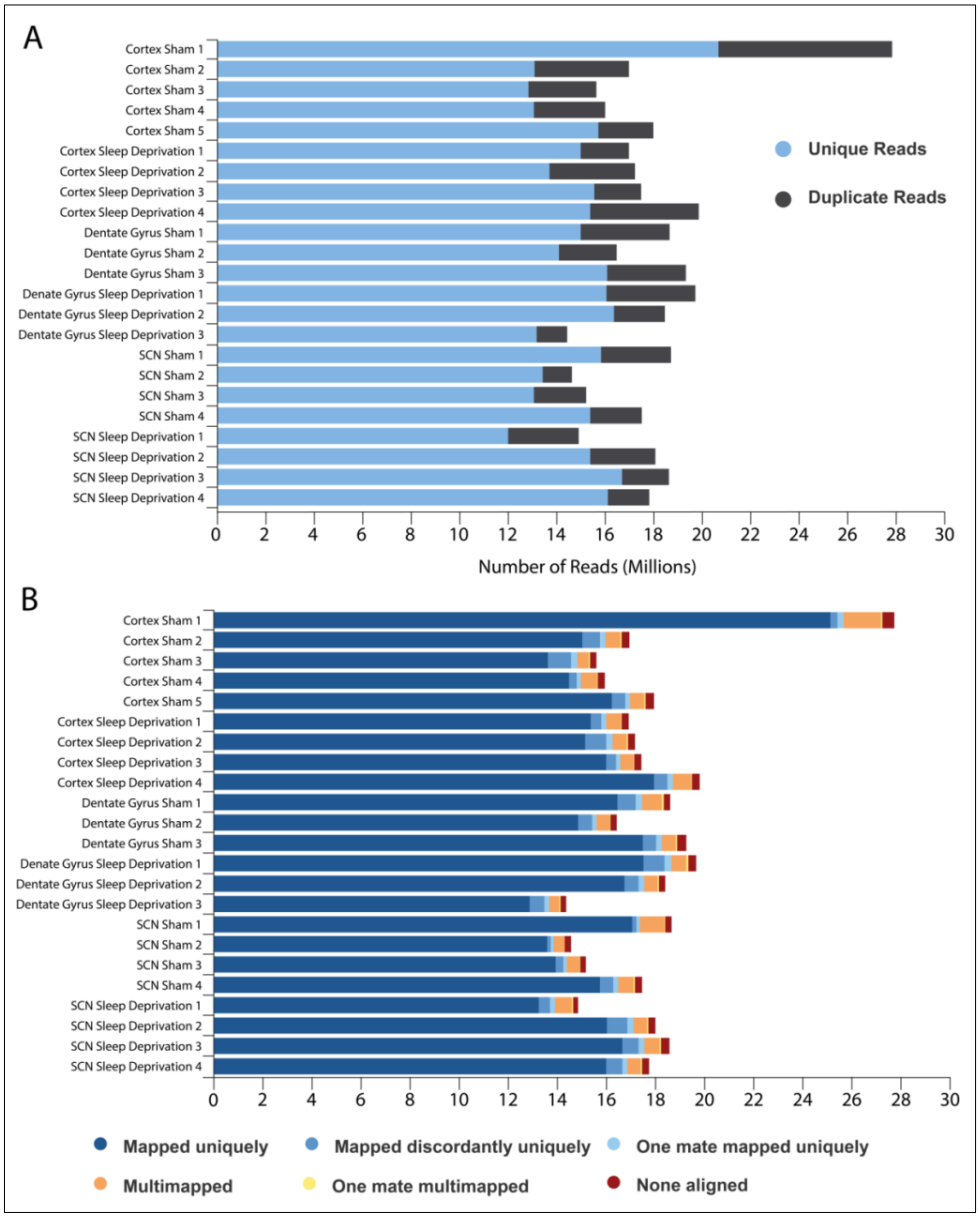


Figure 5.18: Quality control for fragment size and alignment A: Number of duplicates was low, despite multiplexing multiple samples, all libraries had at least 12M reads; **B:** Read mapping statistics to mm10 genome were excellent, with most reads mapping uniquely.

Principle component analysis was carried out to determine if samples cluster into their respective sample conditions and whether any outliers are present. For this experiment samples were again expected to first cluster by brain region, and then by whether they had been sleep deprived or not.

Following PCA analysis, samples clustered primarily by tissue type and to a lesser extent by SD (Figure 5.20). This indicates the correct brain regions were dissected, and SD likely had an effect on gene expression.

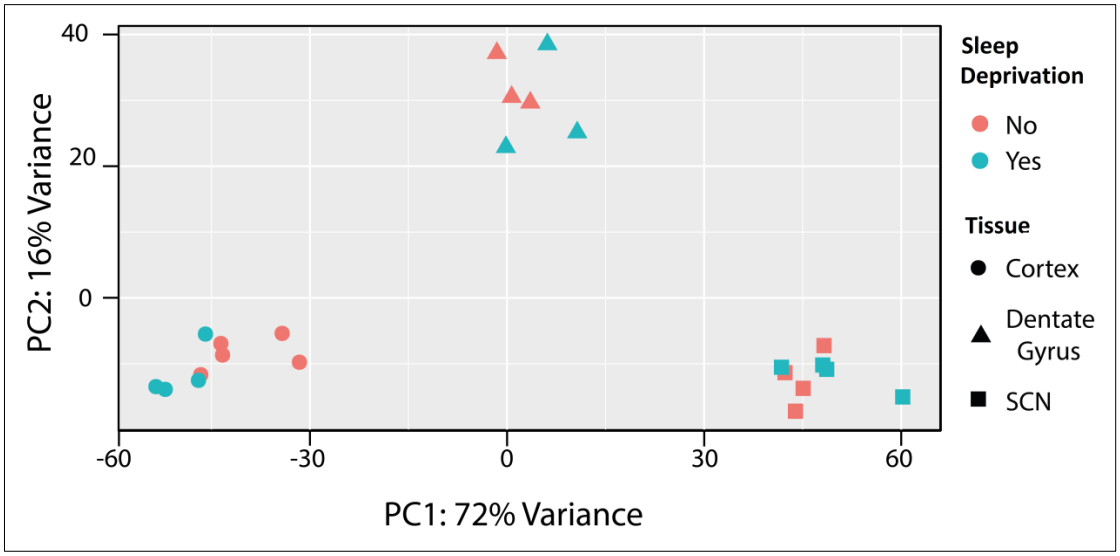


Figure 5.19: PCA plot for read counts from RNA-seq SD experiment. Clustering by both tissue and experimental condition can be seen.

5.4.2: Differential analysis of aligned reads for RNA-seq

Following successful QC and read alignment to the mouse genome, to determine which genes were DE following SD, differential analysis of aligned reads was carried out using DESeq2. Multiple genes across all brain regions were identified whose gene expression had altered following SD (Figure 5.20). Expression of 33 genes in the SCN, 9 in the DG and 147 in the cortex showed a statistically significant change following SD (adjusted P-value < 0.05).

The number of statistically DE genes identified initially appeared small in the SCN and DG, this was likely due to the read depth across the samples following multiplexing 23 samples in one sequencing run. Although read depth passed the required threshold for differential analysis, a higher depth would have improved the power of statistical analysis, and resulted in more statistically significant differential genes being identified. Relaxing the stringency (uncorrected P-value < 0.05) revealed many more differential genes (Figure 5.22), and when viewed in the IGB the majority of these genes were visibly DE. Within the DE genes identified, many were already known to change their expression following SD, including *Arc* and *Homer* in the DG and cortex (Figure 5.23). Both these genes are up regulated in the cortex following SD, giving confirmation of a successful experiment. Interestingly RNA-seq revealed that neither of these genes were expressed in the SCN, demonstrating brain region specific expression of genes related with SD.

In the SCN, *JunB* was shown to be up regulated following SD, and this is particularly notable as JUNB binds AP-1 and CRE motifs (Ryseck and Bravo, 1991), which are important for entrainment of the molecular clock within the SCN. This therefore discerns a further link between SD and entrainment within the SCN. In both the SCN and DG, *Cirbp* was shown to up regulated following SD, with this being of keen relevance, not only due to the action of Cirbp on the molecular clock already described, but also because the chromatin located near to Cirbp appeared to close following SD in the DG. *Egr2*, another gene known to change following SD (Gerashchenko et al., 2017) was also down regulated in the DG, further demonstrating the robustness of this dataset.

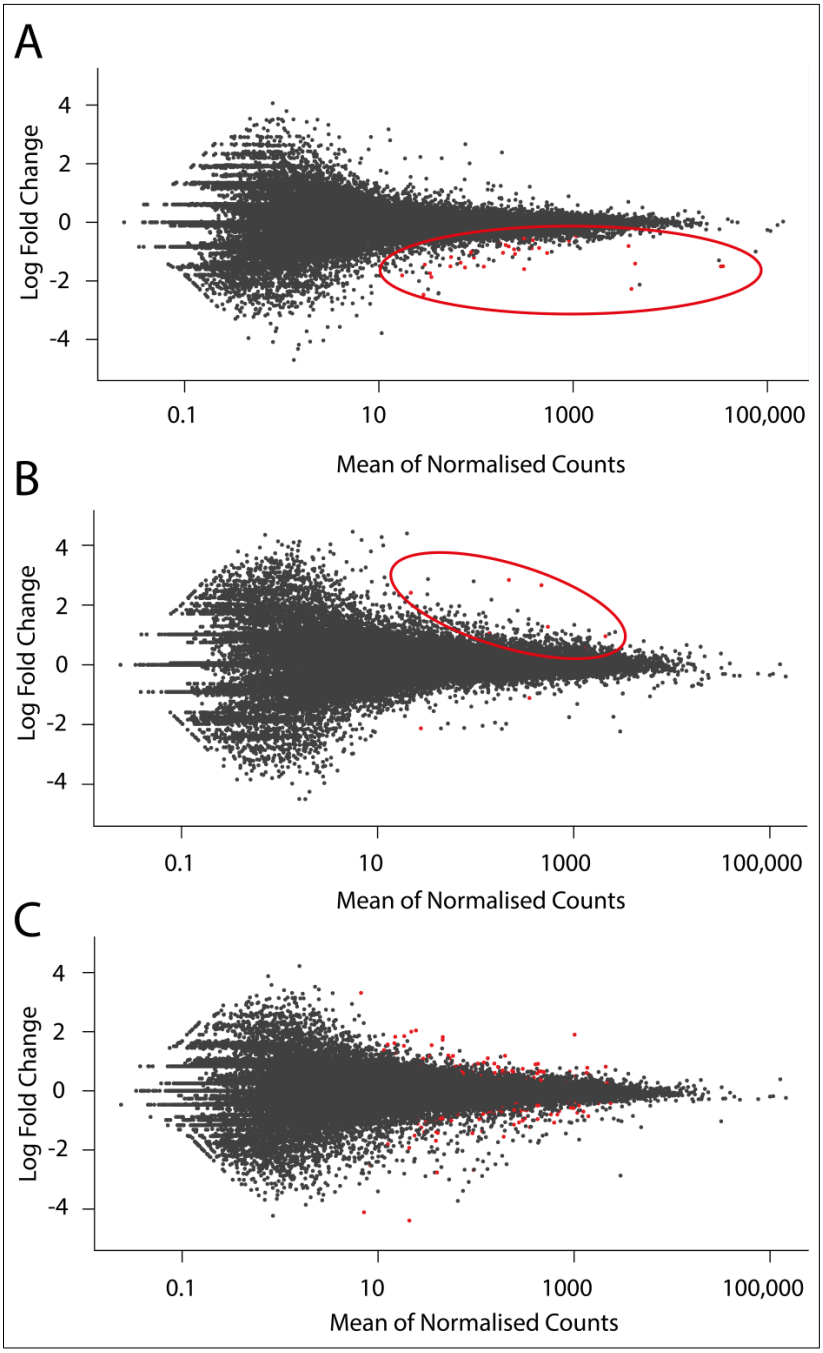


Figure 5.20: Bland–Altman plots for RNA-seq sleep DE analysis. Dots represent genes which were DE following SD; statistically significant genes with an adjusted P-value below 0.05 are highlighted in red.; **A:** SCN; DE statistically significant genes were primarily down regulated; **B:** DG; DE statistically significant genes were primarily up regulated; **C**: Cortex; statistically significant genes were both up and down regulated.

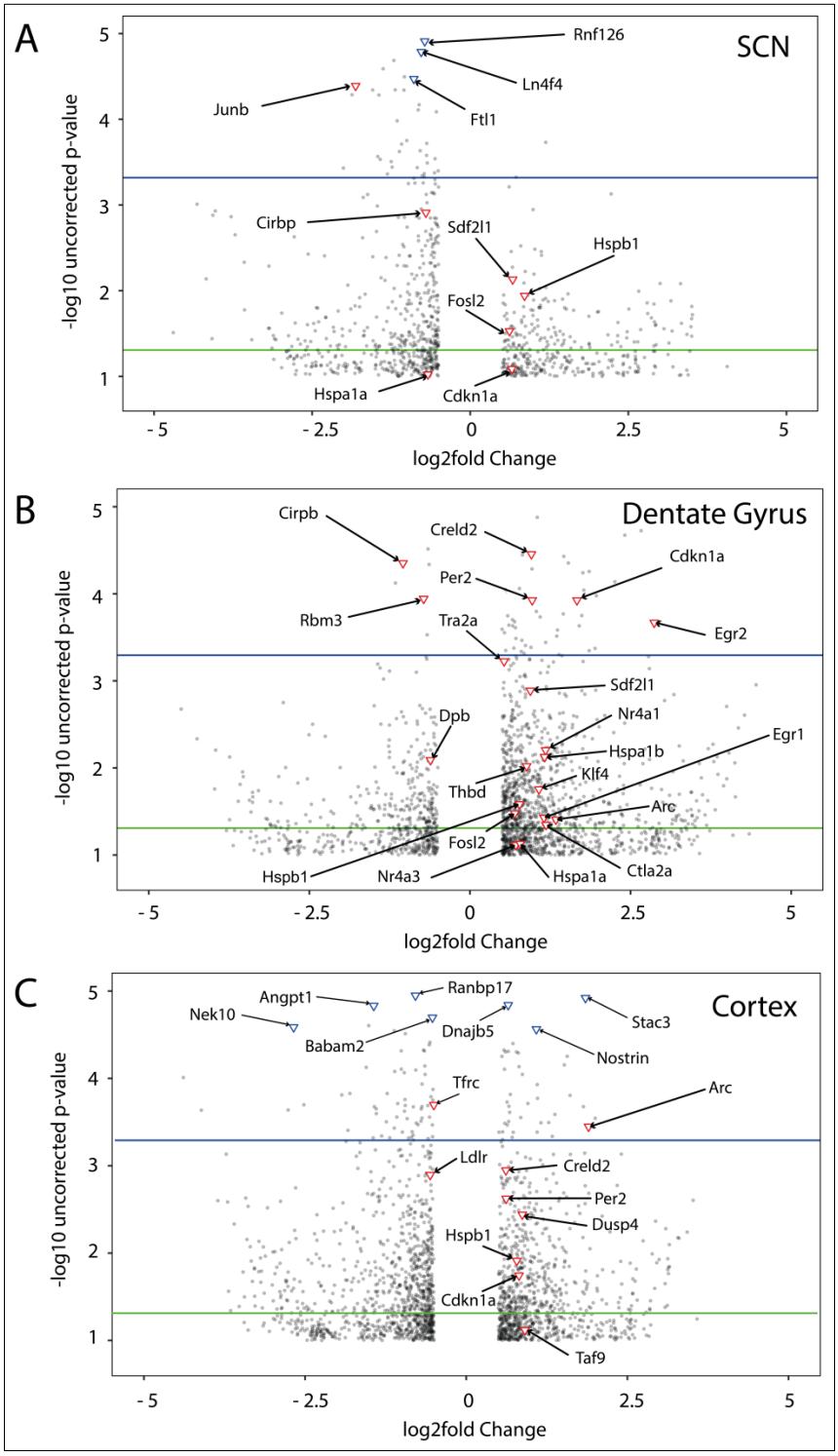


Figure 5.21: Volcano plots showing differential genes following SD analysed by RNA-seq. Points signify a differential gene, blue triangles indicate the most statistically significant differential genes, red triangles show differential genes known to be effected by SD, the green line is the uncorrected P-value threshold of 0.05, and the blue line is the adjusted P-value threshold of 0.05. **A:** SCN; 5 genes associated with SD were statistically significantly DE including *Junb*. **B:** DG; A large number of genes associated with SD were statistically significantly DE including *Egr2, Cirbp* and *Per2*. **C:** Cortex; A large number of genes associated with SD were statistically significantly DE including *Arc*, and *Per2*.

