Bioinformatic approaches to chromatin structure and RNA editing in *Drosophila melanogaster*

A Thesis Presented in Partial Fulfillment of

the Honors Bachelor's Degree

Christa Caggiano

Biological Physics Department

Under the supervision of Michael Rosbash

Brandeis University

May 2017

Contents

List of Figures

List of Tables

Acknowledgments

There are many people I am extremely grateful for help on this thesis:

- Professor Michael Rosbash, for guidance, patience, and support for doing a bioinformatic thesis in his lab.
- Professor Paul Miller for his support on my committee.
- Dr.Kate Abruzzi for reading countless drafts and offering excellent guidance over the years and through life-crises.
- Dr. Reazur Rahman for help with implementing many bioinformatic algorithms and for advice on both these projects.
- Felipe Escobedo, for his collaboration on the ATAC-seq portion of this thesis, and for being a wonderful biologist and friend.
- Arya Boudaie, for proofreading and numerous hours spent debugging and fixing code.
- And my family and friends for support through illness, self-doubt and complaints.

Chapter 1

Chromatin Accessibility of Circadian Regulated Genes in *Drosophila*

Abstract

Nearly all organisms have a circadian rhythm, which is a 24-hour period where physical and behavioral states oscillate in a regular sinusoidal manner. These oscillations are controlled by a molecular clock, which is driven by two key heterodimeric transcription factors: Clock (CLK) and Cycle (CYC). CLK/CYC bind to and activate target genes in a rhythmic manner: CLK/CYC binding is highest in the early night and lowest in the early morning. Recent results suggest that the mammalian homolog of CLK, Bmal1, may function to "open" chromatin to facilitate transcription activation. To test whether this hypothesis is true in Drosophila brains, ATAC-seq was used to examine how chromatin structure changes throughout the day. Surprisingly, initial visual inspection of ATAC-seq data indicated that chromatin accessibility on CLK controlled genes was independent of CLK binding and did not change with time of day. To explore this finding, we used bioinformatic analyses to quantify chromatin conformation on CLK controlled genes (CCGs) throughout the day. Sequencing and peak calling were optimized specifically for ATAC-seq data. Differential analysis was performed on peak data and demonstrated that CLK-regulated genes are not significantly more accessible in the presence of CLK. Large scale coverage analyses further supported the conclusion that chromatin accessibility is independent of CLK binding. More detailed statistical analyses revealed that a more complex mechanism may be opening CCGs, consisting of multiple proteins acting intricately throughout the day.

1.1 Introduction

Circadian rhythms are biological processes that contribute to the oscillatory behaviors experienced by most living organisms. The word circadian comes from a Latin root meaning "about a day." This references the fact that most circadian rhythms are approximately 24-hour cycles. Examples of such oscillatory behavior are sleep, cell production, and hormone production. Circadian rhythms must be endogenous and entrained by an internal timekeeping system. This allows for the organism to anticipate changes in the physical environment (Vitaterna et al. 2016). Circadian rhythms are often correlated with environmental cues- such as day/night transitions- but they do not need these cues to function. When a circadian rhythm matches some external cue, this is referred to as entrainment. If the entrainment of the endogenous clock is not synced with the external environment, this can lead to dramatic changes in the organism's behavior. The study of circadian rhythms, and whether or not they are appropriately synced, is important to the understanding of many biological and bodily processes.

Drosophila melanogaster is an ideal model organism to study circadian rhythms. Unlike mammals, they are easy to cultivate in laboratory conditions and are inexpensive to care for. Many useful laboratory techniques also have been developed to study *Drosophila*. Coupled with the short life cycle of *Drosophila*, this means that genetic manipulations can be created and observed with ease by scientists in a controlled environment (Jennings 2011). This is important, as the entrainment of the circadian clock must be controlled in order to make the study of circadian rhythms meaningful. Furthermore, circadian machinery found in *Drosophila* is conserved to mammals. Core proteins in the *Drosophila* circadian rhythm have analogs in mammals (Table 1) meaning that *Drosophila* can provide useful insights into the study of circadian rhythms at large.

The molecular basis of circadian rhythms has been extensively studied in *Drosophila*. Its molecular clock is driven by an evolutionary conserved transcriptional feedback loop (Rosato et al 2006). Two heterodimeric positive transcription factors, Clock (CLK) and Cycle (CYC), or Clock (CLK) and BMAL1 in mammals, bind to ebox motifs (enhancer regions that are recognized by certain transcription factors) in the promoters of their target clock controlled genes (CCGs) and activate their transcription. In *Drosophila*, two CCGs, *period* and *timeless*, encode proteins that function in a negative feedback loop to repress their own transcription (Figure 1). Once translated, PER and TIM exit the nucleus into the cytoplasm. They then re-enter the nucleus and bind to CLK/CYC on CCG promoters and inhibit transcription. In the early morning, the light activated protein, CRY, drives the degradation of TIM and PER, releasing the repression of transcription and allowing the cycle to start. CLK/CYC act on thousands of CCGs, contributing to a vast transcriptional

output regulated by the circadian clock (Abruzzi et al 2011). This mechanism drives the cycling expression of many different genes, approximately 150 in *Drosophila* (McDonald and Rosbash 2001) (Rodriguez et al 2013).

1.2 Chromatin Dynamics in Circadian Regulated Genes

Cycling gene expression driven by CLK/CYC raises the possibility that chromatin conformation could also change throughout the day. Mammalian CLK has a histone acetyl transferase activity and acts as a pioneering transcription factor, a particular class of transcription factor that can bind to condensed chromatin causing conformation changes (Taylor and Hardin 2008) (Bellet and Sassone-Corisi 2010). These events are associated with RNA polymerase binding, inducing transcriptional activation (Taylor and Hardin 2008). Recent studies in mammalian liver illustrate that chromatin on CCGs undergoes daily oscillations from open, which allows transcription to occur, to closed, where transcription is repressed (Menet et al 2014). BMAL1, the mammalian homolog of CLK, has been shown to open the chromatin in mouse liver, thus allowing transcription to occur (Figure 2). In Drosophila, however, there has been no evidence of CLK as a pioneering transcription factor indicating that cycling could be facilitated by other mechanisms of the circadian machinery. This led us to pose the specific question in Drosophila: Does CLK/CYC function to open chromatin to allow transcriptional activation?

In Drosophila CHIP-seq reveals that CLK binding patterns cycle oscillate over 24-hours (Figure 3). To address whether chromatin accessibility oscillates with CLK binding, we have used an Assay for Transposase Accessible Chromatin (ATAC-seq) to examine chromatin conformation in Drosophila brains (Buenrostro et al 2011). ATAC-seq probes the genome for regions where the chromatin is "open," and would be accessible to transcription factors (Figure 4). Previous experiments show that CLK/CYC bind to the chromatin and activate transcription of CCGs in the evening, causing transcription to be repressed from late night to mid-morning (Abruzzi et al 2011).

Surprisingly, initial visual inspection of ATAC-seq data indicated that chromatin accessibility on CLK controlled genes was independent of CLK binding and did not change with time of day. To explore this finding, we used bioinformatic analyses to quantify chromatin conformation on CCGs throughout the day. Sequencing and peak calling were optimized specifically for ATAC-seq data. Differential analysis was performed on peak data and demonstrated that CLK-regulated genes are not significantly more accessible in the presence of CLK. Large scale coverage analyses further supported the conclusion that chromatin accessibility

is independent of CLK binding. More detailed statistical analyses revealed that a more complex mechanism may be opening CCGs, consisting of multiple proteins acting intricately throughout the day.

1.3 Results

1.3.1 Sequencing, Data Mapping, and Quality Control

ATAC-seq experiments were carried out with three replicates, of 6 time points each (Table 2). The six time points (ZT2, ZT6, ZT10, ZT14, ZT18 and ZT22) were each four hours apart, thus spanning a 24-hour period. In this paradigm, ZT2-ZT10 are time points were the flies are in light, and ZT14-ZT22 are dark time points. These samples were sequenced using single-end Illumina sequencing. Although paired-end sequencing is often recommended for high-throughput ATAC-seq experiments to help resolve biases from PCR duplicates and inaccurate mapping to the genome, it is also more expensive. After comparing paired-end and single-end sequencing, we found that single-end sequencing was sufficient to gain insight into chromatin conformation in Drosophila ATAC-seq libraries. We achieved high quality samples and all libraries used had acceptable coverage (Figure 5).

Reads were trimmed to dynamically to remove adapters using Trim Galore . This approach is widely suggested as a method to reduce bias in some experiments and has been adopted by the ENCODE project (Lee 2015). Trimming reduced adapter read-through and increased percent mapping in our experiments. Trimmed sequencing files were then aligned to the dm3 genome using Bowtie2 and default parameters (Langmead and Saltzber 2012). This produced 6 complete libraries for each of the three replicates (Figure 6).

Library correlation was checked in multiple ways to ensure quality. ATAC-seq libraries must have significant read depth to reduce noise to signal ratios. This is partly because ATAC-seq assays produce high mitochondrial DNA contamination, interfering with the ability of the library to map. In our datasets, two libraries were sequenced twice and data pooled together to have appropriate depth. This increased read depth to 1-2 million reads on average for each of the three libraries in this study. The remaining dataset had sufficient coverage with only one round of sequencing. To test the reproducibility of these three libraries, we used the Irreproducible Discovery Rate (IDR). This method uses a curve to dynamically assess when findings are no longer consistent between replicates (Li et al 2011). All three replicates used in this study were found to be highly consistent with an IDR¡0.3 (Figure 7). We also used Pearson's correlation coefficient analysis to quantify the reproducibility between our samples. All libraries were highly correlated (R¿0.95)

(Table 3).

Libraries were visualized using bigwig files in the Integrated Genome Viewer (Robinson et all 2011). There was no difference between the visualization of the six time points for any of the three datasets studied. Four core circadian genes, *timeless*, *period*, *virile*, and *pdp1*, were examined and results were consistent across these four genes (Figure 8). As this is a qualitative assessment, it leaves the question as to whether there are differences not able to be detected by visual inspection alone.

1.3.2 Peak Calling and Quantification

To confirm or refute this visual finding, we chose to quantify chromatin accessibility. For this study, chromatin openness was defined as how many reads were in each region of interest. Reads often appear in groups, described as peaks, that are distinct from background reads. These are regions of high signal, where the transposase would have access to the chromatin that often correlate with transcription factor binding sites, as seen in our data. Peaks are called computationally; there are several peak calling algorithms available.

Choice of a peak calling algorithm is a crucial step in the analysis of ATAC-seq data, as there is significant variability in the sensitivity and specificity of the various algorithms available (Koohy et al 2014). The original ATAC-seq paper by Buenestro et al., utilized Zero-Inflated Negative Binomial Algorithm (ZINBA) (Rashid et al 2011). While this has been touted as a general purpose peak caller, ZINBA has been demonstrated to have major draw-backs. It is a memory-intensive algorithm, using up to 4.5x as much memory as other comparable peak callers. This can lead to exceedingly long run times – sometimes on the order of days to successfully call peaks. Furthermore, it has been demonstrated in a review of peak callers of DNase-seq data, that ZINBA produces consistently low True Positive Rates compared to other available algorithms (Koohy et al 2014). Since ATAC-seq data is comparable to DNAse-seq data, ZINBA may not be an ideal peak caller for ATAC-seq.