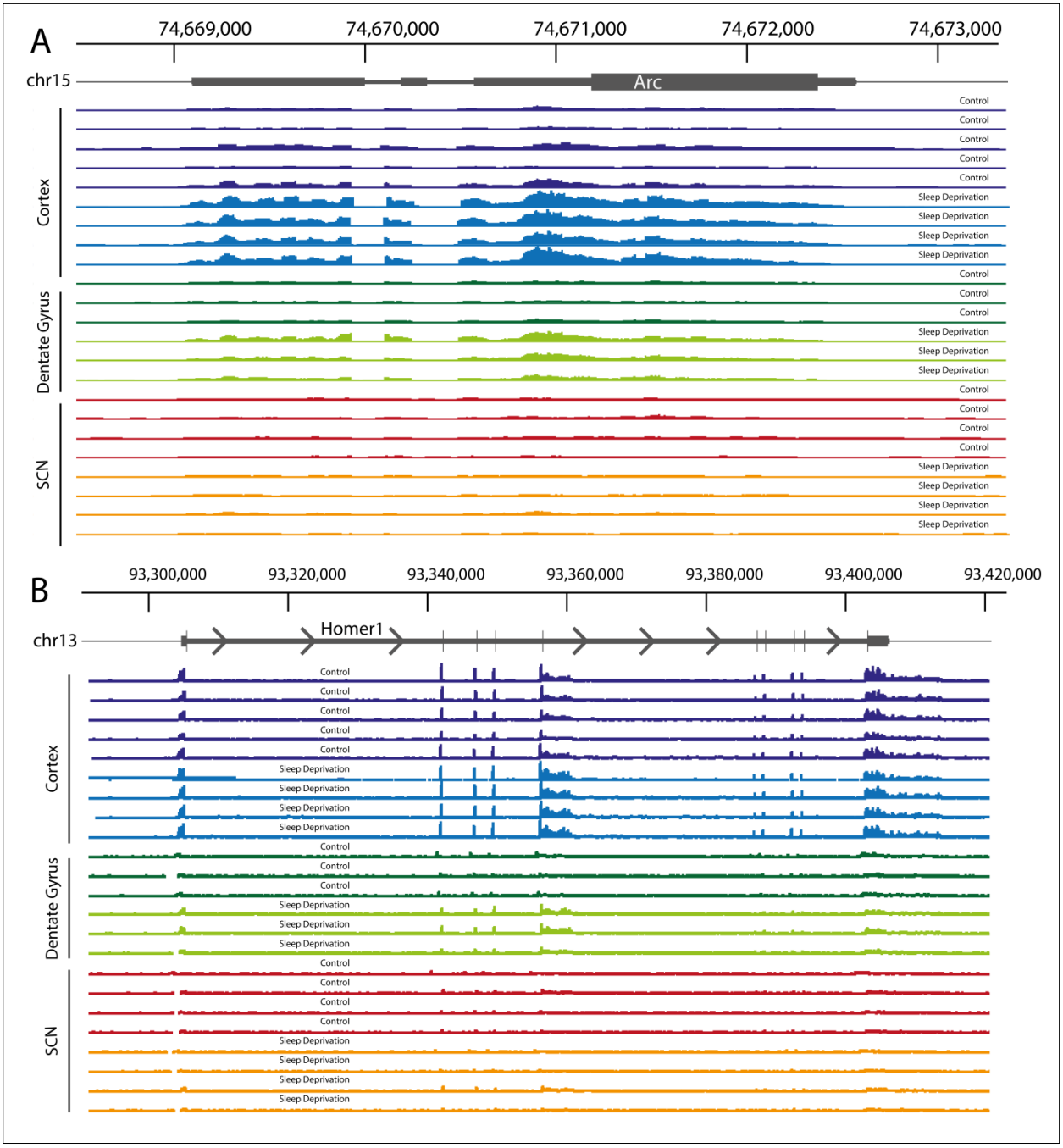


Figure 5.22: RNA-seq read coverage tracks for *Arc and Homer1a*; Coverage tracks are scaled for each individual tissue. **A:** *Arc* transcription was upregulated following SD in the cortex and DG, but was not transcribed in the SCN; **B:** *Homer1a* transcription was upregulated following SD in the cortex and DG, but was not transcribed in the SCN.

5.4.3: Gene set enrichment analysis for RNA-seq

To determine if DE genes shared common functions, or were involved in similar biological pathways, GO BP enrichment analysis was carried out for the top 150 DE genes in each tissue. Analysis revealed clustering across all tissues for GO BP (Figure 5.23), with the main ontologies identified across all brain regions being in line with those identified from gene enrichment analysis from the ATAC-seq experiment, and published data analysed from Hor et al. (2019).

GO BP terms across all brain regions included regulation, metabolism, and response to stimulus, again demonstrating a potential response to stresses caused by SD. For the DG GO BP terms associated with transport clustered very tightly, this GO term was enriched for genes with differential chromatin in the cortex following ATAC-seq, so further verifies a possible response involving the removal of increased toxins and metabolites following SD. GO terms associated with DNA structure were enriched in the SCN, which could potentially indicate DNA structure changing to facilitate repair following DSBs accumulated during SD.

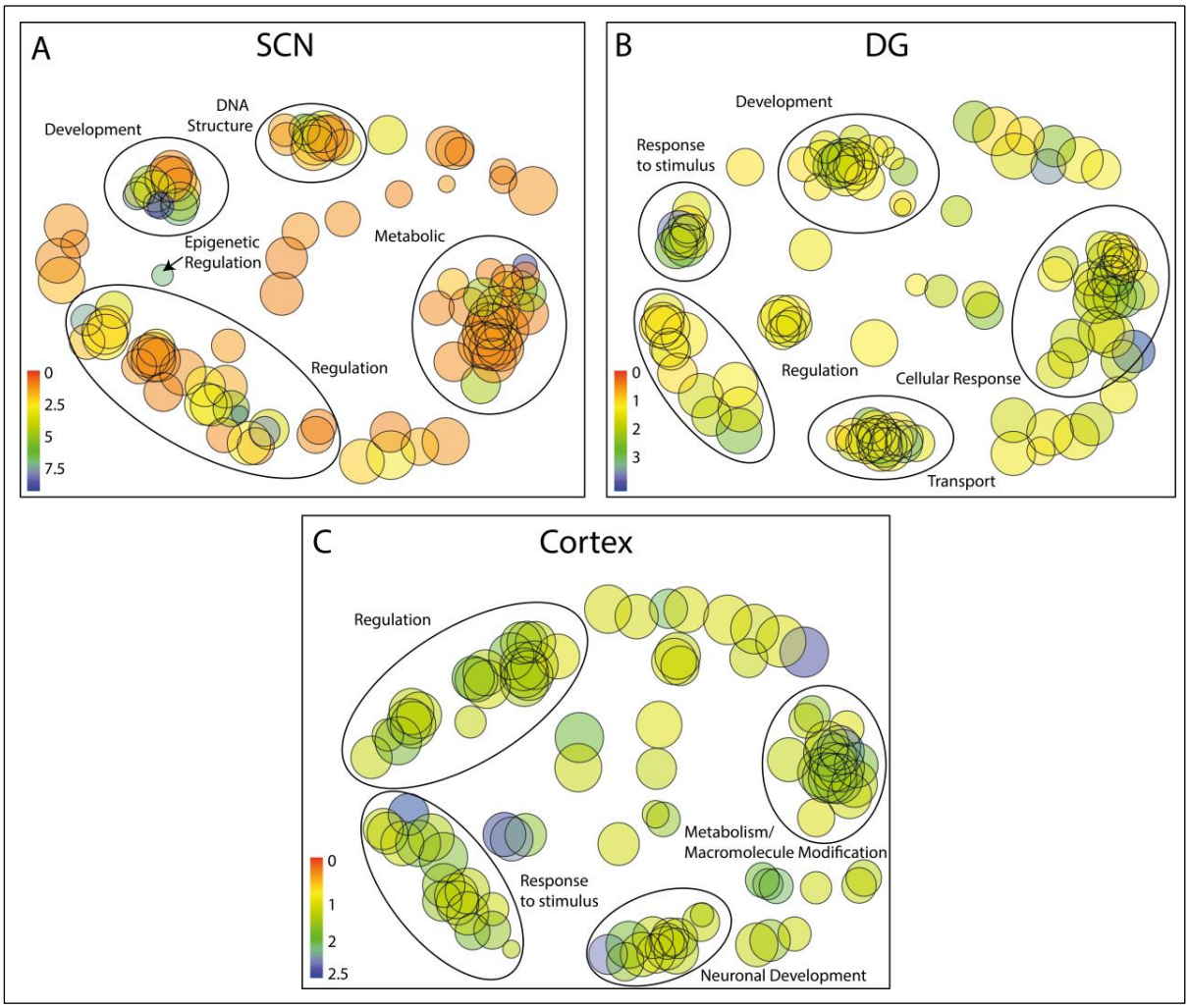


Figure 5.23: GO biological process enrichment for DE genes following SD visualised using Revigo; **A:** SCN; GO BP terms include DNA repair, metabolic and regulation. **B:** DG; GO BP terms include response to stimulus development, transport and cellular response. **C:** Cortex; GO BP terms include regulation, metabolism and neuronal development.

5.4.4: RNA-seq motif analysis

To investigate if DE genes following SD share common motifs in their nearby regulatory regions, motif enrichment analysis was carried out as outlined in section 2.11. Enriched motifs were identified 300bp downstream and 50bp upstream of the top 150 DE genes for all brain regions.

For all brain regions, numerous statistically significant enriched motifs were identified (P-value < 0.05) close to DE genes (Tables 5.6-5.8). Motifs for EGR were identified as being enriched close to DE genes, and this is very intriguing as it has previously been shown in this chapter that EGR2 was down regulated in the DG, but others have previously demonstrated its upregulation in other brain regions (Gerashchenko et al., 2017), revealing possible inter regional signalling at the molecular level between different brain regions following SD.

Furthermore, the binding site for MYB was shown to be enriched in regions close to DE genes in the SCN and DG; MYB is important for regulating the cell cycle, and has been shown to play an important role in DNA repair (Henrich et al., 2017), further supporting the hypothesis for a need to repair excessive DNA DSBs accumulated during SD. ROR was identified as the most statistically significant motif in the SCN; this is particularly remarkable as within the ancillary molecular clock, RORs and REV-ERBs bind to the ROR motif (Figure 1.2), further demonstrating a link between SD and entrainment within the SCN. DE genes within the cortex showed enrichment for RFX motifs, and this is a notable finding as RFX4 has been previously shown to be up-regulated following light exposure exclusively in the SCN and only during the light responsive time of day (Araki et al., 2004). Again this potentially reveals a role for SD in entrainment, and shows interplay across different brain regions following SD at the molecular level.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-10	-2.369e+01	3.57%	0.01%	DUX/ROR
	1e-10	-2.315e+01	24.29%	6.86%	RUNX
	1e-9	-2.272e+01	4.29%	0.04%	ZFX
	1e-9	-2.219e+01	9.29%	0.81%	YY1/RFX
	1e-9	-2.142e+01	7.86%	0.53%	ERG
	1e-9	-2.138e+01	10.71%	1.26%	MYB
	1e-9	-2.122e+01	3.57%	0.02%	ELK/ETV
	1e-8	-1.966e+01	6.43%	0.35%	MAZ/SP1
	1e-8	-1.932e+01	12.86%	2.25%	HIC2
	1e-8	-1.894e+01	5.00%	0.16%	SMAD
	1e-8	-1.874e+01	11.43%	1.80%	DCE
	1e-7	-1.725e+01	6.43%	0.46%	EGR2
	1e-7	-1.725e+01	12.14%	2.29%	NRF/p53
	1e-6	-1.569e+01	26.43%	10.66%	NKX2
	1e-5	-1.330e+01	10.00%	2.07%	MTF1

Table 5.6: Motifs located near to DE genes following SD within the SCN. Enriched motifs include; ROR, an important motif in the ancillary molecular clock; EGR2 whose TF EGR is DE in other brain regions following SD; MYB whose associated TF MYB is involved with the cell cycle, and plays a role in DNA repair.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-9	-2.081e+01	7.07%	0.17%	MED/ZIC
	1e-8	-1.886e+01	5.05%	0.05%	NFYB
	1e-7	-1.829e+01	4.04%	0.01%	EBF/AP-2
	1e-7	-1.773e+01	14.14%	2.05%	MZF1
	1e-7	-1.752e+01	12.12%	1.43%	IRF
	1e-7	-1.632e+01	16.16%	3.11%	BRE
	1e-5	-1.371e+01	22.22%	6.97%	OLIG
	1e-5	-1.283e+01	22.22%	7.36%	FOX
	1e-5	-1.227e+01	10.10%	1.59%	MZF1
	1e-4	-1.113e+01	8.08%	1.08%	EGR1
	1e-4	-1.111e+01	7.07%	0.78%	MYB
	1e-4	-1.095e+01	19.19%	6.44%	OSR/MEF

Table 5.7: Motifs located near to DE genes following SD within the DG. Enriched motifs include; EGR1 whose TFs include EGR which is DE in the DG and across different brain regions following SD; MYB whose associated TF MYB is involved with the cell cycle, and plays a role in DNA repair.