Model-based Analysis of ChIP-Seq or MACS is a popular peak calling algorithm that uses dynamic Poisson distributions to capture variability and to predict peaks more effectively (Zhang et al 2008). This is a strong advantage as it means MACS can capture local biases in sequence, making it more sensitive than other non-dynamic algorithms. It is the recommended peak caller by the authors in the newest version of the ATAC-seq protocol (Buenrostro et al 2015). MACS2, however, has many parameters that have been shown to significantly affect the number of ATAC-seq peaks called (Ackermann et al 2016). To ensure accurate peak calling several options were explored in this study, including regular/broad peak calling, p-value vs q-

value cutoffs, and window extension. Window extension was the only parameter which affected the outcome significantly. The MACS2 documentation suggests that in DNAse-seq data, the window for calling peaks must be extended in both directions from the peak summit to smooth pile-up signals (Liu 2017). Several ATAC-seq pipelines extend their window, including the ENCODE pipeline for the analysis of ATAC-seq which is the recommended protocol for all international groups working as part of this consortium or funded by the National Human Genome Research Institute.

MACS2 peak calling with and without window extension was tested on a sample of our ATAC-seq data at the ZT2 timepoint. Both approaches performed similarly in terms of types of peaks called, mappability, and distance to transcription start-site. As expected, extending the window size by 200bp, significantly decreased the number of peaks called, as it merges smaller peaks together that would be uniquely represented without extending the window size. Additionally, it decreased the number of peaks called per gene (Table 4).

We chose to continue using MACS2 peak calls without window extension because it produced a greater number of peaks and more peaks per gene. Extending the window size appears to be more conservative in calling significant peaks and calls less peaks/gene. In the context of this investigation, more peaks was desired as each peak could represent a different transcription factor binding site. We were particularly interested in CLK binding and cycling. Therefore, a less conservative approach seems warranted in order to capture the sensitivity of transcription factor binding.

1.3.3 Coverage Plots

As ATAC-seq analysis is sometimes biased by high signal to noise ratios, small changes in chromatin accessibility may not be distinguishable in a gene-by-gene basis. Clear trends, however, may appear in aggregate. This is supported in mammalian data, where differences in chromatin accessibility were only seen when the data from hundreds of genes was analyzed in aggregate. To strengthen confidence in our assessment that chromatin is static on circadian genes across the day, we examined ATAC-seq signals and how they change throughout the day in the region surrounding the summit of CLK-binding for a set of CLK-controlled genes.

The summit of CLK peaks was found for the top 200 CLK controlled genes top 100 genes were identified based on the strength of CLK binding from ChIP-seq data (personal communication, W. Luo). All peaks were extended +/-500 base pairs around this summit to produce a clear picture of the chromatin conformation in the region surrounding CLK binding. The total ATAC-seq read counts in these 200 regions were averaged. Data for each of the six time points from all three datasets were pooled together by averaging to provide

a stronger trend in the data. Control data was produced in a similar manner, but used the summed signal around 200 randomly chosen transcription start sites (TSS) using the same data.

Data was fit using a Local Polynomial Regression (LOESS) to reduce the noise observed when the data were plotted. This smoothed our data by producing a weighted quadratic least squares regression over our values. The uniqueness of LOESS is that it a fits a simple regression to small subsets of the data to build a larger function that is truly representative of the variation (Cleveland 1979). We used a modest smoothing parameter of 0.1 in our implementation to prevent over-fitting of data.

Coverage plots of the Top 200 CLK direct targets showed a small difference between dark and light time points. In specific, ZT22 and ZT18 have higher levels of signal than ZT2-ZT10. Interestingly, ZT14 has the least signal of all time points (Figure 9). CLK controlled genesCCGs had significantly higher signal than the TSS controls, but the TSS did not show the same pattern of difference between the time points. In the TSS plot, time points are clustered together except ZT14, which has slightly less signal. This result was in contrast to similar mammalian studies, which had a clear and significant difference in chromatin conformation on CCGs throughout the day.

We chose to next examine smaller windows of +/-200bp. A smaller window size revealed a very clear segregation of time points, with ZT18 and ZT22 having slightly more signal than the other four time points (Figure 10). This difference was hard to interpret as ZT18/ZT22 represent the time of lowest CLK binding and therefore the time of least chromatin accessibility i.e. it was the opposite of the expected result.

Coverage plots of CLK deletion libraries were also generated in order to examine whether the absence of CLK changed the chromatin conformation. At both ZT2 and ZT14, there was virtually no difference between chromatin accessibility in a clock deletion (CLKOUT) and the wild-type control (Figure 11). CYC mutants, cyc01, produced similar results for ZT2 and ZT14, suggesting that neither CYC or CLK is required for these chromatin states.

1.3.4 Differential analysis of chromatin accessibility

To further investigate the role CLK in chromatin conformation, differential peak height analysis was performed on time pointssamples at with the highest CLK binding (ZT14) and the lowest (ZT2ZT2 and ZT14). CLK present and time points without CLK. Specifically, we wanted to quantitatively determine differences between groups with and without CLK. Differences between peak heights can be assessed qualitatively using a genome browser. The peaks aligned with the ebox binding site for CLK but there was no visual difference between peaks across time points. This is a surprising finding as CLK levels have been demonstrated to

cycle. To support or refute this visual finding, we chose to quantity peak differences using software for differential expression analysis. In order to carry out differential analysis the number of reads mapped for each gene needed to be assessed. The greater the number of genomic reads for a region, the more accessible the region is to transposase, thus the more the accessible the chromatin. The FeatureCounts algorithm was used to find read counts for each of our libraries. While HTseq counts is a very popular algorithm, we found it had a much slower runtime than FeatureCounts, and produced nearly identical output to the FeatureCounts algorithm. This aligns with the findings of the authors of FeatureCounts (Liao et al 2015).

This was appropriate for our data, as even though most differential expression analyses were developed for RNA-seq data, the statistical methods for determining abundance are easily generalized because they rely on finding general linear models for regions with differing choices. EdgeR was our chosen differential expression package because it uses quantile-adjusted conditional maximum likelihood to determine differences, a method that has been demonstrated to perform well with small sample pools, ideal for our three replicates (Chen et al 2016). Additionally, EdgeR was a very fast algorithm and produced intuitive output relative to other differential expression packages.

Using FeatureCount generated raw read counts, we compared peaks between ZT2 (peak of CLK expression) and ZT14 (nadir of CLK binding) using edgeR. The output produced an exact test and a generalized linear model that assessed fold change and the significance threshold of that fold change. While all CCGs in this study displayed a fold change, the fold change did not consistently increase or decrease between time points, an unexpected finding in light of known CLK binding patterns. Only one of these peaks had a significant fold change (p¡0.05) (Table 5). In coverage plot data, there was slight difference between ZT22 and ZT6/ZT10. Yet, similar differential analyses between ZT6 and ZT22, or ZT10 and ZT22 yielded a similar result: no significant fold change on the four core circadian genes. Besides these core circadian genes, 18 CCGs in total showed a significant change between ZT2 and ZT14. Eleven of these genes went up with CLK expression. Nonetheless, there are many more CCGs with no significant difference in fold change between time points.

Many non-CCGs changed significantly across circadian time. Gene ontologies of the top peaks with most significant fold changes were examined using DAVID using ATAC-seq peaks from ZT2 as background. The ontologies, however, revealed no clear trend. Both top negative and positive fold change groups were primarily enriched for modestly cytoplasmic ontologies (0.93 and 1.69 respectively). Since genes with a role in cytoplasm are a broad category, this did not shed more light on chromatin conformation changes across circadian time.

1.4 Conclusion

While recent results in mammals suggest a role for CLK in opening chromatin, contributing to overall chromatin accessibility throughout the day in circadian controlled genes, this study demonstrates that there is no clear role for CLK in Drosophila chromatin conformation. It was previously hypothesized that CCG chromatin accessibility would follow CLK binding patterns, such that it would be most open in the night, when CLK levels are highest, and closed in the morning, when CLK levels are low. Coverage plots, however, suggest that while chromatin accessibility does change, it is nearly opposite of what would be expected. ZT14 time points consistently showed the lowest signal of all time points. ZT18 and ZT22 had high signal, which is interesting in light of CLK binding, as these time points have approximately equal CLK binding to the binding at ZT2 and ZT6 time points, yet, ZT2 and ZT6 have much lower signal. This reveals that chromatin accessibility in CCGs is not static, as previously assessed qualitatively, but may be independent of CLK binding. Chromatin accessibility in CLK deletions between ZT2 and ZT14 time points showed virtually no difference compared to a control, which further supports this conclusion. Since CLK deletions did not alter chromatin states, it is not surprising that a CYC deletion showed a similar lack of change, as it is known that CLK and CYC form a heterodimer to interact with chromatin.

Similarly, differential analyses reveal a lack difference between chromatin across circadian time. In the comparison of ZT2 and ZT14, only one peak on core circadian genes changed significantly, and only 18 CCGs changed significantly out of more than 200. Differential analyses between ZT10 and ZT22 and ZT6 and ZT22 yielded similar results, suggesting the lack of difference is not a phenomenon limited to the ZT2/ZT14 time points. While the results of this analysis at first seem to contradict the results of the coverage plots, which indicated a difference between time points, it really seems to indicate the importance of doing large scale analyses. Looking at a gene by gene basis is difficult and perhaps reductive. Genes, including circadian genes, exist in complex networks that are difficult to ascertain from individuals. Coverage plots examined many genes at once, capturing trends that would be difficult to find by hand.

Overall, the results of the coverage plot analyses can suggest that CLK is not a pioneering transcription factor. While chromatin states may not be completely independent of CLK binding, it is unlikely from these results that CLK is causing any chromatin change. This is an unexpected result that may suggest that multiple unknown proteins may be binding to associated circadian enhancer regions, or ebox regions. Clockwork orange (CWO), a transcriptional repressor and circadian pacemaker element, is an example of such a candidate (Kadener et al 2007). Recent work has demonstrated that CWO binds anti-phase to CLK (Zhou et al 2016) (Figure 12). This is extremely compelling as it could partially explain the phenomenon we

witness where chromatin changes only slight over circadian time. A reasonable hypothesis could be that CLK is opening chromatin in the absence of CWO, and vice versa. Preliminary results indicate that in a CWO knockdown, the chromatin is more open relative to a control at ZT2, which is line with this hypothesis, since this is when CWO levels are high. Chromatin is more closed relative to a control in a CWO knockdown at ZT14, however, which would not be expected under this hypothesis. It would be anticipated that high CLK levels would promote chromatin opening regardless of CWO levels (Figure 13). Nonetheless, this example illustrates that chromatin dynamics of CCGs may be more complicated than previously expected.

This is further supported by preliminary motif analysis. CENTIPEDE, an algorithm for detecting transcription factor binding sites, found enrichment of Trithorax-like (Trl) on CCGs. Trl is a GAGA transcription factor that has been implicated in chromatin modification, making it a compelling candidate. Trl is, however, also widely and non-specifically expressed in Drosophila. Preliminary results show that while Trl knockdown does not change chromatin accessibility on CCGs, it may be changing the Drosophila chromatin landscape more widely (Figure 14).

In conclusion, there is no defined role of CLK in circadian chromatin conformation to date. This study presents a coherent bioinformatic workflow for ATAC-seq data. Our workflow can allow for a wide range of potential studies that can further elucidate information on chromatin conformation in Drosophila.

1.5 Methods

1.5.1 Datasets

ATAC-seq data sets were generated on brain samples from male/female Canton-S wildtype, CLK-out and CYC01. Six time points were assayed, ZT2, ZT6, ZT10, ZT14, ZT18, and ZT22. All flies were entrained in a twelve hour light twelve hour dark paradigm. The libraries were prepared according to the ATAC-seq protocol outlined in Buenestro et. al with an empirically derived PCR cycling protocol (Buenrostro et al 2013).

1.5.2 Data processing

Reads were trimmed to and libraries were aligned to the dm3 assembly of the Drosophila genome using default parameters of bowtie2. Unmapped reads were removed. Visualization files in bigWig format were viewed in IGV (Robinson et al 2011).

1.5.3 Sample correlation and concatenation

DeepTools plot coverage was run on bam files from all 6 time points, for each of the 3 libraries. DeepTools was run with default parameters on the Galaxy servers. Quality score box plots were generated for all original fastq files using the FASTX-toolkit. Again, FASTX-toolkit commands were run on a Galaxy server with default parameters. Pearson's Correlation test was performed on all libraries for a given time point. Libraries 1 and 2 were resequenced and the replicates were pooled together by concatenation to achieve the requisite sequencing depth for ATAC-seq.

Irreducible Discovery Rate plots were generated with the batch consistency and batch consistency plot commands from the IDR R-package code. For a given timepoint, Two peak files were compared using the batch consistency command with broadPeak set to False and p-value specified. All possible permutations for a given timepoint were plotted together using the batch consistency plot command. PS plots were converted to PDF using PS2PDF command line arguments.

1.5.4 Peak calling

MACS2 narrowPeaks were generated with ZT2 files set as experimental group and ZT14 files set as control. Input files were bam files. Q values were used with a cutoff of 0.01. For increasing the window size, '–nomodel –shift -100 –extsize 200 parameters were added.

1.5.5 Coverage plots

Using bedtools intersect, reported every CHIP-seq peak from above that corresponded with a CLK direct target. Ranked these peaks by highest MATSCORE as in clk-direct-target csv file. Made three separate bed files, with top 200, top 100, and top 50 peaks according to MATSCORE. The peak center was extended by -500 from the start and +501 from the end to give a total distance of 1001, or 1000 total bp span. Bedfiles included arbitrary strandedness. Uploaded these bed files to separate directories (ie 2/50 for ZT2 top 50, 2/100 for ZT2 top 100 etc.)

A custom script was used to generate read counts for individual base pairs. Custom python scripts to manipulated resulting file. Scripts combined all three files into one data frame, and either summed across all genes for a give bp and then averaged result, or averaged across bp and then averaged replicates. A csv output file was returned.

CSVs across all time points were manually concatenated into excel and made into one csv for R. In R,

signal was plotted against bp distance, -500 to 500 using the basic R plot function. To combat noise, a local polynomial regression was performed. This resulted in a smoothing as it fitted a curve such that a given point was weighted toward the nearest data to x. Smoothness of this curve was controlled by span parameter, α. α was set to a relatively modest 0.2 (ranging from 0 to 1) which gave the curve smoothness without over-fitting. Error could be visualized by using a polygon shading stretching from standard deviation resulting from averaging the three replicates together (calculated in pre-processing of data before import into R).

1.5.6 Differential analysis

Raw reads for ZT2 and ZT14 samples were compared using the Bioconductor R package, edgeR. Three controls, ZT2 timepoints, and three experimental, ZT14 time points, were used. Exact test and general linear model outputs were compared and analyzed.

Chapter 2

Investigating post-transcriptional regulation in *Drosophila* brains

Abstract

After mRNAs are made, they can be post-transcriptionally modified by the enzyme ADAR in a process known as RNA editing. The most common type of RNA editing is adenosine to inosine. This is facilitated by a class of proteins called adenosine deaminases (ADAR) that act on double stranded RNA and catalyze the conversion from A-I. To identify RNA editing sites that change with age in Drosophila, we analyzed RNA-seq libraries. We identified 1445 sites whose editing increases with age and 1015 that decrease. Interestingly, those transcripts that showed increased editing with age, were also identified to be transcripts encoding circular RNAs that increase with age. Several previous studies suggest that ADAR may modulate circle formation by editing the intron regions that bracket circular RNA. The mechanism by which ADAR editing is correlated with circular RNA formation, however, is not clear.To investigate a possible relationship between RNA editing and circle formation we examined whether the RNA editing sites altered splice sites or altered complementarily in the regions surround circular RNAs. We found no significant effect of RNA editing on splice sites but showed that editing may modulate the complementarity of the introns flanking circular RNAs. Those regions flanking circRNAs that are more highly edited with age had a statistically significant ($p=0.03$) higher level of complementarity than pairs without editing events or pairs with editing events that decreased with age. This may be evidence that sequence features that promote editing are also sufficient at promoting circular RNA formation. Though these results are preliminary, they may highlight that RNA

editing and circular RNAs could have a specific function in *Drosophila* brains.

2.1 Introduction

RNA editing is phenomenon seen in eukaryotes, in which an RNA transcript is biochemically altered so its final sequence differs from would be predicted, given genomic DNA. The most common type of RNA editing is adenosine to inosine. This is facilitated by a class of proteins called adenosine deaminases (ADAR) that act on double stranded RNA and catalyze the conversion from A-I. ADAR proteins have a double stranded RNA binding domains and a catalytic domain. In the enzyme active site, a zinc ion coordinates a nucleophilic hydrolytic deamination (Savva et al 2012). The inosine is then read as guanosine by the translation machinery (Stapleton et al 2006). ADAR proteins are highly conserved across species. Mammals encode three ADAR proteins (ADAR1, ADAR2, and ADAR3), while *Drosophila* encodes one, simply ADAR (Savva et al 2012).

While ADAR editing takes place in primarily non-coding regions, it can have specific biological functions. (Morse et al 2002)(Stapleton et al 2006). Human ADAR has been found to act in embedded ALU repeat regions and may act to stabilize or correct mismatched base pairs in the region. (Levanon et al 2014). Other ADAR targets have been found in the 5' and 3' UTR of RNA transcripts, which may affect the stabilization, location or translation of mRNAs (Morse et al 2002). Editing also may play a role in adding splice sites which can affect the stability of the subsequent molecules (Rueter et al 1999). In Drosophila, editing also takes place in exons which can further argue for a role of RNA editing in gene regulation.

ADAR proteins are found disproportionately in the nervous system in both *Drosophila* and in mammals, which leads to a variety of important neurological functions with potential impacts on behavior(Savva et al 2012). In *Drosophila* ADAR targets are enriched in ion channels and protein receptors. Recent work has found that this can alter the structure of fundamental neuron structures, such as synapses (Nainar et al 2016). Splice-site choice and the role ADAR plays in *Drosophila* miRNAs further provides a case for the importance of A-I editing in neuronal gene regulation (Lai 2005)(Behm and Ohman 2016). Some links between ADAR editing and behavior have also been found. ADAR-deficient animals may have increased sleep pressure due to synaptic dysfunction in glutamatergic neurons (Robinson et al 2016). *Period* loss of function alleles are also correlated with an alteration in ADAR editing patterns, while mice deficient in ADAR were found to accumulate abnormal levels of CRY2 (Terajima et al 2016)(Hughes et al 2012). These findings indicate a surprising and complex connection between ADAR and the post-transcriptional regulation

of circadian rhythms.

Interestingly, ADAR has an age dependent role in several organisms (Li et al 2014)(Larsen et al 2016). While this phenomenon has not been well studied in *Drosophila*, preliminary data suggests that editing does occur in an age-specific manner in *Drosophila* brains. Several investigations present a strong link between the presence of circular RNA and editing. A-I editing is highly enriched in the flanking intron regions that bracket circular RNA, suggesting that ADAR could be correlated with circular RNA formation (Ivanov et al 2015). ADAR knockdown experiments have also shown an up regulation of circular RNA formation as ADAR levels decrease (Rybak-Wolf et al 2015). While these are contradicting hypothesis, it is primarily thought that editing acts in opposition to circular RNA formation. As with ADAR editing, circular RNAs are highly enriched in nervous tissue (Westholm et al 2014)(Rybak-Wolf et al 2015). The mechanism by which ADAR is correlated circular RNA formation, however, is not clear.

Circular RNAs are a newly discovered category of RNAs, with properties that are largely unidentified. Unlike linear RNAs, circular RNAs are a covalently closed loop. Circular RNAs were previously thought to be the consequence of incorrect splicing as they were infrequently observed. New high-throughput sequencing methods, particularly next generation RNA-seq, detects circular RNAs in much higher levels that formerly estimated. Around 10% of expressed genes can form circular RNA transcript (Chen 2016). While some circular RNAs may be products of protein coding genes, the circular structure of the RNAs is not conducive to conventional translation machinery. Therefore, most circular RNAs are likely not protein-coding. The ubiquity of circular RNAs, then, is perplexing, posing provocative questions about the formation and function in cells.

Circular RNAs are formed by back splicing to produce a transcript that folds onto itself. Back splicing is a form of non-canonical splicing where the splicing occurs in a reverse order. Although the specifics are not well understood, cis and trans regulatory proteins facilitate back splicing by binding to the precursor RNA and guides a downstream 5' sequence to pair with an upstream 3' acceptor site (Chen 2016). These regions, referred to as flanking regions, border the circle region and are usually intronic. Flanking regions are often significantly longer than typical introns. The flanking regions have been demonstrated to be enriched for repetitive elements at rates nearly two-fold of non-circular transcripts (Lasda and Parker). Repetitive elements, such as ALU repeats in humans, can enable back splicing by increasing the likelihood that the 5' and 3' regions will pair. Recent work suggests that in part, these repetitive elements could be reverse complementary matches, which further promoting circle formation.

Circular RNAs have a number of features that suggest a role in the cell. For instance, circular RNAs

are highly stable. Compared with linear transcripts, which have been observed to have a cellular half-life of about 20 hours, circular RNAs have a half-life of more than 48 hours (Jeck et al 2013). Circular RNAs are notably resistant to many forms of enzymes that degrade linear transcripts, such as RNAse R, which can partially explain this phenomenon. They are also largely cytoplasmic, transported out of the nucleus in a similar fashion to mature mRNA. This is surprising because only a few classes of non-coding RNAs reach the cytoplasm (Chen and Carmichael 2009). The location and the long half of circular RNAs suggests some enduring function for the transcripts, albeit one that is not well understood. Studies in mammals have also demonstrated that circular regions are conserved between paralogs and orthologs of a gene family (Jeck and Sharpless 2014). Conservation of circular RNAs is a remarkable finding that points to some cellular purpose beyond the happenstance, as was previously thought.