Motif	P-Value	Log P-Value	% Targets	% Back-ground	Closest Match
	1e-12	-2.864e+01	5.65%	0.04%	TCF3/ NFIX
	1e-10	-2.379e+01	5.65%	0.08%	ZF
	1e-9	-2.180e+01	4.84%	0.05%	IRF
	1e-9	-2.088e+01	8.87%	0.64%	YY2/HOX
	1e-8	-1.997e+01	7.26%	0.38%	MEF
	1e-7	-1.841e+01	8.06%	0.63%	EGR
	1e-7	-1.623e+01	16.13%	3.90%	ZEB
	1e-6	-1.585e+01	17.74%	4.81%	DUX
	1e-6	-1.516e+01	4.03%	0.10%	SOX
	1e-5	-1.380e+01	4.03%	0.13%	ZNF/HEK
	1e-5	-1.353e+01	31.45%	14.47%	RAR
	1e-5	-1.349e+01	19.35%	6.47%	CRX
	1e-5	-1.294e+01	13.71%	3.57%	RFX
	1e-4	-1.141e+01	23.39%	9.96%	TEF
	1e-4	-1.038e+01	4.84%	0.46%	KLF

Table 5.8: Motifs located near to DE genes following SD within the cortex. Enriched motifs include; EGR1 whose TFs include EGR which is DE across different brain regions following SD; RFX whose TFs include RFX4 which has been shown to be up-regulated following light exposure in the SCN only during the light responsive time of day.

5.5: Comparison of chromatin accessibility with the transcriptome following SD

Following detection of differential gene expression across all brain regions in response to SD, to examine whether changes in gene expression mirrored or were linked to changes in chromatin accessibility, genes close to areas of differential chromatin were compared to DE genes following SD. Genes close to the top 150 areas of differential chromatin by adjusted P-value after SD were compared to the top 150 DE genes by adjusted P-value after SD. For the DG *Cirbp,* and for the SCN *Jag2* were present in both lists, but for the cortex no genes were found in both lists. *Cirbp* presented a particularly interesting finding because chromatin appeared to close following SD, but its expression increased (Figure 5.24).

As previously stated comparisons with the published dataset from Hor et al. (2019) revealed no genes near differential chromatin present in both their published ATAC-seq data set, and genes identified by ATAC-seq here. However, when the published data set was compared to the genes identified by RNA-seq as being DE, 9 genes were present in both lists; *Angpt1, Arl4a, Slit2, Mar-01, Slitrk1, Tbl1xr1, Pcdh20, Cbln4, Galnt9.* A fisher exact test revealed this was unlikely to be by chance (Table 5.8).

	Non differential RNA-seq genes	**Differential RNA-seq genes**
Non differential ATAC genes	24,000	140
Differential ATAC genes	140	9

Overlapping P-Value = < 0.00001

Table 5.9: Contingency table for fisher exact test for RNA-seq cortex DE genes after SD and published data. The number of genes within the cortex following SD with differential chromatin accessibility at ZT6 from the published data set Hor et al. (2019), and genes with differential gene expression, identified by RNA-seq.

The low number of genes identified here to have differential chromatin accessibility and differential expression is likely due to the ATAC-seq experiment lacking replicates and therefore reducing the certainty of differential genes identified. However, comparison with the published Hor et al. (2019) data set does indicate that changes in chromatin structure, can lead to a change in gene expression.

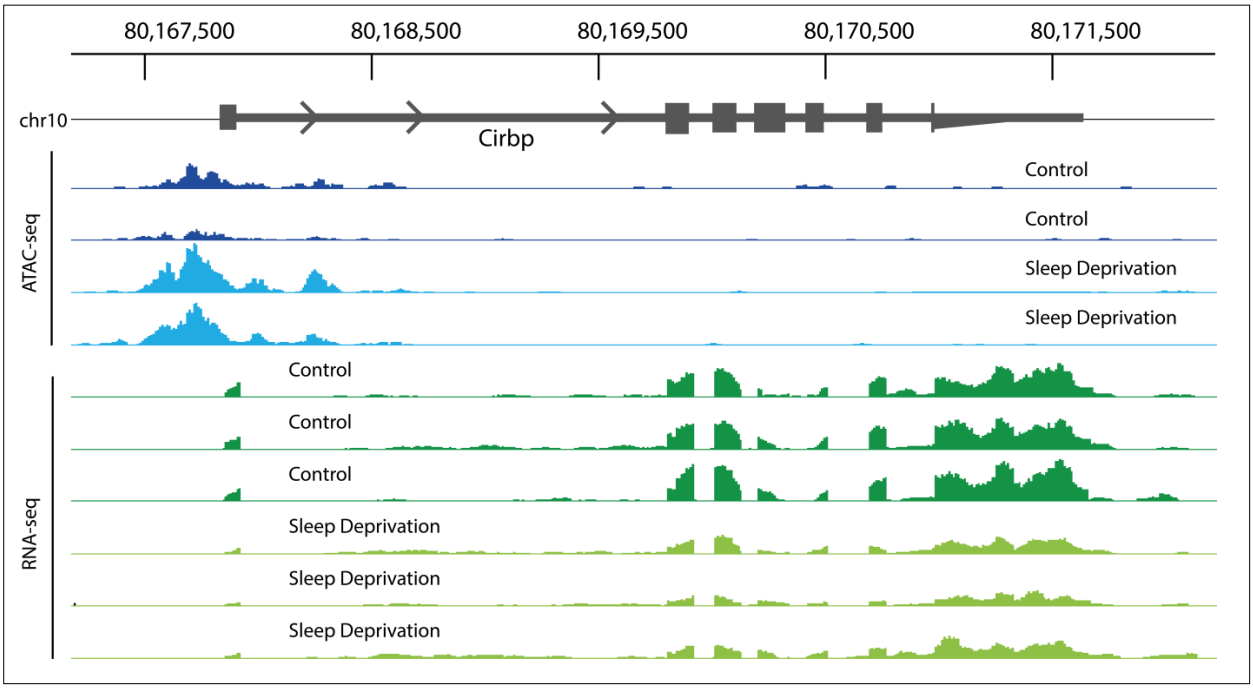


Figure 5.24: Read coverage from ATAC-seq and RNA-seq for *Cirbp*. ATAC-seq tracks and RNA-seq tracks are scaled differently. ATAC-seq revealed chromatin accessibility near to Cirbp increases following SD, but RNA-seq showed its expression decreases.

5.6: Transfection of SD genomic sequences of interest into cultured cells

In the same manner as for the light pulse experiment; further ATAC-seq experiments proved to be unsuccessful due to Covid-19 social distancing restrictions. Therefore a further luciferase assay was carried out, in an attempt to validate results obtained from the SD ATAC-seq first experiment.

From the first ATAC-seq experiment, genomic sequences close to *MiR17hg* were identified as having differential chromatin. *MiR17hg* provided an ideal candidate for further investigation because it encodes for a regulatory micro RNA, which plays an important role in DNA repair, and throughout this project genes involved with DNA repair have repeatedly been highlighted.

Sequences with differential chromatin close to *MiR17hg* were scanned using Jaspar, and found to contain a number of transcription factor binding sites. Motifs again included AP-1, EGR1 and CRE, all important for within the molecular clock. Two regions were chosen to be cloned into a luciferase reporter assay; the whole genomic region of differential chromatin close to *MiR17hg* and the largest of the peaks identified (Figure 5.25).

In the same manner described in chapter 4; the sequences of interested were cloned into pGL4.27 plasmids, before being transfected into mouse NIH3T3 embryonic fibroblast, and human U2OS osteosarcoma cells as described in section 2.8. The cell types and treatments were chosen for the same rational as discussed in chapter 4. A robust chemical mimic for SD is currently unavailable, but because serum treatment activates a wide range of transcription factors, it will give a good indication if these regions are important for transcriptional activation. For this experiment treatment with Forskolin is not intended to determine if the region will respond to light input; this would be difficult to infer because data from the light pulse experiment did not indicate MiR17hg to be switched on following light pulsing. However, the large peak contains a CRE motif; so it may therefore still be activated by the cAMP pathway, and if Forskolin switches on MiR17hg via cAMP, it provides further evidence changes in gene expression following SD may play a role in entrainment, but not necessarily by activation through light input.

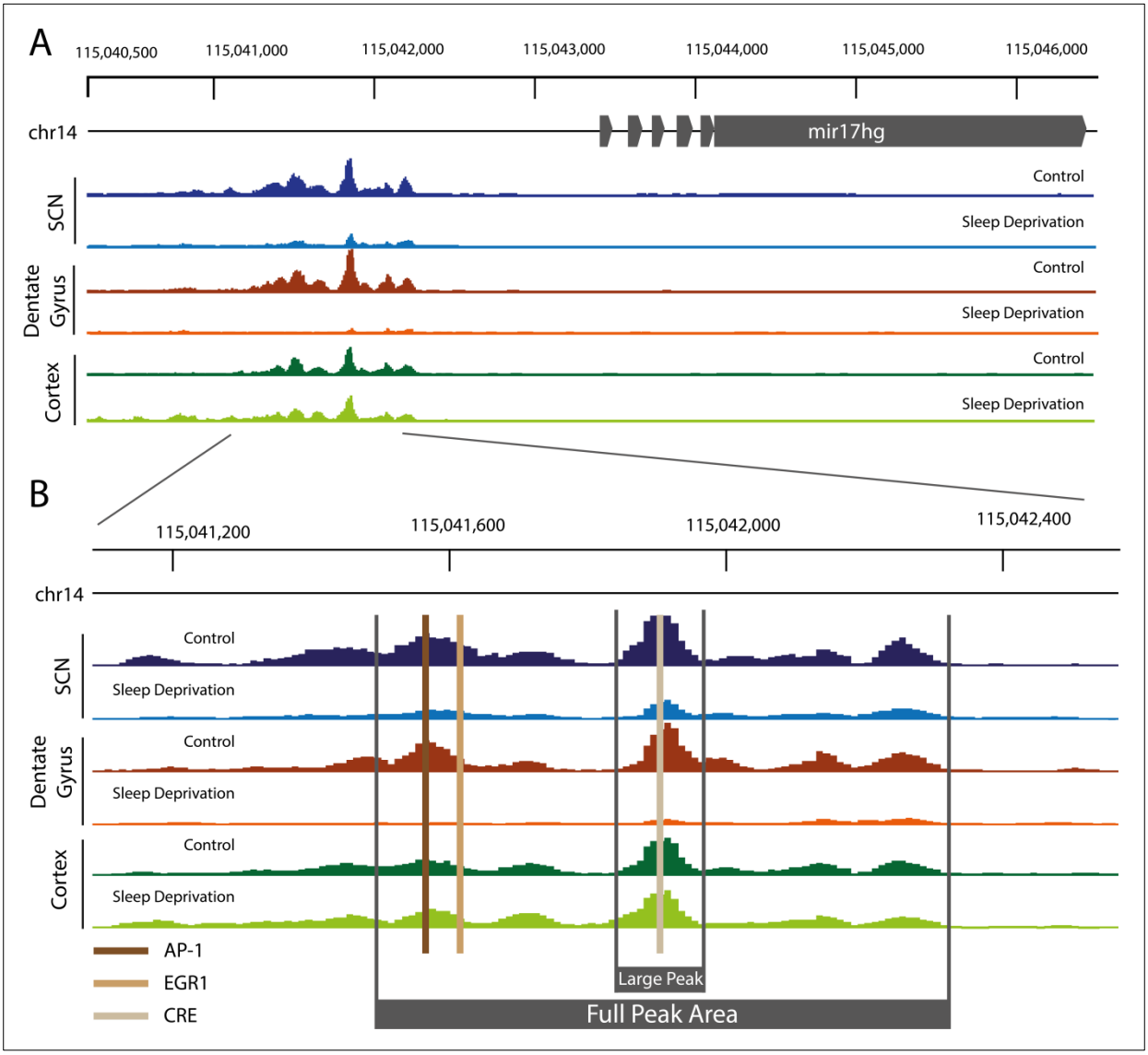


Figure 5.25: ATAC-seq read coverage tracks near to *MiR17hg*. A: Read coverage across the whole differential chromatin near to *MiR17hg* and *MiR17hg* itself; **B:** Close-up of the peak region identified for *MiR17hg*; highlighted are the two genomic regions selected for cloning into cultured cells along with the location of several AP-1, CRE and EGR1 motifs.

Results proved to be positive, both sequences of interest strongly drove production of luciferase following transfection into NIH3T3 and U2OS to a greater extent than control cells (Figures 5.26-A/B and Figures 5.27-A/B). This therefore confirms that the sequences of interest are able to induce transcription, and are likely important in the regulation of transcription for the genes they are located near to.

NIH3T3 cells transfected with either of the sequences of interest showed little additional induction of luciferase after treatment with Forskolin (Figures 5.26-C/D), but cells transfected with either of the sequences of interest showed a substantial increase in luciferase expression following treatment with horse serum (Figures 5.26-E/F). Further analysis confirmed Forskolin had little effect on luciferase expression (Figure 5.26-G), but horse serum greatly increased luciferase expression (Figure 5.26-H).

Following Forskolin treatment U2OS cells transfected with the large *MiR17hg* peak region showed a small blunting of luciferase expression (Figure 5.26-C), but Forskolin had little effect on cells transfected with the full *MiR17hg* peak region (Figure 5.26-D). Serum treatment showed some additional induction in U20S cells transfected with the full *MiR17hg* peak (Figure 5.26-F) region, but had little effect on U20S cells transfected with the large *MiR17hg* peak region (Figure 5.26-E). Closer analysis further validated a blunting of luciferase induction following Forskolin treatment in U2OS cells transfected with the large *MiR17hg* peak region (Figure 5.26-G), but did not validate greater induction following serum treatment in U20S cells transfected with the full *MiR17hg* peak region (Figure 5.26-H).

Overall, similarly to the sequences of interest from the light pulse experiment, all sequences of interest drove luciferase expression without treatments, and the biggest response to treatment was observed in NIH3T3 cells following treatment with horse serum.

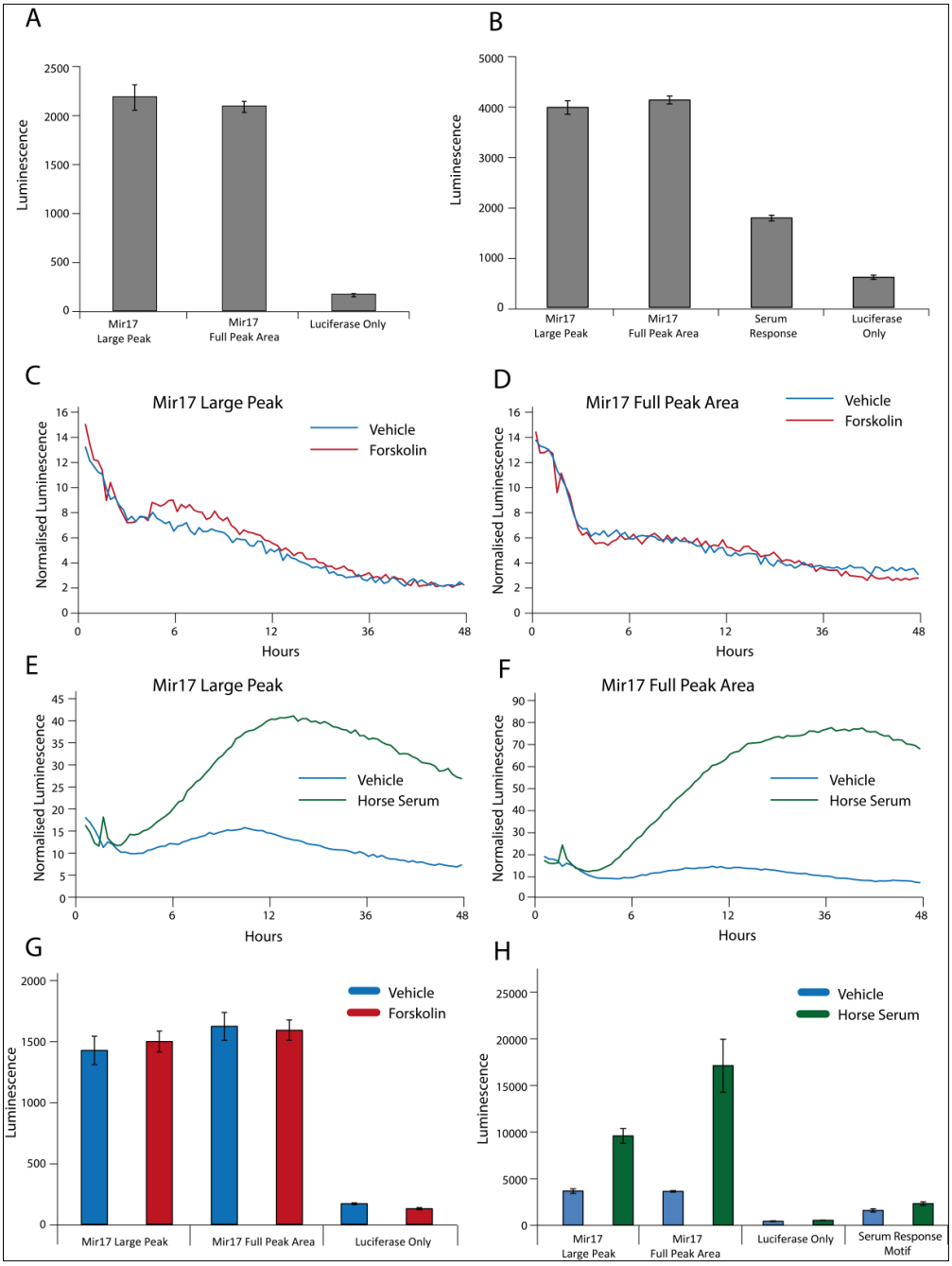


Figure 5.26: Results for luciferase reporter assay using NIH3T3 cultured cells. A: Initial response before Forskolin treatment showed greater induction of luciferase for cells transfection with sequences of interest; **B:** Initial response before serum treatment showed greater induction of luciferase for cells transfection with sequences of interest; **C:** *MiR17hg* large peak region treated with Forskolin shoed blunting of luciferase production; **D:** *MiR17hg* full peak region treated with Forskolin showed no response; **E:** *MiR17hg* large peak region treated with horse Serum showed no response; **F:** *MiR17hg* full peak region treated with horse Serum showed some additional induction of luciferase; **G:** Response after Forskolin treatment showed a response for cells transfected with the *MiR17hg* large peak region; **G:** No Response after treatment with horse Serum.

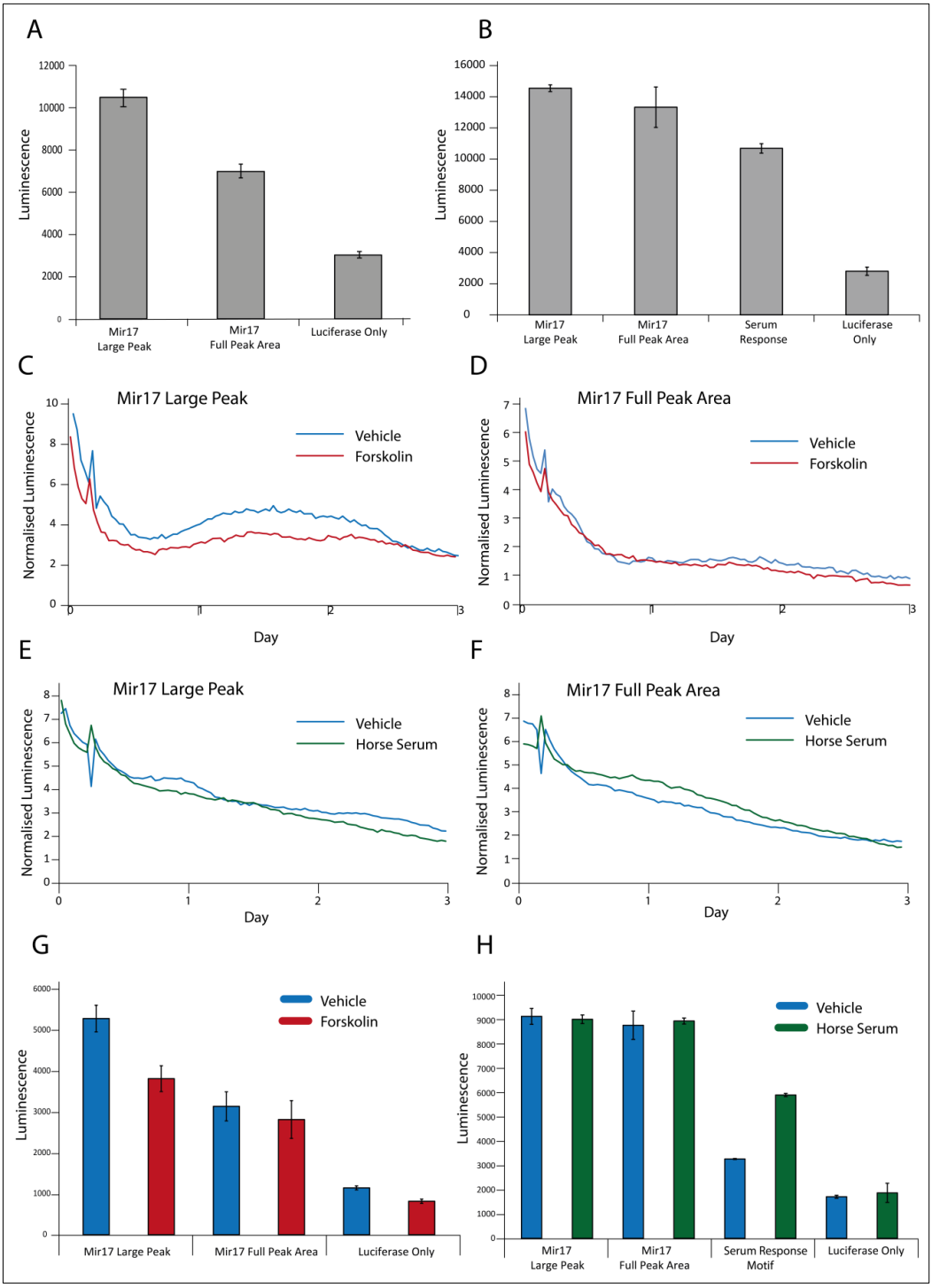


Figure 5.27: Results for luciferase reporter assay using U2OS cultured cells. A: Initial response before Forskolin treatment showed greater induction of luciferase for cells transfection with sequences of interest; **B:** Initial response before serum treatment showed greater induction of luciferase for cells transfection with sequences of interest; **C:** *MiR17hg* large peak region treated with Forskolin shoed blunting of luciferase production; **D:** *MiR17hg* full peak region treated with Forskolin showed no response; **E:** *MiR17hg* large peak region treated with horse Serum showed no response; **F:** *MiR17hg* full peak region treated with horse Serum showed some additional induction of luciferase; **G:** Response after Forskolin treatment showed a response for cells transfected with the *MiR17hg* large peak region; **G:** No Response after treatment with horse Serum.

5.7: Chapter 5 discussion

This chapter aimed to explore how chromatin structure and the transcriptional landscape changes in the SCN, DG, and cortex following SD. To investigate this, ATAC-seq and RNA-seq were successfully carried out, and in response to SD, differential chromatin accessibility and DE genes were identified across all brain regions.

Following ATAC-seq, strong evidence was obtained to confirm that chromatin state between different brain regions is highly dynamic, but it proved more difficult to identify a large quantity of genes with differential chromatin following SD that had strong statistical support (adjusted P-value < 0.05). Despite this, results indicated chromatin near to genes in the SCN and DG showed a tendency to close following SD, while in the Cortex chromatin opened after SD. Additionally, relaxing the stringency of statistical testing revealed many more genes with differential chromatin following SD, which were subsequently confirmed to be differential by visually viewing aligned reads using the IGB. These genes included Cirbp, a known modulator of murine molecular clock gene expression (Hoekstra et al., 2019) and Hspa1b, a heat shock gene induced by SD (Ackermann et al., 2013) whose nearby chromatin closed in the DG following SD. Furthermore, data analysed from Hor et al. (2019), helped validate results for the cortex, indicating a larger number of genes show an increase in chromatin accessibility following SD.

Downstream GSEA between the different brain regions identified both similarities and differences in the functions of the genes close to differential chromatin. When the function of the genes identified was scrutinized for enrichment, the SCN and DG appeared to show enrichment for GO terms associated with transcriptional regulation. This may indicate that transcriptional regulation is changing following SD, allowing different genes to be switched on, potentially a reaction to the stress SD places on the brain. Furthermore, within the cortex GO terms associated with cellular transportation were enriched, which may indicate the removal of excess toxins and other metabolites following an extended period of SD (Eugene and Masiak, 2015).

RNA sequencing proved to be highly effective, with genes known to change their expression following SD, such as *Homer1a, Arc, Per2 and Cirbp* (Cirelli and Tononi, 2000, Terao et al., 2006)

identified as being induced. Differential analysis revealed further DE genes following SD across all brain regions, but many of these genes failed to retain statistical significance following correction for multiple comparisons. This was likely due to the large number of samples multiplexed at one time, reducing the number of reads available for differential analysis using DESeq2. This problem would easily be resolved by sequencing the RNA-seq libraries to a greater depth, the increased read numbers would help to increase statistical power, and likely enable DESeq2 to identify a greater number of statistically significant DE genes.

Chromatin accessibility across all brain regions changed near to and within promoter regions of genes, in areas expected to be crowded with cis-regulatory elements (Figure 5.9). This strongly suggests that chromatin likely plays an important role in the regulation of gene expression. However, when DE genes following SD were compared to genes with differential chromatin, very few genes were identified in both lists. This may suggest that chromatin state is not having an effect on gene transcription, but this would seem unlikely to be the case due to the importance of the physical state of DNA in the initiation of transcription, if a promoter is closed it cannot be transcribed. It is therefore far more likely that the complexity of gene expression is the culprit; when a gene is switched off at the chromatin level, RNA can no longer be transcribed, but RNA already present will remain within the nucleus or cytoplasm for an unknown amount of time, with RNA being degraded at varying lengths of time (Sachs, 1993). Additionally, the promoter regions of many early response genes (ERGs) are never fully closed and remain poised for transcription whether the gene is being expressed, or not (Sng et al., 2004), therefore ATAC-seq may incorrectly identify chromatin near to ERGs as being open, when in fact it is partially closed. Consequently, up regulation of the expression of ERGs may not be reflected in chromatin state because promoter regions were incorrectly identified as already being open.

It is also possible that the low number of replicates from the ATAC seq experiment, and high level of multiplexing carried out for the RNA-seq, limited the results obtained and potentially missed overlap between differential chromatin accessibility and transcription. When differential genes following SD were compared to genes with differential chromatin following SD at ZT6 from published data Hor et

al. (2019), 9 genes were found in both lists. Although ZT6 is not a direct comparison to CT6, this further validates that chromatin state is likely to have an effect on the transcriptome following SD.

Furthermore, GSEA and motif analysis were carried out for data from both ATAC-seq and RNA-seq, the pathways and motifs identified each time were consistent, with many similarities observed. Comparable GSEA and motif results between ATAC-seq and RNA-seq data indicate both that data generated is likely to be accurate, and there must be some cross over between chromatin state and eventual gene transcription. Further analysis of data from Hor et al. (2019) identified similar enriched GO terms within the cortex following SD to those identified from my own ATAC dataset, helping to validate results obtained from my own experiment.

Further evidence was attained in line with current literature to support the hypothesis that SD can have an effect on the SCNs ability to entrain the molecular clock. The motif for ROR, an important motif within the ancillary arm of the molecular clock was found to be enriched amongst promoter regions of genes whose expression changed following SD. Additionally, the expression of JUNB a protein which binds the AP-1 and CRE motifs important for entrainment within the SCN, was also up regulated following SD in the SCN. Increased levels of JUNB will potentially have an effect on core clock proteins such as PER1 by binding the AP-1 motif in its promoter. Motifs for EGR were also enriched in the SCN, and RNA-seq revealed EGR is down regulated in the DG following SD, current literature indicates EGR is up regulated in other brain regions (Gerashchenko et al., 2017), this demonstrates potential cross regional signalling at the molecular level.

Within the DG, genes associated with the mitotic cell cycle were found to be partially enriched, and although the data establishing this enrichment is limited, the molecular clock is intricately related to the cell cycle (Pregueiro et al., 2006) (Matsuo et al., 2003). Therefore enrichment of these genes does seem sensible, and on this premise gives a basis for further analysis. Due to differential analysis having little statistical power, the identification of further genes associated with the cell cycle was difficult, but despite this, by visually interrogating the ATAC-seq read coverage tracks of less statistically significant differential chromatin, *MiR-17hg* was identified as a further gene whose nearby chromatin closed fully in the DG, and partially in the SCN following SD (Figure 5.8-A). *MiR-*

17hg encodes for a cluster of microRNAs, which are involved in the post-transcriptional regulation of gene expression by affecting both the stability and translation of mRNAs (Mogilyansky and Rigoutsos, 2013). RNA-seq carried out was unable to determine if *MiR-17HG* was expressed in any brain region, so to further investigate *MiR-17HG,* two regions of differential chromatin near to *MiR-17hg* were cloned into U2OS and NIH3T3 cultured cells for a luciferase reporter assay. Data indicated both sequences strongly drove luciferase expression in both cells lines, therefore demonstrating their importance in the activation of transcription. Following treatment with horse serum, luciferase production was greatly increased within NIH3T3 cells transfected with either of the genomic regions. Horse serum activates a wide range of molecular clock genes (Balsalobre et al., 1998), therefore providing further evidence *MiR17hg* may play an important following SD.