Despite this precursory knowledge of circular RNAs, there is not an agreed upon function. Alternative splicing in general is thought to be a gene regulatory mechanism, as it can generate protein diversity. In this hypothesis, the formation of circular RNAs is intimately tied with function. There is also some evidence that circular RNAs act as microRNA sponges. In some cases, circular RNAs have microRNA target sites that can bind with mRNAs, and thus negatively deregulate protein expression. While this hypothesis has been experimentally demonstrated for some specific circular RNAs and bioinformatics analyses suggest that exons in circular regions may be conserved for microRNA binding, it is unclear whether circular RNAs exist to the high levels required for noticeable RNA regulation. Other studies provide evidence that because of the presence of circular RNAs in the cytoplasm, some particular classes of circular RNAs may bind to cytoplasmic proteins and prevent their entry to the nucleus. This occurrence, as with most of the models of circular RNA function, has not been widely researched, which undercuts any definitive conclusions that can be made.

The connection between ADAR editing and circular RNA formation is compelling. In particular, age dependent editing and circular RNA accumulation presents an intriguing possibility. Recent work suggests that aged heads (¿20 days) produced the highest number of circular RNAs out of any tissue (Figure 14A).]These results were consistent across a number of RNA transcripts studied (Figure 14D-E). This could possibly be related to circular RNA stability, as circular RNA is more stable than linear transcripts, they could accumulate over time. In this study, age dependent editing in *Drosophila* was examined. To identify RNA editing sites that change with age in Drosophila, we analyzed RNA-seq libraries. We identified 1445 sites whose editing increases with age and 1015 that decrease. Interestingly, those transcripts that showed increased editing with age, were also identified to be transcripts encoding circular RNAs that increase with age. Several pre-

vious studies suggest that ADAR may modulate circle formation by editing the intron regions that bracket circular RNA The mechanism by which ADAR editing is correlated with circular RNA formation, however, is not clear.To investigate a possible relationship between RNA editing and circle formation we examined whether the RNA editing sites modulated splice sites or altered complementarily in the regions surround circular RNAs. We found no significant effect of RNA editing on splice sites but showed that editing may modulate the complementarity of the introns flanking circular RNAs. Those regions flanking circRNAs that are more highly edited with age had a statistically significant (p=0.03) higher level of complementarity than pairs without or pairs with editing events that decreased with age. This may be evidence that sequence features that promote editing are also sufficient at promoting circular RNA formation. Though these results are preliminary, they may highlight that RNA editing and circular RNAs could have a specific function in *Drosophila* brains.

2.2 Results

2.2.1 Datasets

RNA-seq data in this investigation was generated using old (35-40 days old) and young (3-5 days old) *Drosophila* brains. Data was mapped to the dm6 genome using the spliced aligner STAR (Dobin et al 2013). Editing sites were identified and data was sorted into three categories: editing sites that increase with age, editing sites that decrease with age, and editing sites that do not change with age. Editing sites were required to have at least 5% editing in one sample and greater than ten reads in both samples. This identified approximately 8,000 editing sites in total, where 1445 sites increase with age and 1015 decrease with age. 500 transcripts in *Drosophila* had editing events that increase with age, and 504 contained editing events that decreased with age.

These transcripts show incredibly high literature enrichment for circular RNAs ($p=10.0^{-100}$) (Personal communication, K. Abruzzi) (Ashwal-Fluss et al 2014). Of these, 212 editing events of total 1445 editing events that increase with age were found in Ashwal-Fluss et al, examining the competition of circular RNA and pre-mRNA splicing.

Since both circles and ADAR editing are enriched in neuronal tissue, and both are simultaneously increasing with age, could age dependent editing be associated with circle formation? To answer this question, circular RNA data was obtained from Westholm et al and converted to dm6 coordinates (Table 6). Coordinates of known circular RNAs were converted to a BED file and used throughout this investigation. From

this dataset, 516 of editing sites that increase with age, or 35%, are also in transcripts that encode circular RNA that increase with age. Only 203 editing sites that decrease with age are in circular RNAs transcripts that increase with age, just 19%. This is comparable with only 16% of sites that do not change with age, or 915 editing sites of 5557 total sites, that are in circles that increase with age. The proportion of editing sites in circles is statistically significant different between groups ($p<0.0001$).

This correlation lead us to question what mechanism is leading to this correlation. In particular, could editing be affecting splicing, or changing RNA structure in order to promote circular RNA formation?

2.2.2 Alternative splicing variants in age dependent circular RNA

Since circular RNA formation is particularly linked to alternative splicing, we investigated whether editing was affecting splicing in an age dependent manner. Splice site variants are defined as a transcript that is produced because a mutation occurs at the site, thus altering splicing and the production of a mature RNA. If editing that increases or decreases with age is affecting splicing, this could be a possible mechanism to explain the correlation of circular RNAs with editing events that increase with age.

The ENSEMBL variant effect predictor (VEP) tool was used to predict the location and consequence of editing events (McLaren et al 2016). Editing datasets that increase, decrease, and do not change with age were annotated with the nucleotide changes seen and run on the VEP webserver. Data was run with default parameters that did not filter and returned all consequences for all variants.

Only a very small number of editing sites were found in splice sites, less than 1% of all variants for each dataset (Table 7) (Figure 16). While this is seemingly insignificant, the majority of splice site variants were found to be age-dependent. Of these age-dependent splice variants, 67% of those that increase with age (22 editing sites) were located at splice sites that flank known circular RNAs. 82% of those editing sites that decrease with age (34 sites) were located at splice sites that flank known circular RNAs (Table 8). However, these 34 sites were only contained in 5 unique transcripts, making this finding less significant.

This result is more compelling in light of gene ontology (GO) analysis. Using DAVID, splice-site variants were queried. While the only significant ($p<0.01$) GO cluster for all variants and variants produced by editing that increases with age was "localization", variants produced by editing that decreases with age had significant GO clusters enriched for synaptic plasticity (Table 9). All five of the variants associated with circular RNA function were included in this cluster. Unlike localization, synaptic plasticity, or the ability of synapses to strengthen or weaken over time, is a very specific biological function important to brain development. This could suggest that editing is acting in combination with circular RNA formation

to regulate processes in young brains.

2.2.3 Investigating intronic regions flanking circular RNA

The effects of editing on sequences flanking circular RNA regions was also examined. Since these regions are often intronic regions that are enriched for repetitive elements, they are excellent candidates for ADAR editing (Ivanov et al 2014). In particular, we investigated whether age dependent editing can affect the introns flanking circular RNAs in a way to promote circular RNA formation.

Intronic regions that flank circular RNAs were found. Two regions were generated for each circle, on the 5' and 3' ends of the region. While there is some evidence in mammals that flanking regions are long and repetitive, recent research in Drosophila suggests that introns may be shorter than previously thought (Westholm et al 2014). This motivated restricting intronic regions to 300bp. We also anticipated complementarity and folding analysis, where accuracy is often limited by length of nucleotide sequence.

These flanking intronic regions were then analyzed for whether they contained editing sites. Three datasets were generated: flanking intronic regions with any editing (1339 pairs), flanking intronic regions with editing sites that increase with age (345 pairs), and flanking intronic regions with editing that decreases with age (208 pairs). This compares with 36,875 flanking pairs around all circular RNA from our dataset in total, meaning that editing occurs in approximately 3% of all circular RNA flanking regions. While this is not a substantial portion, out of circular RNA transcripts that increase with age, there are 110 flanking pairs, or 21% that contain editing events, out of 524 total flanking pairs. The percentage of circular RNA transcripts that increase with age that contain editing more than doubles to 252 flanking pairs out of 524 total if the flanking region is extended to 500bp. This less conservative analysis suggests that upwards of 50% of age-dependent circular RNAs could be associated with editing events.

Initial complementarity was performed using pairwise sequence alignment, where the 5' and 3'intronic region around the circular RNA was analyzed for complementarity. We chose the BLAST algorithm to implement our pairwise sequence alignment (Camacho et al 2009). BLAST is a heuristic algorithm that subsections input sequences into smaller words. These query words are compared to the target sequences and matches are identified. If BLAST detects a match, the alignment is extended in both directions in order to and a score that is higher than some minimum threshold (Altschul et al 1990). Scoring is calculated by awarding points for matches, and penalizing for mismatches or gaps. In this way, two sequences with high complementarity would be considered a high scoring pair, and could be directly compared to sequences of the same length that are lower scoring.

BLAST is less accurate than a dynamical programming approach to pairwise sequence alignment, such as the EMBOSS Water implementation of the Smith-Waterman algorithm, in determining precise alignments (Rice et al 2000). Importantly, however, BLAST is a much faster and less memory intensive algorithm. This has the advantage of being able to test many sequences very quickly, an advantage in this investigation with large numbers of flanking pairs. Holistic indications of complementarity were also acceptable for comparing complementarity between edited and non-edited flanking pair datasets.BLAST was performed by hand on twenty randomly selected pairs for both flanking pairs with editing that increases with age and for flanking pairs without editing. Max bit scores (a normalized score that estimates the magnitude of search space) were recorded for all queries.

Preliminary analyses that suggested a significant difference between complementarity in a small sample, motivated performing pairwise complementarity analysis on all flanking pairs that contain editing events that increase or decrease with age. A custom script was written that mined the UCSC database and retrieved fasta files for each pair of flanking introns around circles. The fasta files were then used as the input of a BLAST query. Any intron with multiple editing events was only tested once. The max bit score was return for all pairs. This analysis was done on flanking pairs around circles with editing events that increase with age, editing events that decrease with age, and those with no editing.Flanking pairs around circles with editing events that increase with age was had a higher max score than non-edited pairs (Table 10). This result was modestly statistically significant (p=0.02). There was not statistically significant difference between pairs with editing events that decrease with age compared with non-edited pairs (p¿0.05). Again, it is worth noting that both groups with editing events had a higher level of variability than data without editing.

Next, flanking pairs with editing sites that increase with age around circular RNA transcripts that also increase with age were analyzed for whether editing events were stability promoting. Stability was assessed using the UNAFold algorithm for RNA secondary structure. This algorithm encompasses three popular methods for secondary structure prediction: mFold, Vienna RNAfold and sFold (Zucker 2003) (Gruber et al) (Ding et al 2004). It computes both minimum and suboptimal foldings, as well as partion functions that allow exact base pair calculations in combination with stochastic sampling (Markham and Zucker 2008).

In our investigation, minimum free energy (MFE), the point at which the RNA structure was most stable, was calculated. This is a thermodynamic quantity in which the more negative the value, the more stable the structure is. MFE was first calculated for sequence files of flanking pairs (of 300bp each) not containing an editing event, and then the sequence of the same flanking pairs with the editing encompassed. A small

sample of 10 flanking pairs was examined. Of these, 7 structures were shown to increase stability with editing. Two did not change with editing, while only one structure became destabilized with editing (Table 11).

2.2.4 Sequence composition and motif prediction

The differences between complementarity in intronic pairs that have editing events that increase with age and without editing could suggest some fundamental difference in regions that border circles with editing events. This was explored further using motif analysis.

Simple tallying of the nucleotide composition in intronic regions that border circles without editing and with editing that increases with age was identical (Table 16). While this indicates that the actual diversity of nucleotides is not different between groups, MEME motif analysis suggests that the arrangement of these nucleotides is different. MEME motif analysis acts by running two complementarity motif discovery analysis on input files, to produce output on possible motif enrichment (Machanick and Bailey 2011).

Intronic pairs around circles without editing were enriched for motifs, but these motifs were not present in more than 27% of the pairs. Intronic pairs with editing, however, were enriched for a particular motif, GCAGCAGC, in 46%, or 49 of 105 total pairs (Figure 18). This motif did not contain editing sites, however, it was consistently located within 40bp of an editing event./uences are not well understood, this is an interesting trend that could be studied further.

2.3 Conclusion

While much is still not known about the formation of circular RNAs, this study explores several ways that age-dependent ADAR editing may help to facilitate circular RNA transcripts that increase with age. Variant effect prediction results were largely inconclusive in demonstrating a connection between editing and splice site variants. While age dependent editing does make up a majority of splice-site variants, and that these editing events are occurring in circular RNA splice sites, the number of these occurrences is extremely small compared to editing events as a whole. The small size of this phenomenon indicates that while it could be feasible ADAR editing is encouraging some circular formation, it is likely not a significant mechanism. This is consistent with an explanation of circle RNAs as a by-product of regulation, and not necessarily an important biological process in their own right.

Complementarity analysis, however, contradicts this conclusion. Flanking regions around circular RNAs

were studied, as these regions have been demonstrated to be key in facilitating circular RNA formation. Our results show that while these flanking regions did not have strong degrees of complementarity, as previously suggested by Westholm et al, there was a variation in complementarity between flanking pairs that contained editing and flanking pairs without editing. Flanking pairs with editing that increased with age had a statistically significant (p=0.03) higher level of complementarity than pairs without or pairs with editing events that decreased with age. This may be evidence that sequence features that promote editing are also sufficient at promoting circular RNA formation. Motif analysis, which finds flanking pairs with editing that increases with age enriched for a motif near editing sites can also bolster this claim. Furthermore, around 50% of the flanking pairs around circular RNAs that increase with age have editing events. This is a strong indication that there exist distinct commonalities between editing that increases with age and circular RNA that increases with age in *Drosophila* brains.

A possible mechanism that may explain these commonalities is that ADAR editing may act to stabilize circular RNA transcripts. While this is expected given that ADAR must act on double stranded RNA which is inherently more stable than single stranded RNA, the connection with age-dependent circular RNAs is novel. As editing increases with age, it could be acting in the flanking regions of circular RNAs, to stabilize the structure so that instead of decaying, like linear RNA, more can accumulate in the brain with time. Other models may be that in general, post-transcriptional processing becomes disregulated with time, and that both ADAR editing and back splicing levels increase. ADAR editing has been repeatedly demonstrated to be important to regulating gene products and ultimately, may play a role in animal behavior. Though much is left to be said about the function of circular RNAs, this conclusion may highlight that circular RNAs could have a specific function in *Drosophila* brains.

2.4 Methods

2.4.1 RNA-seq data generation

RNA-seq data was generated using CS WT *Drosophila* brains. Young flies were 3-5 days old and aged flies were greater than 35 days old. Libraries were sequenced on Next-seq 75bp reads (Illumina).

2.4.2 Data processing

RNA-seq data was aligned to the dm6 transcriptome using STAR (Dobin et al 2013). Only unique reads were considered and PCR duplicates were removed. Raw read count and read per million counts were

generated for each gene.

2.4.3 Editing analysis

Editing analysis was performed to detect A-I editing events. These editing sites were then analyzed on how they changed with age. Editing sites were required to have more than ten reads. Sites meeting this criteria were then filtered so that editing events were only an A base call on the top or bottom strand, such that our pipeline only considered ADAR editing. Only sites with no genomic variation were considered to increase confidence in the accuracy of our prediction. Editing was required to be at least 5% in either editing sites that increase or decrease with age datasets. The ratio of editing in young to old brains was calculated. Proportions larger than 1.5 were considered to be editing that decreased with age, and proportions less than 1.5 were considered to be editing events that increased with age.

2.4.4 Variant effect prediction

Input datasets for the ENSEMBL variant effect predictor software were generated using custom perl scripts that transformed counts of base pair per editing site into VCF format. Strandedness was forced on samples using strand information from the UCSC database. Three datasets were generated- all editing sites, editing sites that increase with age, and editing sites that decrease with age.

Input files were then run on the http://www.ensembl.org/Tools/VEP server using default parameters, the dm6 build of the *Drosophila* genome, and ENSEMBL transcript database. Output was analyzed separately using R.

2.4.5 Generation of intronic flanking regions around circular RNAs

Introns flanking known circular RNA were found using UCSC table browser and bedtools. dm6 introns were generated using the table browser and the following parameters Assembly- dm6 Group- genes and gene predictions Track- refSeq Table- refGene Output- bed Create one bed ouput per- introns

Flanking regions were generated from known circular RNA bed file (Westholm et al 2014) and known introns were compared to flanking regions and reported using bedtools left outer join. Verified that these regions were intronic using bedtools intersect and bedfile of dm6 introns from UCSC genome browser

2.4.6 Complementarity algorithm

Fasta files were obtained from UCSC genome browser of flanking regions around circRNA. BLASTN algorithm (with word size of 7) was used to quantitatively measure complementarity of flanking intron pairs (Altschul et al 1990). Flanking intron pairs were divided into two categories: those with editing that changed with age (either decrease or increased) and those without editing. The UCSC genome database was massively queired for dm6 fasta files of flanking pairs of interest. The fasta files were used as the input for the BLAST API and the blast scores were recorded for all flanking pairs. Hits from BLAST were sorted by max score (bits) and most significant (smallest e-value) – used these criteria to return top hit. Calculated mean, median, and SD for the max score.

2.4.7 Stability prediction

Fasta files for flanking pairs with editing events that increase with age were found. These fasta files were folded using UNAfold 3.8. The minimum free energy structure was recorded. Editing events were manually implemented in the fasta files, and the structures were re-recorded.

Appendices

Appendix A

Figures

Figure **??**

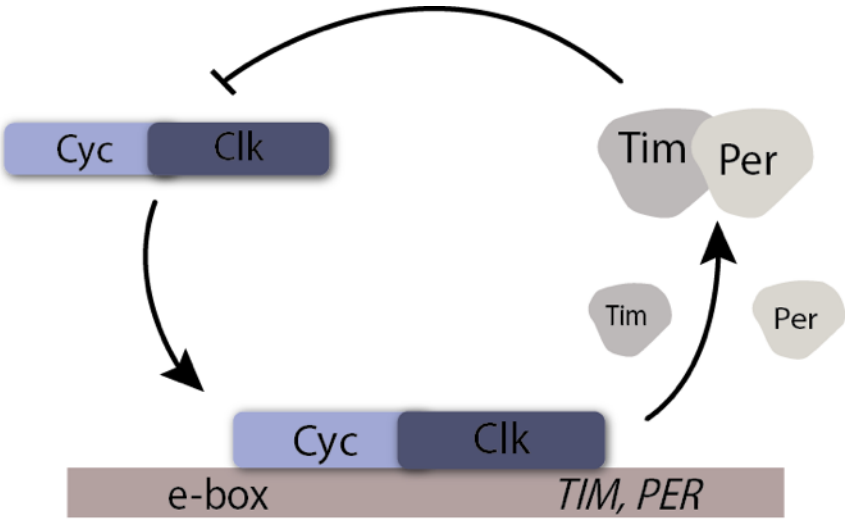

Figure 1: CYC/CLK form a heterodimeric transcription factor. The CYC/CLK protein complex then binds to the ebox region of *tim* or *per*, activating transcription. TIM and PER proteins are translated and exit the nucleus. They accumulate in the cytoplasm and form a dimer. The TIM/PER dimer then re-enters the nucleus, cause the realse of CYC/CLK, repressing transcription.

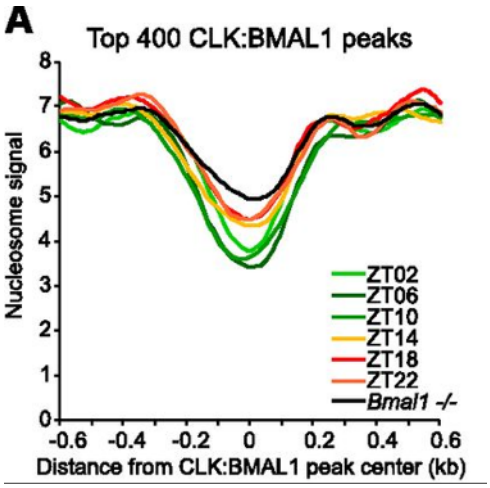

Figure 2: Nucleosome signal around the CLK/BMAL1 binding location in for 400 CLK controlled genes in mammals. Magnitude of nucleosome signal correlates with strength of CLK binding, unlike what is seen in ATAC-seq plots (figure from Menet et al 2014).

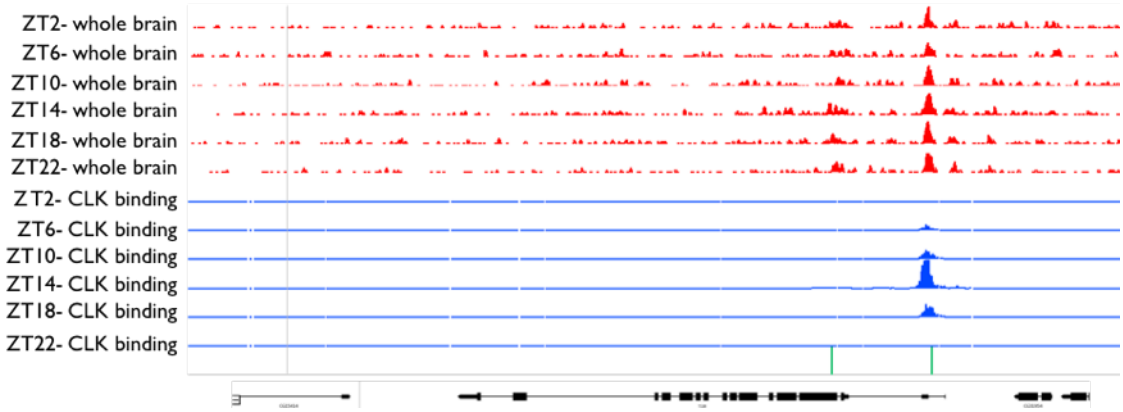

Figure 3: (in blue) CLK CHIP-seq reveals that CLK binding levels are oscillatory. CLK binds at the ebox of the tim gene, indicated by the gene lines. Transcription occurs from left to right across the gene. CLK binding lLevels are the highest during the night, peaking at ZT14, and lowest during the early morning, ZT2. In contrast ATAC-seq reveals a constant signal on the ebox of tim (in red). This aligns with CLK binding patterns, but does not oscillate, as expected.