Closing of nearby chromatin would be expected to switch off production of MiR17hg; as RNA-seq was unable to determine if MiR17hg was being expressed under any condition its downregulation could not be confirmed. Nonetheless, if it is indeed suppressed, reduced levels of MiR17hg would have a positive downstream effect on the transcription of a number of other proteins, through de-repression of its targets. In fact, MiR17hg is already known to be very important in the regulation of hippocampal function, playing a vital role in genome maintenance (Fang et al., 2017). Therefore the Repression of MiR17hg transcription may consequently act to switch on DNA repair mechanisms that would usually be repressed, and may help to mitigate DNA damage caused by excess SD. Indeed SD has already been shown to increase DNA double stranded breaks in Zebrafish (Zada et al., 2019), as shown in Figure 5.28, hence repression of MiR17hg could therefore help mitigate the build-up of double stranded breaks.

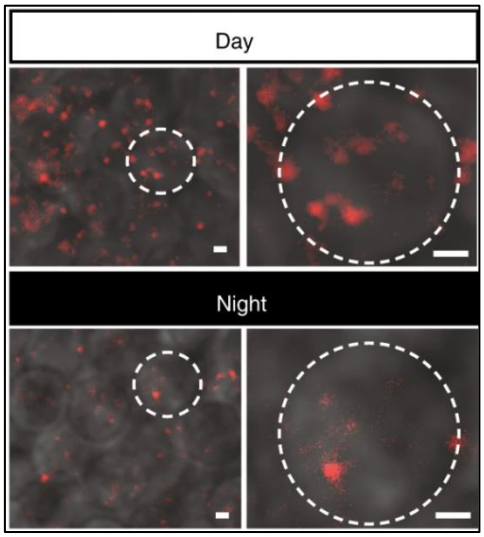


Figure 5.28: DNA Double Stranded breaks. From Zada, Bronshtein et al. 2019; γH2AX is used as a marker of DNA double stranded breaks, double stranded breaks can clearly be seen accumulating during the day, and then disappearing at night during sleep. Dashed circle indicates a single neuron. Scale bar = 1 μm

With the exception of decreased chromatin accessibility close to the *MIR17HG* cluster of micro-RNAs, further evidence of gene expression changes to genes involved with epigenetic modification was not obtained from this investigation. Neither the expression of, nor chromatin located near to genes involved with epigenetic modification such as HKMTs appeared to change following SD. With the significant amount of literature indicating the importance of epigenetic modification following SD (Lahtinen et al., 2019, Ruifeng et al., 2016), it is likely the limitations of data obtained here, are the reason further evidence was not obtained.

This chapter has clearly demonstrated both chromatin accessibility and gene expression change following SD, and has shown both the expression and chromatin accessibility near to genes known to be important for SD differs following SD. Owing to limitations with statistical power, to determine if the chromatin sate is indeed having an effect on the transcriptome, and if chromatin near to genes and small RNAs such as *MiR17hg* is closing following SD, a further ATAC-seq experiment must be conducted when Covid-19 restrictions have been lifted, this experiment ideally would have 4 replicates for each brain region with and without SD.

Data from Hor et al. (2019) has demonstrated chromatin state in the cortex state to be highly dynamic, therefore a further more robust ATAC-seq experiment would likely find a greater number of genes with differential chromatin nearby following SD, and these areas of differential chromatin would be easier to validate statistically. Following completion of a further ATAC-seq experiment, the RNA-seq libraries would then be sequenced to a greater depth enabling differential genes following SD to be easily validated statistically, and give a more robust gene list to compare to ATAC-seq genes with differential chromatin.

Furthermore, due to the complexities of transcriptional regulation additional RNA-seq experiments will need to be carried out to determine exactly how the transcriptome is changing following SD in the SCN, DG and Cortex. Ideally, both nascent RNA, and micro RNA would be extracted and sequenced; nascent RNA from the nucleus would give a better idea how transcription is changing immediately following SD, and discount any RNA persisting within the cytosol, and extracting micro RNAs would enable the expression of micro RNAs such as the MiR17hg cluster to be determined following SD.

Chapter 6

Role of circadian regulation in the immune response to Covid-19

6.1: Introduction

The COVID-19 pandemic, caused by SARS-CoV-2 coronavirus, is a global health problem, leading to unprecedented public health challenges. SARS-CoV-2 primarily infects cells of the respiratory tract, by binding human angiotensin-converting enzyme (ACE2) (Wan et al., 2020), and infection leads to a wide range of respiratory problems and can often result in pneumonia.

Throughout this project, the importance of an organism's capability to correctly align to its external environment through precise entrainment of circadian rhythms has been reported. The downstream result of this intricate alignment helps an organism to function successfully within its external environment, and part of this includes its immune response to threats, such as those from viral infections. In fact a number of recent studies have shown a role for the circadian clock to regulate host susceptibility to viral infection (Borrmann et al., 2020); influenza-A infection of arrhythmic mice, which lack the circadian component BMAL1, and therefore a correctly functioning molecular clock, results in higher viral replication and elevated inflammatory response (Edgar et al., 2016).

Bioinformatic analysis was carried out as part of collaboration, aiming to demonstrate the importance concerning circadian regulation of ACE2 in lung epithelial cells. The study aimed to show that silencing BMAL1 or treatment with the synthetic REVERB agonist SR9009 reduces ACE2 expression and inhibits SARS-CoV-2 entry and therefore RNA replication.

The aim of this small bioinformatic study was to help substantiate the initial hypothesis of the project, to demonstrate that currently available data supports the idea that circadian rhythms are vitally important to an organism's response to viral infection, and therefore likely play a role in the response to SARS-CoV-2. In-silico bioinformatic analysis was carried out to locate *Bmal1* motifs near to SARS-Cov-2 interacting genes; these locations were then verified using published CHIP-seq *Bmal1* data to determine the most notable target genes.

The bioinformatic analysis aimed to answer the following questions:

- Is it possible to identify Bmal1 binding targets near to known SARS-Cov-2 interacting genes?
- Are any targets identified further supported using CHIP-seq data of known Bmal1 binding sites?
- Does bioinformatic analysis substantiate the overall hypothesis of the project, that circadian rhythms likely play an important role in a host's response to SARS-Cov-2 infection?

6.2: Methods

Gene lists were downloaded from (Gordon et al., 2020), and converted into numeric entrez species specific gene I.Ds. A motif specific search against the human genome was carried out using HOMER as described in section 2.7, promoter regions of SARS-Cov-2 interacting genes were scanned for the following circadian related motifs identified from the Jaspar database (Fornes et al., 2020):

Bmal1 'core' motif: **CACGTG**

Mouse Bmal1 motif: 

Human CLOCK motif: 

The human CLOCK motif was included because a human BMAL1 motif was not available from Jaspar, and evidence of CLOCK biding near to SARS-Cov-2 interacting genes would further support the hypotheses, for the same reason, the mouse BMAL1 motif was used in the absence of the human BMAL1 motif. The motif profiles were converted to HOMER compatible motifs as outlined in section 2.7, and are shown below:

>bmal1 ets 8

0.001	0.997	0.001	0.001
0.997	0.001	0.001	0.001
0.001	0.997	0.001	0.001
0.001	0.001	0.997	0.001
0.001	0.001	0.001	0.997
0.001	0.001	0.997	0.001

>mouse-bmal1 ets 8

0.251000	0.229000	0.261000	0.259000
0.246000	0.125000	0.558000	0.071000
0.061938	0.043956	0.142857	0.751249
0.003000	0.997000	0.000000	0.000000
0.953000	0.016000	0.010000	0.021000
0.002000	0.965000	0.009000	0.024000
0.048000	0.003000	0.946000	0.003000
0.054000	0.058000	0.051000	0.837000
0.009009	0.004004	0.938939	0.048048
0.255000	0.331000	0.146000	0.268000

>human-clock ets 8

0.526807	0.081585	0.216783	0.174825
0.696429	0.137987	0.144481	0.021104
0.024775	0.966216	0.004505	0.004505
0.955457	0.011136	0.033408	0.000000
0.033827	0.906977	0.000000	0.059197
0.085288	0.000000	0.914712	0.000000
0.006787	0.011312	0.011312	0.970588
0.000000	0.006803	0.972789	0.020408
0.001445	0.219653	0.158960	0.619942
0.206977	0.297674	0.090698	0.404651

The binding affinity for each motif was altered in-silico using the Homer Log odds detection threshold as described in section 2.7. For each motif, a strict, moderate, and loose search was carried out; this essentially indicates how exact the match with the motif had to be, and therefore an estimate of binding affinity.

Following identification of SARS-Cov-2 interacting genes with a BMAL1 motif, this list of genes was compared to the list of genes containing a BMAL1 motif that had been experimentally verified using CHIP-seq from (Beytebiere et al., 2019); this search was carried out using R-studio, the code is shown below in Figure 71.

```
```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

#Load in Data
```{r}
chiplist <- read.csv(file ="C:\\Uni\\DPhil\\Thesis\\Chapter\ 7\ Covid-19\\
COVID-STUFF\\CHipData\\bmal1_chipgenes.csv")
covidlist <- read.csv(file="C:\\Uni\\DPhil\\Thesis\\Chapter\ 7\ Covid-19\\
COVID-STUFF\\To-send\\bmal1-mouse-moderate-genelist.csv")
```

#Ensure both lists are in the same format and contain only unique entries
```{r}
chiplist <- tolower(unique(chiplist$Gene.Name))
covidlist <- tolower(unique(covidlist$query))
```