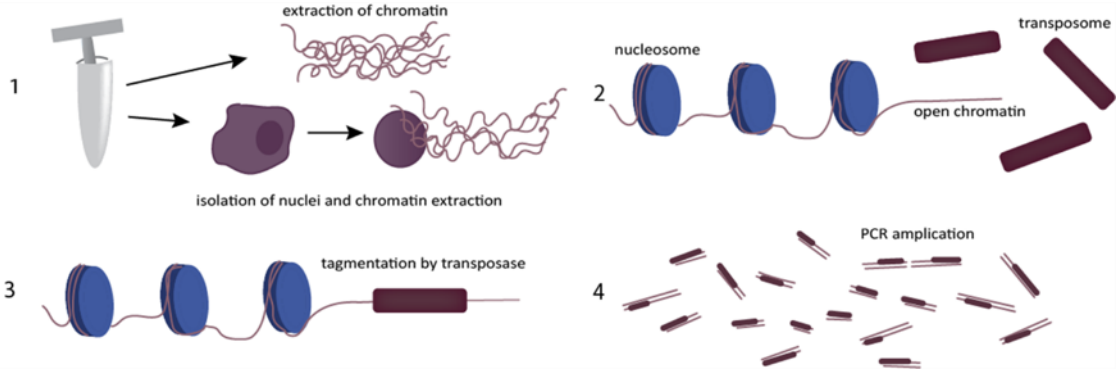

Figure 4: (1) Chromatin was extracted from Drosophila brains. (2) Chromatin was incubated with transposome and (3) tagmented in areas where chromatin is open. (4) Tagmented chromatin underwent several rounds of PCR amplification and was sequenced (Buenrostro et al 2013).

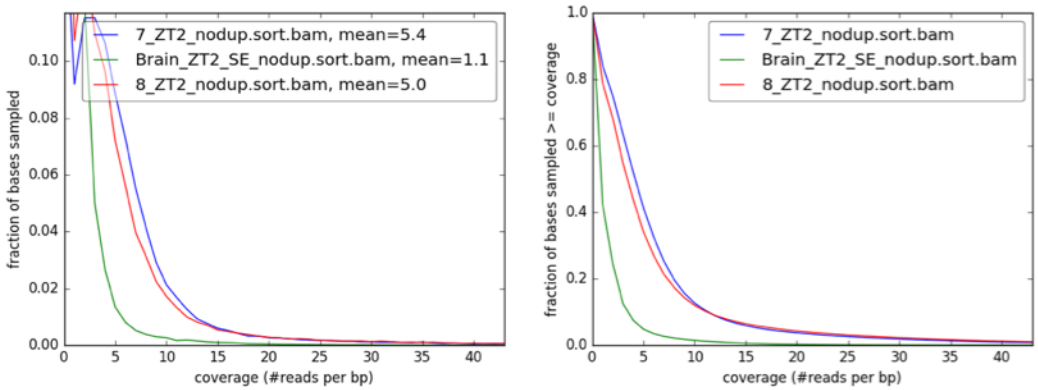

Figure 5: All three data sets had approximately 5-10 reads per base pair. Libraries 1 and 2 have a mean of 5.4 and 5.0 reads per base pair. This is within the acceptable range for high-throughput sequencing data. Library 3 had poor coverage, with a mean of 1.1 reads per base pair, which could be explained by this library being the only non-pooled dataset in this study. This supports our conclusion that single-ended sequencing provides acceptable coverage for the purposes of this study.

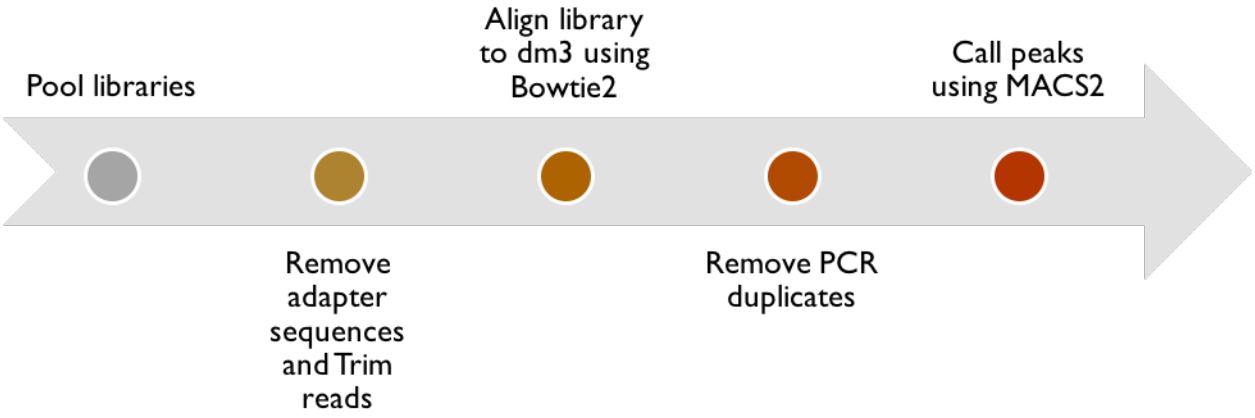

Figure 6: Libraries are pooled together to achieve significant coverage, adapter sequences are removed, the libraries are aligned using Bowtie2, and PCR duplicates are removed. Finally, peaks are called using MACS2.

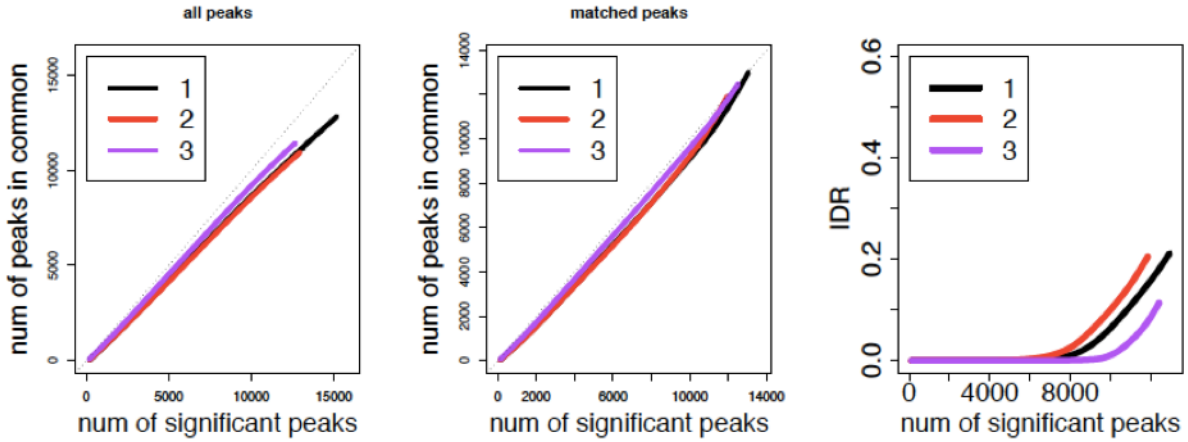

Figure 7: (a) number of peaks in common versus the number of significant peaks from all peaks from all replicates. There is a high degree of matched peaks, suggesting all libraries have concordance in the number of peaks that produce significant reads. (b) number of peaks in common versus number of matched peaks, or peaks present in all samples. Nearly all peaks shared by all three datasets are significant suggesting a high level of reproducibility between the sets. (c) number of significant peaks versus the irreducible discovery rate for all three datasets. The reproducibility of the datasets begins to decay at 8000 significant peaks, indicating that the majority of peaks in the samples are reproducible.

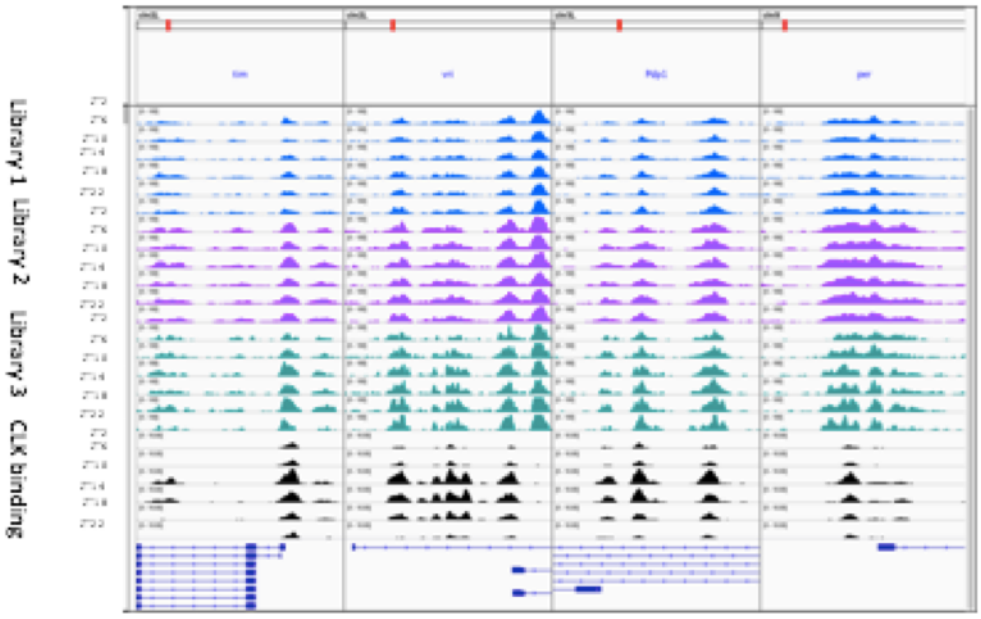

Figure 8: In each of the three datasets used in this investigation, chromatin accessibility did not appear visually different between time points for core circadian genes. This contrasts with CLK binding patterns, which have a clear visual difference between all time points on all genes studied.

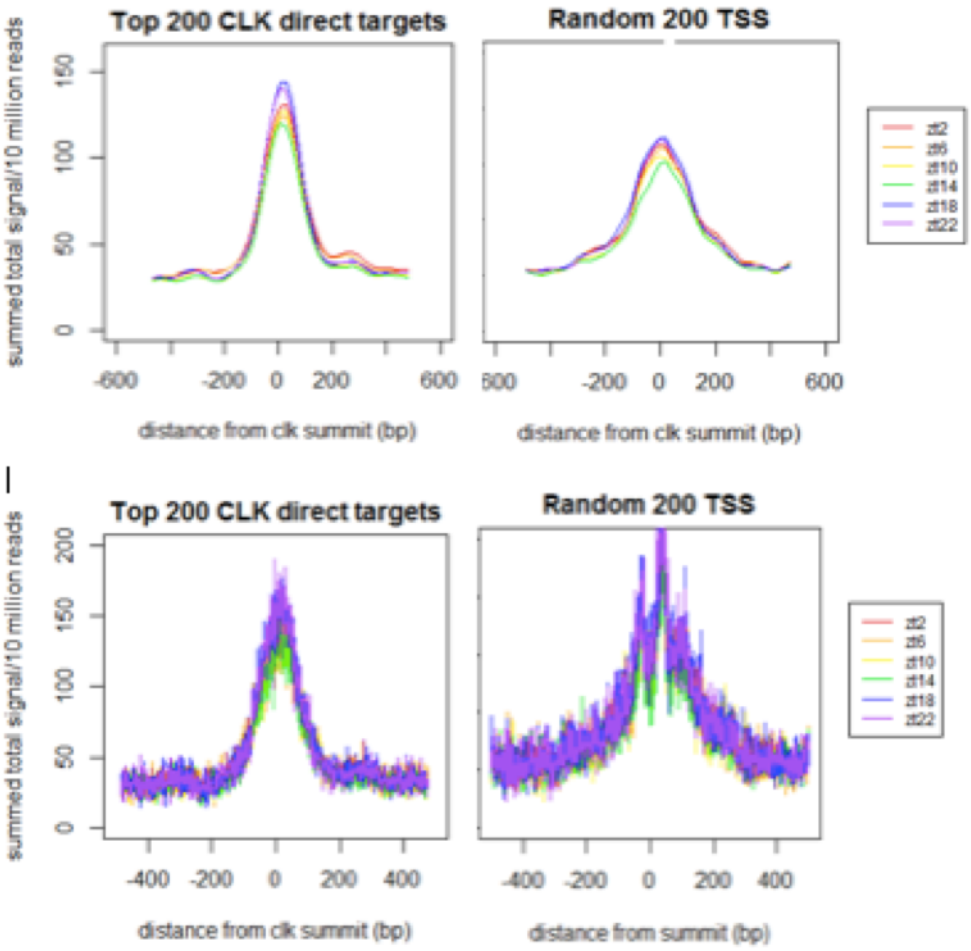

Figure 9: Average Summed signal (10 million reads) versus distance from the CLK binding summit for (a) 200 CLK direct targets and (b) 200 random transcription start sites. ZT14 appears to have slightly less signal, however, there is a high degree of error (c and d)

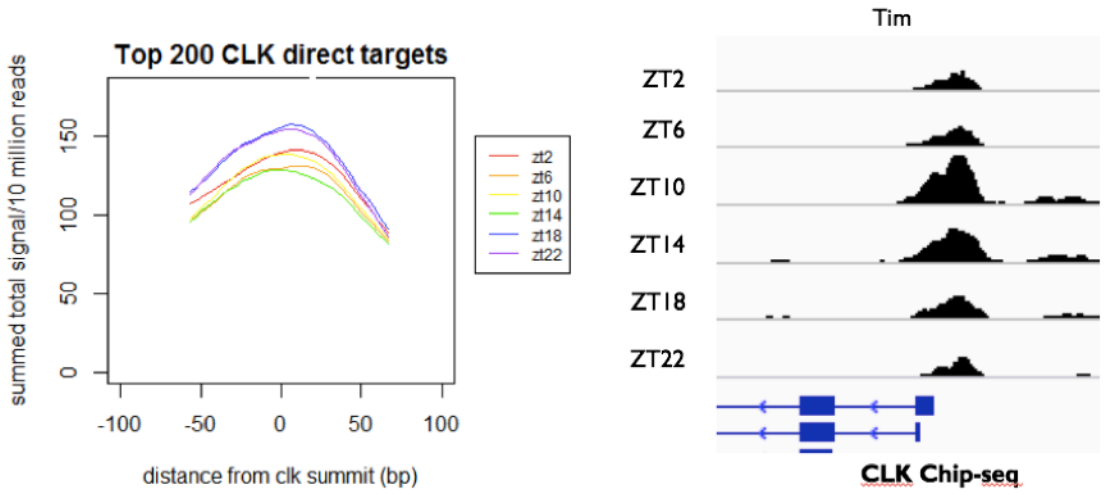

Figure 10: (a) there is a slight difference between summed total signal of the 6 time points for 200 CLK direct targets, but it does not correlate with (b) CLK binding patterns

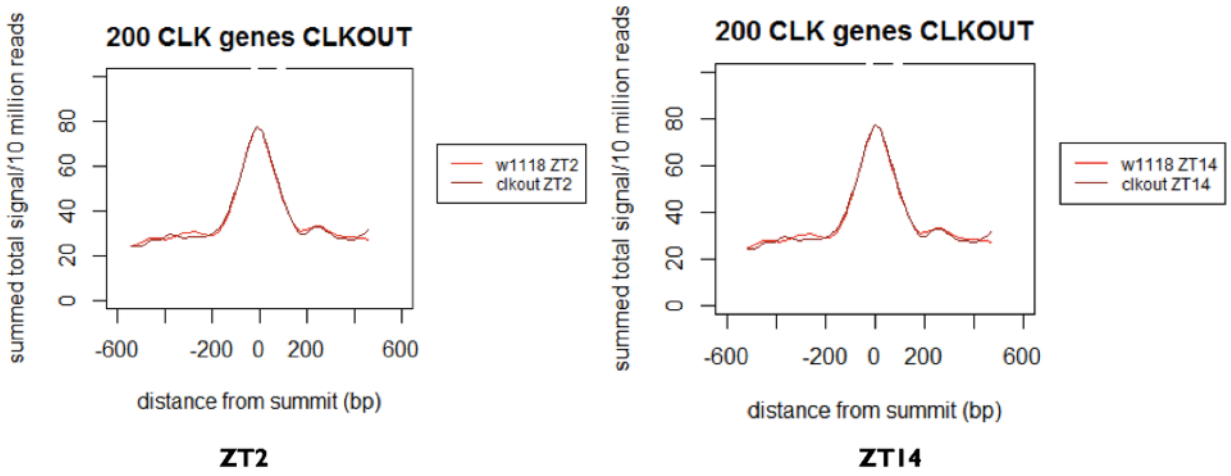

Figure 11: (a) ZT2 and (b) ZT14 time points are nearly identical in a CLK deletion. This could suggest that chromatin accessibility is not dependent on CLK binding

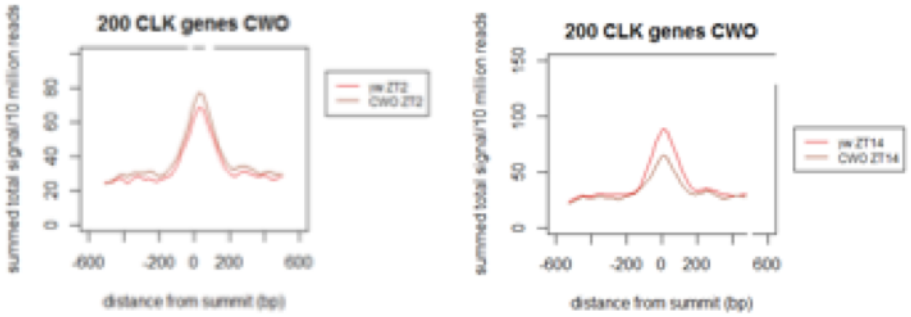

Figure 12: Preliminary data shows that in a CWO mutant, chromatin is (a) more open than a wild type at ZT2 but (b) substantially less open than wild type at ZT14. This could suggest a complex relationship between CLK and CWO as transcription factors working in opposition to open CCGs

Figure 13: Transcription factor foot printing analysis suggests that Trl may be binding to CCGs. Preliminary analysis of a Trl knockdown finds that (a) chromatin accessibility may not be changing on CCGs in particular, but that (b) Trl deficiency increases chromatin accessibility throughout the *Drosophila* genome.

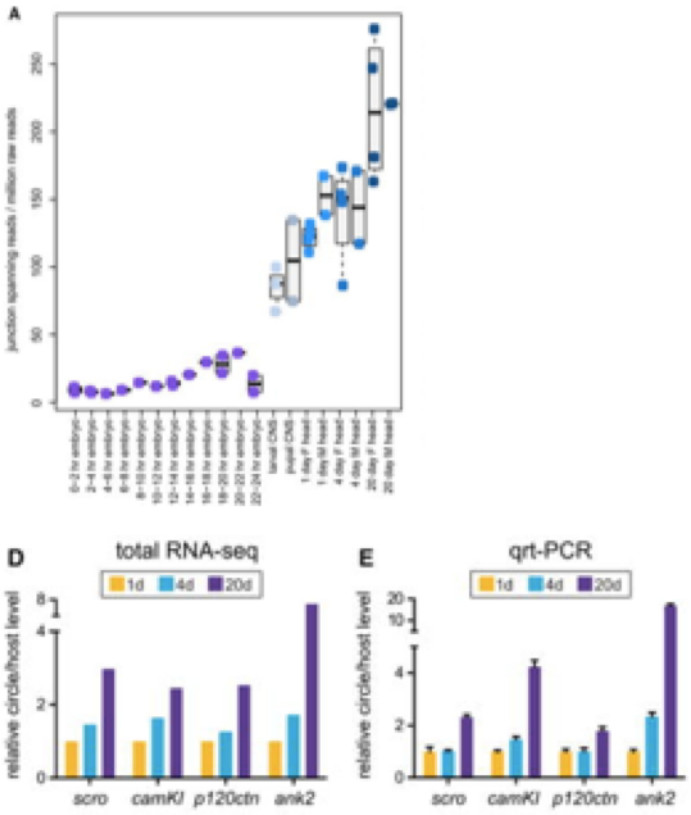

Figure 14: Circular RNA levels have been demonstrated experimentally to (a) increase substantially overtime in *Drosophila* brains by examining the number of junction spanning reads used in the detection of back splicing events that indicate circular RNA formation. This has been verified using total RNA-seq and qt-PCR for four genes of interest over 1 day, 4 day, and 20 day intervals (figure from Westholm et al 2014).