#Get genes in both lists
```{r}
inboth <- as.data.frame(chiplist %in% covidlist) # Get boolean for both lists
genesinboth <- which (inboth == TRUE) #Get only True values, and corresponding vector ID's

x <- chiplist[genesinboth] # Use vector ID's to get genes from covid chip list
Y <- covidlist[genesinboth] # Use vector ID's to get genes from covid interacting list

length(x) == length(Y) #Sanity check
print(x) #Get genes
```
```

Figure 6.1: R-code to determine which genes have a BMAL1 motif in the SARS-Cov-2 gene list determined in-silico, and also have an experimentally verified BMAL1 from CHIP-seq data.

6.3: Results

Initial motif specific searching proved to be successful, and all motifs were identified close to many SARS-Cov-2 interacting genes, Table 6.1 gives a summary of the results:

	Bmal1 core	**Mouse Bmal1**	**Human CLOCK**
Loose	96	253	198
Moderate	0	89	46
Strict	0	21	6

Table 6.1: Results for motif specific searching against BMAL1 core, Mouse BMAL1 and Human CLOCK. with loose, moderate and strict settings applied.

Following specific motif analysis, the genes obtained containing the mouse BMAL1 motif nearby using the loose settings were compared to the BMAL chip-seq data set; the gene set derived using loose settings was used because it was being compared to experimentally verified data. Following comparison, 94 genes were found to be in both lists, and a fisher exact test determined that the chance of the genes appearing in both lists, from a pool of 24,000 mouse genes was unlikely to be by chance, this is shown in Table 6.2 below:

	Non SARS-Cov-2 genes	**SARS-Cov-2 genes**
Non BMAL1 CHIP genes	24,000	253
BMAL1 CHIP genes	6143	94

Overlapping P-Value = 3.1e-03

Table 6.2: Contingency table for fisher exact test for SARS-Cov-2 interacting genes lists with a BMAL motif nearby, and gene lists for experimentally verified chip-seq binding sites.

Several of the genes identified in both list were viewed in the integrative genome browser, and the BMAL1 binding site predicted by in-silico searching, was found to be in the same location as the peaks, this is shown for *Rrp9* in Figure 7.2 below:

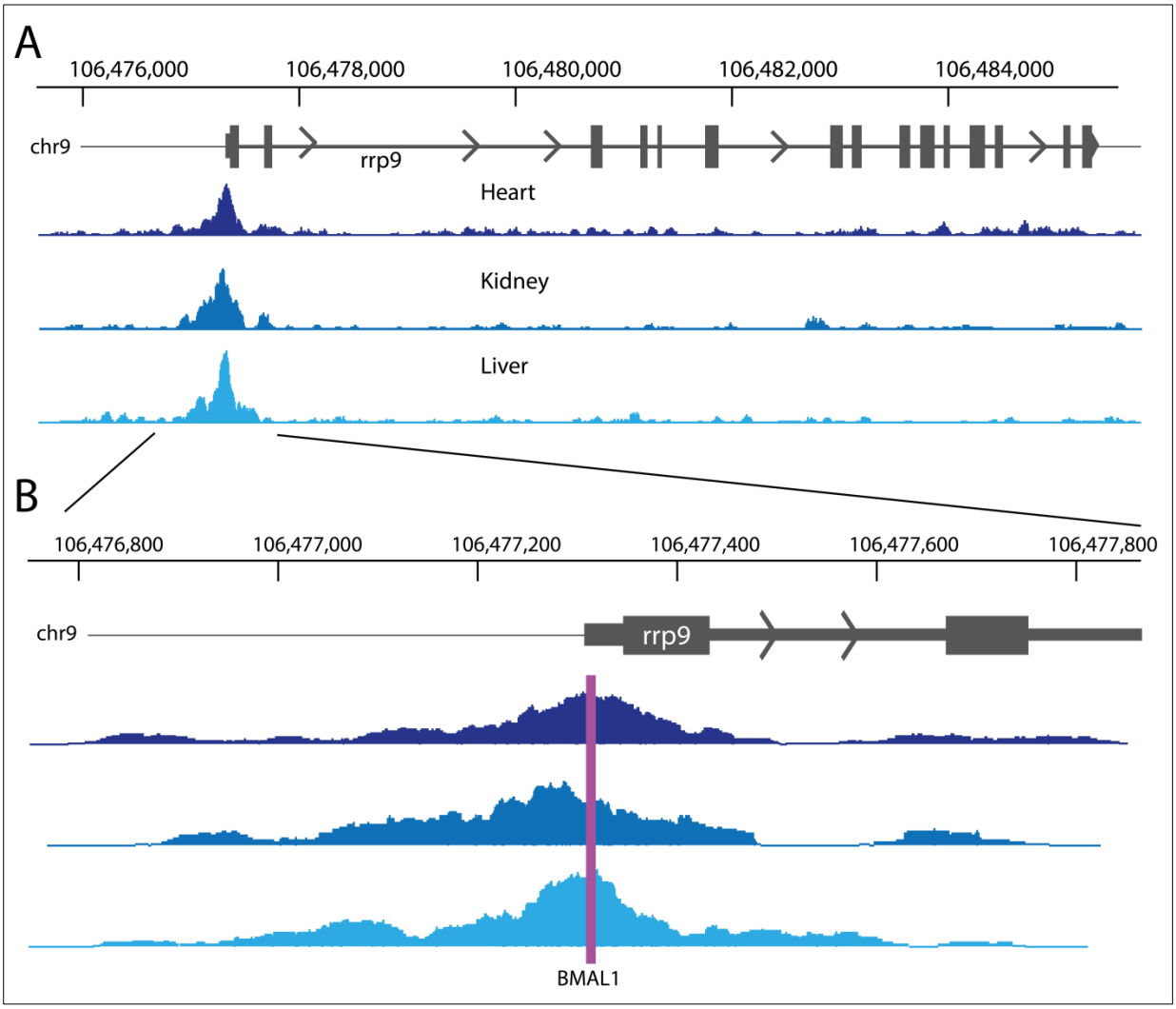


Figure 6.2 Chip-seq binding tracks for BMAL1 from heart, kidney and liver. A: whole *Rrp9* gene area; **B:** Chip-seq peaks near to *Rrp9*, the in-silico BMAL1 motif has been highlighted.

6.4: Chapter 6 Discussion

BMAL1 motif binding sites were identified near to numerous SARS-Cov-2 interacting genes, and a large number of these genes could be verified using the BMAL binding CHIP-seq data set. This therefore indicates that BMAL1 binding likely plays a role in many of the SARS-Cov-2 interacting genes, and in turn demonstrates the molecular clock is anticipated to regulate host susceptibility to viral infection. Indeed many of the genes found in both the SARS-Cov-2 interacting gene list and the CHIP-seq data set have well established roles in the immune response. For example the protein product of *Rrp9* shown in Figure 7.2 is involved with RNA-processing (Bouhaddou et al., 2020), and is differentially phosphorylated during infection.

The aforementioned therefore sufficiently validates the initial hypothesis, and substantiates that a larger project to demonstrate the importance of circadian regulation in production of ACE2 within lung epithelial cells. To that end experiments have since been carried out using Calu-3 human epithelial lung cells, and have shown a significant reduction of SARS-CoV-2 (Victoria 01/20, B1.1.7, B1.135 strains) replication in Bmal1 silenced Calu-3 cells compared to parental cells (Figure 6.2). Furthermore, treatment with SR9009, a compound that enhances REV-ERB activity and therefore suppresses BMAL1 reduced SARS-CoV-2 replication.

This small bioinformatic project successfully helped to confirm a hypothesis which has since been validated, and aims to help the ongoing global fight against Covid-19. The research presented here along with the subsequent experiments carried out, are currently under review with Cell Reports.

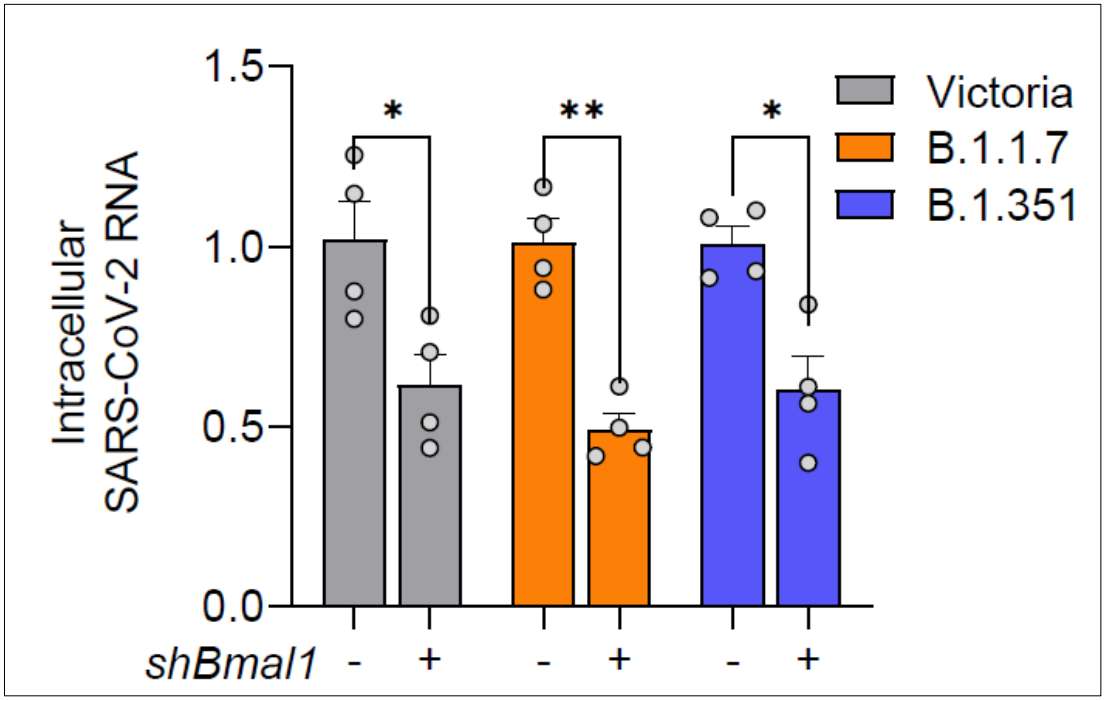


Figure 6.3: Silencing BMAL1 reduces SARS-CoV-2 replication: Across all three strains of SARS-CoV-2 tested, the silencing of BMAL1 reduces intracellular SARS-CoV-2 RNA.

Chapter 7

Final Discussion

This project aimed to refine the published ATAC-seq protocol to work with small amounts of tissue obtained from mouse SCN, and then subsequently use this technique to investigate changes in chromatin accessibility following light pulsing in the SCN at CT8 and CT16, and following 6 hours of SD in the SCN, DG and cortex. Despite ATAC-seq often being a difficult technique to master, the protocol was successfully refined, and several ATAC-seq datasets were successfully generated. Chromatin accessibility in the SCN following light pulsing was compared to the SCN transcriptome following light pulsing using published data from Jagannath et al. (2013), and chromatin accessibility in the SCN, DG and cortex after SD was compared to the transcriptome by generating a RNA-seq dataset following SD.

The data generated revealed several interesting findings; chromatin accessibility in the SCN was found to change following light pulsing, but only at CT16 when the SCN is already known to be the most responsive to light (Khalsa et al., 2003). SD was also found to alter chromatin structure across all brain regions investigated, in the SCN and DG chromatin primarily closed, while in the cortex, and in agreement with published data from Hor et al. (2019), chromatin primarily opened.

Comparisons made between changes in chromatin state and gene expression did not produce compelling evidence to support the hypothesis that changes in chromatin state will have an effect on gene expression. Following light pulsing a number of genes were identified whose nearby chromatin state and gene expression changed; indicating chromatin state may be effecting expression (Figure 5.24). However, following SD, comparisons between chromatin state and the transcriptome were more limited, which could be as a consequence of the ATAC-seq experiment only having two replicates per sample and therefore limited statistical power. However, I believe it is far more likely that the lack of correlation between changes in chromatin accessibility and transcript abundance in both cases is caused by the complexities of transcription and the time points selected. Indeed, the

chromatin of ERGs often only ever changed to a partially closed state, and following the closing of chromatin, mRNA already transcribed may persist for extended time periods (Sachs, 1993). Therefore, it can easily be seen how linking chromatin state to gene expression can be very challenging. Despite this, comparisons made between GSEA of genes located close to differential chromatin and DE genes following SD revealed they share similar GO terms, clearly the differential overlap between chromatin and gene expression indicates that these changes will impact a similar group of pathways and biological processes.

Through the annotation of genes close to differential chromatin, many notable genes across all experiments were identified. Following light pulsing at CT16, ATAC-seq revealed chromatin close to *Csnk1e* and *Nfil3* (both components of the molecular clock (Figure 1.2)) appeared to close, and after SD the chromatin of *Cirbp* (known to adjust clock-gene expression (Hoekstra et al., 2019)) appeared to open; yet RNA-seq revealed its expression fell. In addition novel genes such as *MafG* and the micro-RNA cluster *MiR17hg* were discovered following light pulsing and SD respectively. Both these provide interesting leads for follow up experimentation; as MAFG is known to bind AP-1 motifs (Kerppola and Curran, 1994) it may play a role in entrainment, and MiR17hg being important for DNA repair (Xie et al., 2019) could possibly have a role in the recovery from SD.

Despite successfully modifying the published ATAC-seq protocol to work with small amounts of SCN tissue and generating several datasets, throughout this project ATAC-seq has proved to be a difficult technique to assay chromatin structure, and has at times given incomplete results. While differential analysis using DeSeq2 for ATAC-seq and RNA-seq frequently failed to correctly identify areas of differential chromatin or DE genes, it also occasionally determined genes or chromatin to be differential when they were not. Repeatedly, statistical testing had to be relaxed to maximise the discovery of differential areas of chromatin or genes, then read alignments had to be visualised using the IGB to determine if areas of differential chromatin or gene expression were in fact changing across the different samples conditions.

Undesirably, due to Covid-19 further ATAC-seq datasets could not be generated, and therefore results obtained could not be validated using the same experimental techniques. To overcome this, several

genomic areas of interest identified from differential chromatin close to interesting genes were transfected into cultured cells along with a luciferase reporter. All sequences strongly drove luciferase expression in both mouse NIH3T3 cells and U2OS human cells, demonstrating their importance with regard to initiating transcription. Furthermore, subsequent treatment with horse serum significantly increased luciferase production in NIH3T3 cells transfected with any of the sequences of interest. With serum shock being known to induce circadian genes (Balsalobre et al., 1998), this indicates the sequences of interest may play a role in either entrainment or recovery following SD.

NGS techniques have gained great popularity over the last decade, and give the promise of generating vast quantities of data across the whole genome from a single experiment. This project has no doubt demonstrated these are powerful techniques which will generate large amounts of data, however data analysis can often be complex and at times prone to errors. No doubt techniques for analysing NGS data will continue to evolve, and over time will become far more capable of analysing data at the whole genome level. It is likely methods will become more complex and machine learning techniques will be employed to analyse data in such a way that would be impossible using current techniques. In fact such techniques are currently being investigated (Jabeen et al., 2018), and already proving to be successful, but until these methods are widely peer reviewed and accepted, and computational power meets the demands of more demanding analytical methods, the accepted methodologies and programs, such as DESeq2, will continue to be most widely used. While waiting for analytical techniques to improve, this project has demonstrated that results obtained using NGS methods should always be viewed in the light of the capabilities of current techniques. Results from NGS experiments should themselves not to be used to paint the entire picture, but should instead be used to explore and generate ideas for future hypotheses, and importantly they should be fully corroborated using alternative, but complementary, molecular biology techniques.

References

Ackermann, K., Plomp, R., Lao, O., et al. 2013. Effect of sleep deprivation on rhythms of clock gene expression and melatonin in humans. *Chronobiology International,* 30**,** 901-9.

Adey, A., Morrison, H. G., Asan, et al. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biology,* 11**,** R119.

Aguilar-Arnal, L., Hakim, O., Patel, V. R., et al. 2013. Cycles in spatial and temporal chromosomal organization driven by the circadian clock. *Nature Structural & Molecular Biology,* 20**,** 1206-13.

Alhola, P. & Polo-Kantola, P. 2007. Sleep deprivation: Impact on cognitive performance. *Neuropsychiatric Disease and Treatment,* 3**,** 553-567.

Allfrey, V. G., Faulkner, R. & Mirsky, A. E. 1964. Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proceedings of the National Academy of Sciences of the United States of America,* 51**,** 786-794.

Altman, Brian j., Hsieh, Annie l., Sengupta, A., et al. 2015. MYC disrupts the circadian clock and metabolism in cancer cells. *Cell Metabolism,* 22**,** 1009-1019.

Amemiya, H. M., Kundaje, A. & Boyle, A. P. 2019. The ENCODE blacklist: Identification of problematic regions of the genome. *Scientific Reports,* 9**,** 9354.

Andrews, S. FastQC A quality control tool for high throughput sequence data. *http://www.bioinformatics.babraham.ac.uk/projects/fastqc/*.

Araki, R., Takahashi, H., Fukumura, R., et al. 2004. Restricted expression and photic induction of a novel mouse regulatory factor X4 transcript in the suprachiasmatic nucleus. *Journal of Biological Chemistry,* 279**,** 10237-42.

Ausió, J., Paz, A. M. D. & Esteller, M. 2014. MeCP2: the long trip from a chromatin protein to neurological disorders. *Trends in Molecular Medicine,* 20**,** 487-498.

Azzi, A., Dallmann, R., Casserly, A., et al. 2014. Circadian behavior is light-reprogrammed by plastic DNA methylation. *Nature Neuroscience,* 17**,** 377-82.

Balsalobre, A., Brown, S. A., Marcacci, L., et al. 2000. Resetting of circadian time in peripheral tissues by glucocorticoid signaling. *Science,* 289**,** 2344.

Balsalobre, A., Damiola, F. & Schibler, U. 1998. A serum shock induces circadian gene expression in mammalian tissue culture cells. *Cell,* 93**,** 929-37.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., et al. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics,* 27**,** 1691-1692.

Beytebiere, J. R., Trott, A. J., Greenwell, B. J., et al. 2019. Tissue-specific BMAL1 cistromes reveal that rhythmic transcription is associated with rhythmic enhancer-enhancer interactions. *Genes & development,* 33**,** 294-309.

Bodnar, J. W. & Bradley, M. K. 1996. A chromatin switch. *Journal of Theoretical Biology,* 183**,** 1-7.

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

Borbély, A. A. 1982. A two-process model of sleep regulation. I. Physiological basis and outline. *Human Neurobiology,* 1**,** 195-204.

Borrmann, H., Mckeating, J. A. & Zhuang, X. 2020. The circadian clock and viral infections. *Journal of Biological Rhythms,* 36**,** 9-22.

Bouhaddou, M., Memon, D., Meyer, B., et al. 2020. The global phosphorylation landscape of SARS-CoV-2 infection. *Cell,* 182**,** 685-712.e19.

Bozek, K., Relógio, A., Kielbasa, S. M., et al. 2009. Regulation of clock-controlled genes in mammals. *PLOS ONE,* 4**,** e4882.

Buenrostro, J., Wu, B., Chang, H. & Greenleaf, W. 2015. ATAC-seq: A method for assaying chromatin accessibility genome-wide. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.],* 109**,** 21.29.1-21.29.9.

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., et al. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods,* 10**,** 1213-1218.

Buxton, O. M. & Marcelli, E. 2010. Short and long sleep are positively associated with obesity, diabetes, hypertension, and cardiovascular disease among adults in the United States. *Social Science & Medicine,* 71**,** 1027-1036.

Ceasar, S. A., Rajan, V., Prykhozhij, S. V., et al. 2016. Insert, remove or replace: A highly advanced genome editing system using CRISPR/Cas9. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research,* 1863**,** 2333-2344.

Cedernaes, J., Osler, M. E., Voisin, S., et al. 2015. Acute sleep loss induces tissue-specific epigenetic and transcriptional alterations to circadian clock genes in men. *The Journal of Clinical Endocrinology and Metabolism,* 100**,** E1255-61.

Cgat Mapping. https://github.com/cgat-developers/cgat-flow/tree/master/cgatpipelines/tools/pipeline_mapping.

Cgat. Peak Calling. *https://github.com/cgat-developers/cgat-flow/tree/master/cgatpipelines/tools/pipeline_peakcalling* [Online]. [Accessed].

Chen, C. & Bazan, N. G. 2005. Lipid signaling: Sleep, synaptic plasticity, and neuroprotection. *Prostaglandins & Other Lipid Mediators,* 77**,** 65-76.

Chung, I.-M., Ketharnathan, S., Kim, S.-H., et al. 2016. Making sense of the tangle: insights into chromatin folding and gene regulation. *Genes,* 7.

Cirelli, C. & Tononi, G. 1998. Differences in gene expression between sleep and waking as revealed by mRNA differential display. Published on the World Wide Web on 30 March 1998. *Molecular Brain Research,* 56**,** 293-305.

Cirelli, C. & Tononi, G. 1999. Differences in brain gene expression between sleep and waking as revealed by mRNA differential display and cDNA microarray technology. *The Journal Sleep Research,* 8 Suppl 1**,** 44-52.

Cirelli, C. & Tononi, G. 2000. Differential expression of plasticity-related genes in waking and sleep and their regulation by the noradrenergic system. *The Journal of neuroscience : the official journal of the Society for Neuroscience,* 20**,** 9187-9194.

Cirelli, C. & Tononi, G. 2008. Is sleep essential? *PLOS Biology,* 6**,** e216.

Crosio, C., Cermakian, N., Allis, C. D. & Sassone-Corsi, P. 2000. Light induces chromatin modification in cells of the mammalian circadian clock. *Nature Neuroscience,* 3**,** 1241-7.

Davis, C. A., Hitz, B. C., Sloan, C. A., et al. 2018. The Encyclopedia of DNA elements (ENCODE): data portal update. *Nucleic acids research,* 46**,** D794-D801.

Davis, C. J., Bohnet, S. G., Meyerson, J. M. & Krueger, J. M. 2007. Sleep loss changes microRNA levels in the brain: a possible mechanism for state-dependent translational regulation. *Neuroscience letters,* 422**,** 68-73.

Delorme, J. E., Kodoth, V. & Aton, S. J. 2019. Sleep loss disrupts Arc expression in dentate gyrus neurons. *Neurobiology of Learning and Memory,* 160**,** 73-82.

Derrien, T., Johnson, R., Bussotti, G., et al. 2012. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Research,* 22**,** 1775-89.

Desbarats, L., Gaubatz, S. & Eilers, M. 1996. Discrimination between different E-box-binding proteins at an endogenous target gene of c-myc. *Genes and Development,* 10**,** 447-60.

Ditacchio, L., Le, H. D., Vollmers, C., et al. 2011. Histone lysine demethylase JARID1a activates CLOCK-BMAL1 and influences the circadian clock. *Science,* 333**,** 1881.

Duan, R., Liu, X., Wang, T., et al. 2016. Histone acetylation regulation in sleep deprivation-induced spatial memory impairment. *Neurochemical Research,* 41**,** 2223-2232.

Edgar, R. S., Stangherlin, A., Nagy, A. D., et al. 2016. Cell autonomous regulation of herpes and influenza virus infection by the circadian clock. *Proceedings of the National Academy of Sciences,* 113**,** 10085.

Eugene, A. R. & Masiak, J. 2015. The neuroprotective aspects of sleep. *MEDtube science,* 3**,** 35-40.

Ewels, P., Magnusson, M., Lundin, S. & Käller, M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics,* 32**,** 3047-8.

Fang, B., Everett, L. J., Jager, J., et al. 2014. Circadian enhancers coordinate multiple phases of rhythmic gene transcription in vivo. *Cell,* 159**,** 1140-52.

Fang, L.-L., Wang, X.-H., Sun, B.-F., et al. 2017. Expression, regulation and mechanism of action of the miR-17-92 cluster in tumor cells (Review). *International journal of molecular medicine,* 40**,** 1624-1630.

Felsenfeld, G., Boyes, J., Chung, J., et al. 1996. Chromatin structure and gene expression. *Proceedings of the National Academy of Sciences,* 93**,** 9384.

Fornes, O., Castro-Mondragon, J. A., Khan, A., et al. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research,* 48**,** D87-D92.

Forni, D., Pozzoli, U., Cagliani, R., et al. 2014. Genetic adaptation of the human circadian clock to day-length latitudinal variations and relevance for affective disorders. *Genome Biology,* 15**,** 499.

Foster, R. G., Hughes, S. & Peirson, S. N. 2020. Circadian Photoentrainment in Mice and Humans. *Biology,* 9**,** 180.

Franken, P. 2013. A role for clock genes in sleep homeostasis. *Current Opinion in Neurobiology,* 23**,** 864-872.

Freedman, M. S., Lucas, R. J., Soni, B., et al. 1999. Regulation of mammalian circadian behavior by non-rod, non-cone, ocular photoreceptors. *Science,* 284**,** 502-4.

Freeman, D., Sheaves, B., Goodwin, G. M., et al. 2017. The effects of improving sleep on mental health (OASIS): a randomised controlled trial with mediation analysis. *The Lancet Psychiatry,* 4**,** 749-758.

Freese, N. H., Norris, D. C. & Loraine, A. E. 2016. Integrated genome browser: visual analytics platform for genomics. *Bioinformatics,* 32**,** 2089-95.

Gau, D., Lemberger, T., Von Gall, C., et al. 2002. Phosphorylation of CREB Ser142 regulates light-induced phase shifts of the circadian clock. *Neuron,* 34**,** 245-53.

Gerashchenko, D., Pasumarthi, R. K. & Kilduff, T. S. 2017. Plasticity-related gene expression during eszopiclone-induced sleep. *Sleep,* 40.

Gery, S., Gombart, A. F., Yi, W. S., et al. 2005. Transcription profiling of C/EBP targets identifies Per2 as a gene implicated in myeloid leukemia. *Blood,* 106**,** 2827-2836.

Gill, K. 2016. Night working is a boom industry. *Labour Force Survey.* TUC.

Ginty, D., Kornhauser, J., Thompson, M., et al. 1993. Regulation of CREB phosphorylation in the suprachiasmatic nucleus by light and a circadian clock. *Science,* 260**,** 238-241.

Goodstadt, L. 2010. Ruffus: a lightweight Python library for computational pipelines. *Bioinformatics,* 26**,** 2778-2779.

Gordon, D. E., Jang, G. M., Bouhaddou, M., et al. 2020. A SARS-CoV-2 protein interaction map reveals targets for drug repurposing. *Nature,* 583**,** 459-468.

Goryshin, I. Y. & Reznikoff, W. S. 1998. Tn5 in vitro transposition. *Journal of Biological Chemistry,* 273**,** 7367-74.

Gräff, J., Kim, D., Dobbin, M. M. & Tsai, L.-H. 2011. Epigenetic regulation of gene expression in physiological and pathological brain processes. *Physiological Reviews,* 91**,** 603-649.

Guzmán-Marín, R., Suntsova, N., Stewart, D. R., et al. 2003. Sleep deprivation reduces proliferation of cells in the dentate gyrus of the hippocampus in rats. *The Journal of physiology,* 549**,** 563-571.

Halazonetis, T. D., Georgopoulos, K., Greenberg, M. E. & Leder, P. 1988. c-Jun dimerizes with itself and with c-Fos, forming complexes of different DNA binding affinities. *Cell,* 55**,** 917-24.

Hannibal, J., Møller, M., Ottersen, O. P. & Fahrenkrug, J. 2000. PACAP and glutamate are co-stored in the retinohypothalamic tract. *The Journal of Comparative Neurology,* 418**,** 147-55.

Hardin, P. E., Hall, J. C. & Rosbash, M. 1990. Feedback of the Drosophila period gene product on circadian cycling of its messenger RNA levels. *Nature,* 343**,** 536-540.

Harms, E., Kivimäe, S., Young, M. W. & Saez, L. 2004. Posttranscriptional and posttranslational regulation of clock genes. *Journal of Biological Rhythms,* 19**,** 361-373.

Hattar, S., Lucas, R. J., Mrosovsky, N., et al. 2003. Melanopsin and rod-cone photoreceptive systems account for all major accessory visual functions in mice. *Nature,* 424**,** 76-81.

Hauer, M. H. & Gasser, S. M. 2017. Chromatin and nucleosome dynamics in DNA damage and repair. *Genes and Development,* 31**,** 2204-2221.

Heintzman, N. D., Stuart, R. K., Hon, G., et al. 2007. Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome. *Nature Genetics,* 39**,** 311-318.

Heinz, S., Benner, C., Spann, N., et al. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Molecular Cell,* 38**,** 576-89.

Henrich, S. M., Usadel, C., Werwein, E., et al. 2017. Interplay with the Mre11-Rad50-Nbs1 complex and phosphorylation by GSK3β implicate human B-Myb in DNA-damage signaling. *Scientific Reports,* 7**,** 41663.

Hoekstra, M. M. B., Emmenegger, Y., Hubbard, J. & Franken, P. 2019. Cold-inducible RNA-binding protein (CIRBP) adjusts clock-gene expression and REM-sleep recovery following sleep deprivation. *eLife,* 8**,** e43400.

Hor, C. N., Yeung, J., Jan, M., et al. 2019. Sleep–wake-driven and circadian contributions to daily rhythms in gene expression and chromatin accessibility in the murine cortex. *Proceedings of the National Academy of Sciences,* 116**,** 25773-25783.

Horvath, S. & Raj, K. 2018. DNA methylation-based biomarkers and the epigenetic clock theory of ageing. *Nature Reviews Genetics,* 19**,** 371-384.

Hower, I. M., Harper, S. A. & Buford, T. W. 2018. Circadian Rhythms, Exercise, and Cardiovascular Health. *Journal of circadian rhythms,* 16**,** 7-7.

Husse, J., Hintze, S. C., Eichele, G., et al. 2012. Circadian clock genes Per1 and Per2 regulate the response of metabolism-associated transcripts to sleep disruption. *PLOS ONE,* 7**,** e52983.

Jabeen, A., Ahmad, N. & Raza, K. 2018. Machine learning-based state-of-the-art methods for the classification of RNA-seq data. *In:* DEY, N., ASHOUR, A. S. & BORRA, S. (eds.) *Classification in BioApps: Automation of Decision Making.* Cham: Springer International Publishing.

Jagannath, A., Butler, R., Godinho, Sofia i., et al. 2013. The CRTC1-SIK1 pathway regulates entrainment of the circadian clock. *Cell,* 54**,** 1100-1111.

Jagannath, A., Varga, N., Dallmann, R., et al. 2021. Adenosine integrates light and sleep signalling for the regulation of circadian timing in mice. *Nature Communications,* 12**,** 2113.

Jia, D., Jurkowska, R. Z., Zhang, X., et al. 2007. Structure of Dnmt3a bound to Dnmt3L suggests a model for de novo DNA methylation. *Nature,* 449**,** 248-251.

Kamakaka, R. T. & Thomas, J. O. 1990. Chromatin structure of transcriptionally competent and repressed genes. *The EMBO Journal,* 9**,** 3997-4006.

Kelley, R. I. 1973. Isolation of a histone IIb1–IIb2 complex. *Biochemical and Biophysical Research Communications,* 54**,** 1588-1594.

Kerppola, T. K. & Curran, T. 1994. Maf and Nrl can bind to AP-1 sites and form heterodimers with Fos and Jun. *Oncogene,* 9**,** 675-84.

Kervezee, L., Cuesta, M., Cermakian, N. & Boivin, D. B. 2018a. Simulated night shift work induces circadian misalignment of the human peripheral blood mononuclear cell transcriptome. *Proceedings of the National Academy of Sciences,* 115**,** 5540.

Kervezee, L., Kosmadopoulos, A. & Boivin, D. 2018b. Metabolic and cardiovascular consequences of shift work: The role of circadian disruption and sleep disturbances. *European Journal of Neuroscience,* 51.

Khalsa, S. B. S., Jewett, M. E., Cajochen, C. & Czeisler, C. A. 2003. A phase response curve to single bright light pulses in human subjects. *The Journal of Physiology,* 549**,** 945-952.

Koike, N., Yoo, S. H., Huang, H. C., et al. 2012. Transcriptional architecture and chromatin landscape of the core circadian clock in mammals. *Science,* 338**,** 349-54.

Kornberg, R. D. 1974. Chromatin structure: a repeating unit of histones and DNA. *Science,* 184**,** 868-871.

Kornhauser, J. M., Nelson, D. E., Mayo, K. E. & Takahashi, J. S. 1990. Photic and circadian regulation of c-fos gene expression in the hamster suprachiasmatic nucleus. *Neuron,* 5**,** 127-34.

Kuna, S. T., Maislin, G., Pack, F. M., et al. 2012. Heritability of performance deficit accumulation during acute sleep deprivation in twins. *Sleep,* 35**,** 1223-1233.

Lahtinen, A., Puttonen, S., Vanttola, P., et al. 2019. A distinctive DNA methylation pattern in insufficient sleep. *Scientific Reports,* 9**,** 1193.

Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biology,* 10**,** R25.

Leone, M. J., Marpegan, L., Duhart, J. M. & Golombek, D. A. 2012. Role of proinflammatory cytokines on lipopolysaccharide-induced phase shifts in locomotor activity circadian rhythm. *Chronobiology International,* 29**,** 715-23.

Levenson, J. M., O'riordan, K. J., Brown, K. D., et al. 2004. Regulation of histone acetylation during memory formation in the hippocampus. *Journal of Biological Chemistry,* 279**,** 40545-59.

Li, H., Handsaker, B., Wysoker, A., et al. 2009. The sequence alignment/map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

Liao, Y., Smyth, G. K. & Shi, W. 2013. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research,* 41**,** e108.

Lieber, M. R. 2010. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annual Review of Biochemistry,* 79**,** 181-211.

Lin, J. T., Kornhauser, J. M., Singh, N. P., et al. 1997. Visual sensitivities of nur77 (NGFI-B) and zif268 (NGFI-A) induction in the suprachiasmatic nucleus are dissociated from c-fos induction and behavioral phase-shifting responses. *Molecular Brain Research,* 46**,** 303-10.

Liu, Y., Zhou, J. & White, K. P. 2014. RNA-seq differential expression studies: more sequence or more replication? *Bioinformatics,* 30**,** 301-304.

Lonze, B. E. & Ginty, D. D. 2002. Function and regulation of CREB family transcription factors in the nervous system. *Neuron,* 35**,** 605-23.

Lourida, I., Hannon, E., Littlejohns, T. J., et al. 2019. Association of lifestyle and genetic risk with incidence of dementia. *JAMA,* 322**,** 430-437.

Love, M. I., Huber, W. & Anders, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15**,** 550.

Maret, S., Dorsaz, S., Gurcel, L., et al. 2007. Homer1a is a core brain molecular correlate of sleep losss. *Proceedings of the National Academy of Sciences,* 104**,** 20090.

Martínez De Paz, A., Sanchez-Mut, J. V., Samitier-Martí, M., et al. 2015. Circadian cycle-dependent MeCP2 and brain chromatin changes. *PLOS ONE,* 10**,** e0123693.

Massart, R., Freyburger, M., Suderman, M., et al. 2014. The genome-wide landscape of DNA methylation and hydroxymethylation in response to sleep deprivation impacts on synaptic plasticity genes. *Translational Psychiatry,* 4**,** e347.

Matsuo, T., Yamaguchi, S., Mitsui, S., et al. 2003. Control mechanism of the circadian clock for timing of cell division in vivo. *Science,* 302**,** 255-9.

Meijer, J. H. & Schwartz, W. J. 2003. In search of the pathways for light-induced pacemaker resetting in the suprachiasmatic nucleus. *Journal of Biological Rhythms,* 18**,** 235-49.

Millar-Craig, M., Bishop, C. & Raftery, E. B. 1978. Circadian variation of blood-pressure. *The Lancet,* 311**,** 795-797.

Minors, D. S., Waterhouse, J. M. & Wirz-Justice, A. 1991. A human phase-response curve to light. *Neuroscience Letters,* 133**,** 36-40.

Mitxelena, J., Apraiz, A., Vallejo-Rodríguez, J., et al. 2018. An E2F7-dependent transcriptional program modulates DNA damage repair and genomic stability. *Nucleic acids research,* 46**,** 4546-4559.

Miyazaki, S., Liu, C. Y. & Hayashi, Y. 2017. Sleep in vertebrate and invertebrate animals, and insights into the function and evolution of sleep. *Neuroscience Research,* 118**,** 3-12.

Mogilyansky, E. & Rigoutsos, I. 2013. The miR-17/92 cluster: A comprehensive update on its genomics, genetics, functions and increasingly important and numerous roles in health and disease. *Cell Death & Differentiation,* 20**,** 1603-1614.

Murillo-Rodríguez, E., Arankowsky-Sandoval, G., Barros, J. A., et al. 2019. Sleep and neurochemical modulation by DZNep and GSK-J1: potential link with histone methylation status. *Frontiers in Neuroscience,* 13.

Naidoo, N., Giang, W., Galante, R. J. & Pack, A. I. 2005. Sleep deprivation induces the unfolded protein response in mouse cerebral cortex. *Journal of Neurochemistry,* 92**,** 1150-7.

Narasimamurthy, R., Hunt, S. R., Lu, Y., et al. 2018. CK1δ/ε protein kinase primes the PER2 circadian phosphoswitch. *Proceedings of the National Academy of Sciences,* 115**,** 5986.

Nedeltcheva, A. V., Kessler, L., Imperial, J. & Penev, P. D. 2009. Exposure to recurrent sleep restriction in the setting of high caloric intake and physical inactivity results in increased insulin resistance and reduced glucose tolerance. *The Journal of clinical endocrinology and metabolism,* 94**,** 3242-3250.

Ng, S. S., Yue, W. W., Oppermann, U. & Klose, R. J. 2008. Dynamic protein methylation in chromatin biology. *Cellular and Molecular Life Sciences,* 66**,** 407.

Obrietan, K., Impey, S. & Storm, D. R. 1998. Light and circadian rhythmicity regulate MAP kinase activation in the suprachiasmatic nuclei. *Nature Neuroscience,* 1**,** 693-700.

Ode, K. L. & Ueda, H. R. 2020. Phosphorylation hypothesis of sleep. *Frontiers in Psychology,* 11**,** 2700.

Ohayon, M. M., Lemoine, P., Arnaud-Briant, V. & Dreyfus, M. 2002. Prevalence and consequences of sleep disorders in a shift worker population. *Journal of Psychosomatic Research,* 53**,** 577-583.

Okano, M., Bell, D. W., Haber, D. A. & Li, E. 1999. DNA methyltransferases Dnmt3a and Dnmt3b are essential for de novo methylation and mammalian development. *Cell,* 99**,** 247-257.

Oki, M., Aihara, H. & Ito, T. 2007. Role of histone phosphorylation in chromatin dynamics snd its implications in diseases. *In:* KUNDU, T. K., BITTMAN, R., DASGUPTA, D., ENGELHARDT, H., FLOHE, L., HERRMANN, H., HOLZENBURG, A., NASHEUER, H. P., ROTTEM, S., WYSS, M. & ZWICKL, P. (eds.) *Chromatin and Disease.* Dordrecht: Springer Netherlands.

Parsons, M. J., Brancaccio, M., Sethi, S., et al. 2015. The regulatory factor ZFHX3 modifies circadian function in SCN via an AT motif-driven axis. *Cell,* 162**,** 607-21.

Pattaroni, C. & Jacob, C. 2013. Histone methylation in the nervous system: functions and dysfunctions. *Molecular Neurobiology,* 47**,** 740-56.

Pembroke, W. G., Babbs, A., Davies, K. E., et al. 2015. Temporal transcriptomics suggest that twin-peaking genes reset the clock. *eLife,* 4**,** e10518.

Perelis, M., Marcheva, B., Moynihan Ramsey, K., et al. 2015. Pancreatic β cell enhancers regulate rhythmic transcription of genes controlling insulin secretion. *Science,* 350.

Phillips, A. J. K., Vidafar, P., Burns, A. C., et al. 2019. High sensitivity and interindividual variability in the response of the human circadian system to evening light. *Proceedings of the National Academy of Sciences,* 116**,** 12019.

Porterfield, V. M., Piontkivska, H. & Mintz, E. M. 2007. Identification of novel light-induced genes in the suprachiasmatic nucleus. *BMC Neuroscience,* 8**,** 98.

Poulsen, R. C., Hearn, J. I. & Dalbeth, N. 2021. The circadian clock: a central mediator of cartilage maintenance and osteoarthritis development? *Rheumatology,* 60**,** 3048-3057.

Pregueiro, A. M., Liu, Q., Baker, C. L., et al. 2006. The Neurospora checkpoint kinase 2: A regulatory link between the circadian and cell cycles. *Science,* 313**,** 644-9.

Preitner, N., Damiola, F., Lopez-Molina, L., et al. 2002. The orphan nuclear receptor REV-ERBalpha controls circadian transcription within the positive limb of the mammalian circadian oscillator. *Cell,* 110**,** 251-60.

Quigley, M. E. & Yen, S. S. C. 1979. A mid-day surge in cortisol levels. *The Journal of Clinical Endocrinology & Metabolism,* 49**,** 945-947.

Ralph, M. R. & Menaker, M. 1988. A mutation of the circadian system in golden hamsters. *Science,* 241**,** 1225-7.

Ramírez, F., Ryan, D. P., Grüning, B., et al. 2016. deepTools2: A next generation web server for deep-sequencing data analysis. *Nucleic Acids Research,* 44**,** W160-5.

Rapkin, L. M., Anchel, D. R. P., Li, R. & Bazett-Jones, D. P. 2012. A view of the chromatin landscape. *Micron,* 43**,** 150-158.

Refinetti, R. & Menaker, M. 1992. The circadian rhythm of body temperature. *Physiology & Behavior,* 51**,** 613-637.

Riedel, C. S., Georg, B., Jørgensen, H. L., et al. 2018. Mice lacking EGR1 have impaired clock gene (BMAL1) oscillation, locomotor activity, and body temperature. *Journal of Molecular Neuroscience,* 64**,** 9-19.

Risso, D., Ngai, J., Speed, T. P. & Dudoit, S. 2014. Normalization of RNA-seq data using factor analysis of control genes or samples. *Nature Biotechnology,* 32**,** 896-902.

Roenneberg, T. & Merrow, M. 2002. Life before the clock: Modeling circadian evolution. *Journal of Biological Rhythms,* 17**,** 495-505.

Ruifeng, D., Liu, X., Wang, T., et al. 2016. Histone acetylation regulation in sleep deprivation-induced spatial memory impairment. *Neurochemical Research,* 41.

Russell, W. M. S. W. M. S. & Burch, R. L. R. L. 1959. Principles of humane experimenal technique.

Ryseck, R. P. & Bravo, R. 1991. c-JUN, JUN B, and JUN D differ in their binding affinities to AP-1 and CRE consensus sequences: effect of FOS proteins. *Oncogene,* 6**,** 533-42.

Sachs, A. B. 1993. Messenger RNA degradation in eukaryotes. *Cell,* 74**,** 413-421.

Schwartz, W. J., Tavakoli-Nezhad, M., Lambert, C. M., et al. 2011. Distinct patterns of Period gene expression in the suprachiasmatic nucleus underlie circadian clock photoentrainment by advances or delays. *Proceedings of the National Academy of Sciences,* 108**,** 17219-17224.

Shi, S.-Q., Ansari, T., Mcguinness, O. P., et al. 2013. Circadian disruption leads to insulin resistance and obesity. *Current biology : CB,* 23**,** 372-381.

Sng, J. C., Taniura, H. & Yoneda, Y. 2004. A tale of early response genes. *Biological and Pharmaceutical Bulletin,* 27**,** 606-12.

Soshnev, A. A., Ishimoto, H., Mcallister, B. F., et al. 2011. A conserved long noncoding RNA affects sleep behavior in Drosophila. *Genetics,* 189**,** 455-468.

Su, Y., Shin, J., Zhong, C., et al. 2017. Neuronal activity modifies the chromatin accessibility landscape in the adult brain. *Nature Neuroscience,* 20**,** 476-483.

Supek, F., Bošnjak, M., Škunca, N. & Šmuc, T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLOS ONE,* 6**,** e21800.

Surrallés, J., Puerto, S., Ramírez, M. J., et al. 1998. Links between chromatin structure, DNA repair and chromosome fragility. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis,* 404**,** 39-44.

Suzuki, A., Yanagisawa, M. & Greene, R. W. 2020. Loss of Arc attenuates the behavioral and molecular responses for sleep homeostasis in mice. *Proceedings of the National Academy of Sciences,* 117**,** 10547.

Takahashi, J. S. 2017. Transcriptional architecture of the mammalian circadian clock. *Nature Reviews Genetics,* 18**,** 164-179.

Takeuchi, J., Shannon, W., Aronin, N. & Schwartz, W. J. 1993. Compositional changes of AP-1 DNA-binding proteins are regulated by light in a mammalian circadian clock. *Neuron,* 11**,** 825-836.

Team, R. C. 2017. R: A language and environment for statistical computing. *R Foundation for Statistical Computing.*

Terao, A., Steininger, T. L., Hyder, K., et al. 2003. Differential increase in the expression of heat shock protein family members during sleep deprivation and during sleep. *Neuroscience,* 116**,** 187-200.

Terao, A., Wisor, J. P., Peyron, C., et al. 2006. Gene expression in the rat brain during sleep deprivation and recovery sleep: an Affymetrix GeneChip® study. *Neuroscience,* 137**,** 593-605.

Thomas, J., Overeem, S. & Claassen, J. a. H. R. 2019. Long-term occupational sleep loss and post-retirement cognitive decline or dementia. *Dementia and Geriatric Cognitive Disorders,* 48**,** 105-112.

Thompson, C., Wisor, J., Lee, C.-K., et al. 2010. Molecular and anatomical signatures of sleep deprivation in the mouse brain. *Frontiers in Neuroscience,* 4**,** 165.

Trivedi, M. S., Holger, D., Bui, A. T., et al. 2017. Short-term sleep deprivation leads to decreased systemic redox metabolites and altered epigenetic status. *PLOS ONE,* 12**,** e0181978.

Vitaterna, M. H., Ko, C. H., Chang, A. M., et al. 2006. The mouse Clock mutation reduces circadian pacemaker amplitude and enhances efficacy of resetting stimuli and phase-response curve amplitude. *Proceedings of the National Academy of Sciences of the United States of America,* 103**,** 9327-32.

Voigt, R. M., Forsyth, C. B. & Keshavarzian, A. 2019. Circadian rhythms: a regulator of gastrointestinal health and dysfunction. *Expert Review of Gastroenterology & Hepatology,* 13**,** 411-424.

Wan, Y., Shang, J., Graham, R., et al. 2020. Receptor recognition by the novel coronavirus from wuhan: an analysis based on decade-long structural studies of SARS coronavirus. *Journal of Virology,* 94.

Wang, H., Liu, Y., Briesemann, M. & Yan, J. 2010. Computational analysis of gene regulation in animal sleep deprivation. *Physiological Genomics,* 42**,** 427-436.

Welsh, D. K., Takahashi, J. S. & Kay, S. A. 2010. Suprachiasmatic nucleus: cell autonomy and network properties. *Annual review of physiology,* 72**,** 551-577.

White, A. J., Kresovich, J. K., Xu, Z., et al. 2019. Shift work, DNA methylation and epigenetic age. *International Journal of Epidemiology,* 48**,** 1536-1544.

Williams, N. M., O'donovan, M. C. & Owen, M. J. 2005. Is the dysbindin gene (DTNBP1) a susceptibility gene for schizophrenia? *Schizophrenia Bulletin,* 31**,** 800-5.

Wisor, J. P., O'hara, B. F., Terao, A., et al. 2002. A role for cryptochromesin sleep regulation. *BMC Neuroscience,* 3**,** 20.

Wu, H., Dunnett, S., Ho, Y.-S. & Chang, R. C.-C. 2019. The role of sleep deprivation and circadian rhythm disruption as risk factors of Alzheimer's disease. *Frontiers in Neuroendocrinology,* 54**,** 100764.

Xia, L., Ma, S., Zhang, Y., et al. 2015. Daily variation in global and local DNA methylation in mouse livers. *PLOS ONE,* 10**,** e0118101.

Xie, L., Huang, R., Liu, S., et al. 2019. A positive feedback loop of SIRT1 and miR17HG promotes the repair of DNA double-stranded breaks. *Cell Cycle,* 18**,** 2110-2123.

Xu, Y., Guo, W., Li, P., et al. 2016. Long-range chromosome interactions mediated by cohesin shape circadian gene expression. *PLOS Genetics,* 12**,** e1005992.

Yagita, K. & Okamura, H. 2000. Forskolin induces circadian gene expression of rPer1, rPer2 and dbp in mammalian rat-1 fibroblasts. *FEBS Letters,* 465**,** 79-82.

Yamazaki, S., Numano, R., Abe, M., et al. 2000. Resetting central and peripheral circadian oscillators in transgenic rats. *Science,* 288**,** 682.

Yoo, S.-S., Hu, P. T., Gujar, N., et al. 2007. A deficit in the ability to form new human memories without sleep. *Nature Neuroscience,* 10**,** 385-392.

Yoshitane, H., Asano, Y., Sagami, A., et al. 2019. Functional D-box sequences reset the circadian clock and drive mRNA rhythms. *Communications Biology,* 2**,** 300.

Zada, D., Bronshtein, I., Lerer-Goldshtein, T., et al. 2019. Sleep increases chromosome dynamics to enable reduction of accumulating DNA damage in single neurons. *Nature Communications,* 10**,** 895.

Zhang, R., Lahens, N. F., Ballance, H. I., et al. 2014. A circadian gene expression atlas in mammals: Implications for biology and medicine. *Proceedings of the National Academy of Sciences,* 111**,** 16219-16224.

Zhang, Y., Liu, T., Meyer, C. A., et al. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biology,* 9**,** R137.

Zhou, X., Maricque, B., Xie, M., et al. 2011. The human epigenome browser at washington university. *Nature Methods,* 8**,** 989-990.