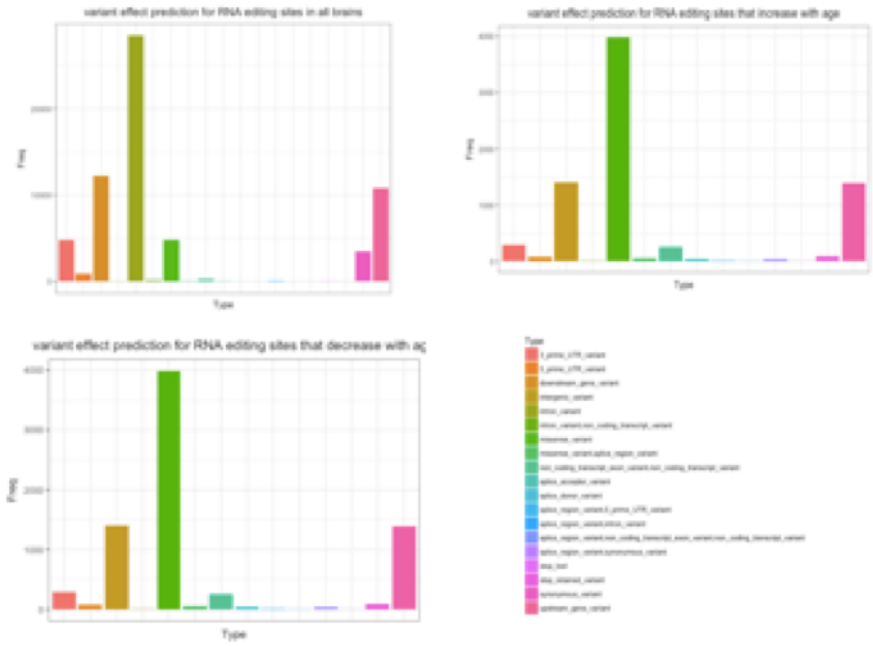

Figure 15: Variant effect prediction in editing sites that (a) do not change with age (b) increase with age and (c) decrease with age finds that the primary variant effect is intronic. In editing sites that increase (b) and decrease (c) with age there are high levels of missense variants. A small minority of variants are splice-site variants

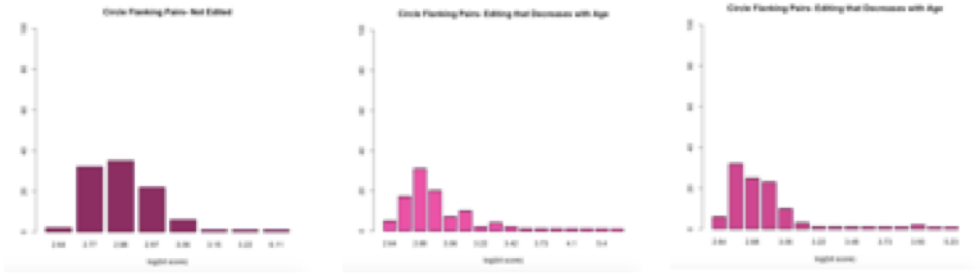

Figure 16: Complementarity, determined using max bit score obtained from pairwise sequence alignment shows that flanking pairs around circular RNA with editing that increases with age (c) is significantly higher than pairs with no editing (a) or pairs with editing that decrease with age.

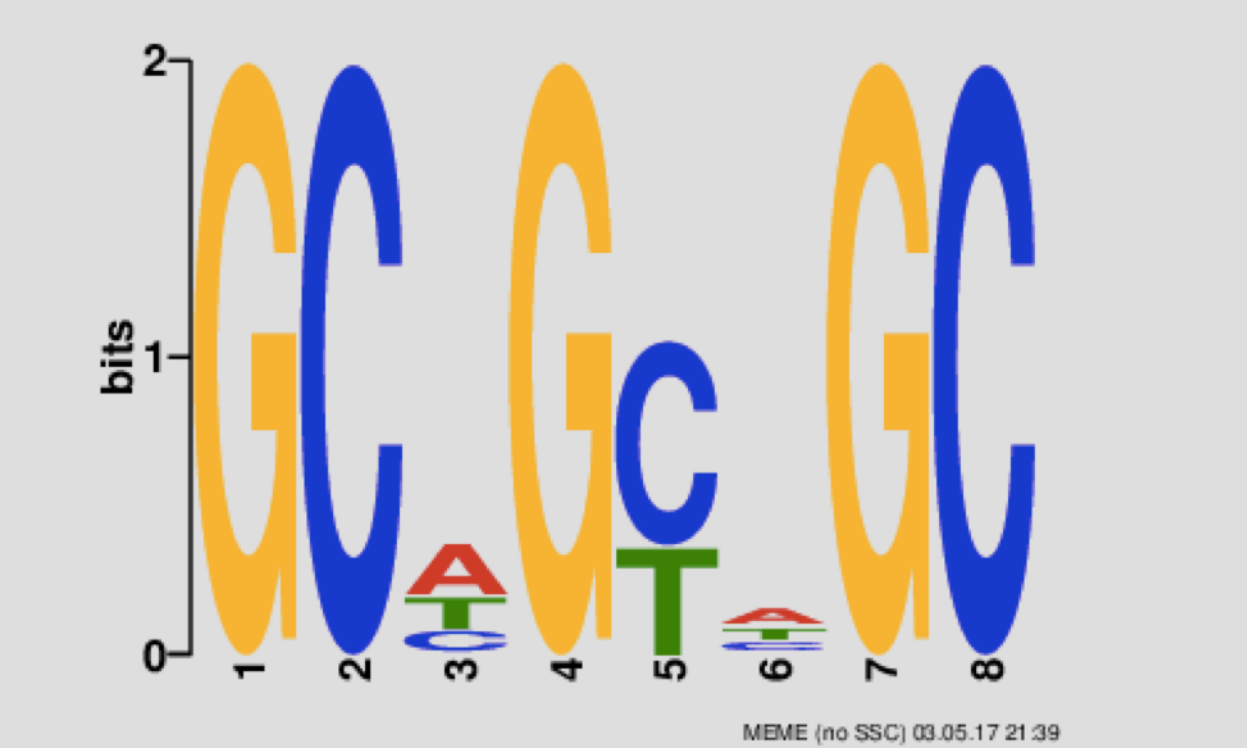

Figure 17: Motif analysis finds a strong preference for the motif GCAGCAGC in more than 50% of flanking pairs around circular RNA that contain editing sites that increase with age

Appendix B

Tables

Drosophila CLK genes	Mammalian CLK genes
period(per)	Period 1 (Per1)
	Period 2 (Per2)
	Period 3 (Per3)
timeless(tim)	Timeless(Tim)
doubletime (dbl)	Casein kinase 1
clock(Clk)	Circadian locomotor output cycle kaput (Clock)
cycle(cyc)	Bmal1/MOP3
cryptochrome(cry)	Cryptochrome 1
	Cryptochrome 2
virile(vri)	NII3/E4BP4
par domain protein 1 (pdp1)	

Table 1 Several core circadian proteins are conserved between Drosophila and mammals. The Drosophila CLK/CYC heterodimer is analogous to the CLK/BMAL1 complex in mammals.

	Genotype	Time points	Data Type	Read length
Library 1	CS WT	6	Pooled	200bp
Library 2	CS WT	6	Pooled	150bp
Library 3	CS WT	6	Individual	150bp

Table 2 Three ATAC-seq libraries, consisting of six time points each were used in this investigation. Library 1 and library 2 were samples resequenced to obtain higher coverage and then pooled together.

		Rep1	Rep2	Rep3
Library 1	R	-	-	-
	R Standard Error			
Library 2	R	0.98	-	-
	R Standard Error	<< 0.01		
Library 3	R	0.97	0.96	-
	R Standard Error	<< 0.01	<< 0.01	

Table 3 Via a Pearson's correlation coefficient test, all libraries in this study were highly correlated for all time points.

	Default Parameters	Extending window size +/-200 bp
Total number of peaks called:	25733	1027
Average distance to TSS:	1242	1570
Percent mapping to protein coding gene:	83	82
Percent of significant peaks between ZT2 and ZT14:	6	7.0
Percent of unique peaks:	4	84

Table 4 Peaks called using MACS2 default parameters and extending window size by 200bp. Extending the window size called less peaks, but less peaks per gene.

Gene at ebox	logFold Change	P-Value
tim	0.028	0.82
per	0.086	0.52
vri	0.59	0.18
vri	-0.54	0.0077
Pdp1	0.16	0.33

Gene at ebox	LogFold change	P-Value
Pdp1	0.082	0.58

Table 5 Fold change and significance thresholds for core circadian genes between ZT2 and ZT14 using both peak calling methods. (Top) default parameters called more relevant circadian peaks than were called (bottom) extending the window size. Neither method produced a large number of CCGs with a significant fold change.

	Number of sites
Editing sites that increase with age	1445
Editing sites that decrease with age	1015
All editing sites	7996

Table 6 Number of total ADAR A-I editing sites found, editing sites that increase in old brains, and editing sites that decrease in old brains

	% Splice Site Variant	% Splice site variants that change with age	increase with age	decrease wit
All editing sites	0.015% (99 sites)	86	34	41
Increase with age	0.32% (33 sites)	-	-	-
Decrease with age	0.57% (44 sites)	-	-	-

Table 7 Percent and number of editing sites that are predicted to be splice site variants. The actual number of splice site variants is very small, however, of the 99 sites associated with splice site variants, the majority are associated with age.

	% Splice site editing events at either 5' or 3' splice site around circRNAs
All editing sites	99- 70%
Increase with age	22/33 - 67%
Decrease with age	36/44 - 82%

Table 8 Of the small number of splice site variants, the majority are associated with circular RNA formation. Splice sites that change with age seem to be specifically associated with editing sites that decrease with age

Category:	Enriched for:
All splice site variants	Localization (p=5.14e-4)
Splice sites that decrease with age	Synaptic plasticity (p=0.0073)
Splice sites that increase with age	Localization (p=0.019)

Table 9 While splice site variants that are not age associated and increase with age are only associated with a localization gene ontology cluster, all splice site variants caused by editing that decreases with age are associated with synaptic plasticity.

Flanking pair with:	Mean max value	SD
Increases with age	25.98	35.71
Decreases with age	22.51	22.91
No editing	18.09	2.45

Table 10 Of all flanking pairs around circular RNAs that increase with age, there was a significantly higher level of complementarity in the pair than all pairs that contain events that decrease with age or pairs with no editing.

flanking pair		before editing	with editing	change
chr3L:10567920-10568420	chr3L:10570424-10570924	-152.3	-154.9	-2.6
chr2L:21383956-21384456	chr2L:21389079-21389579	-84.9	-87.2	-2.3
chr4:301851-302351	chr4:303364-303864	-107.8	-107.8	0
chr4:1251530-1252030	chr4:1252562-1253062	-106.3	-110.6	-4.3
chrX:14990990-14991490	chrX:15003926-15004426	-194.8	-195.1	-0.3
chr2L:21383956-21384456	chr2L:21385397-21385897	-84.9	-87.2	-2.3
chr4:1250329-1250829	chr4:1259724-1260224	-117.6	-117.3	0.3
chr2L:21712270-21712770	chr2L:21717672-21718172	-121.6	-123.7	-2.1
chr3L:4577418-4577918	chr3L:4580159-4580659	-122.3	-122.3	0
chrX:17932619-17933119	chrX:17934416-17934916	-159.8	-161.2	-1.4

Table 11 Of a random sample of 10 flanking pairs containing editing events that increase with age around circular RNA also associated with age, editing events seem to be modestly stabilizing secondary RNA structure from minimum free energy analysis (where the more negative the value, the more stable the structure).

Bibliography

[1] Yunshun Chen, D. McCarthy, M. Ritchie, M. Robinson, and G. K. Smyth, "edgeR: differential expression analysis of digital gene expression data."

[2] M.-S. Cheung, T. A. Down, I. Latorre, and J. Ahringer, "Systematic bias in high-throughput sequencing data and its correction by BEADS," vol. 39, no. 15, pp. e103–e103.

[3] W. S. Cleveland, "Robust locally weighted regression and smoothing scatterplots," vol. 74, no. 368, pp. 829–836.

[4] O. Tataroglu and P. Emery, "Studying circadian rhythms in drosophila melanogaster," vol. 68, no. 1, pp. 140–150.

[5] G. dos Santos, A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby, J. Thurmond, D. B. Emmert, and W. M. Gelbart, "FlyBase: introduction of the drosophila melanogaster release 6 reference genome assembly and large-scale migration of genome annotations," vol. 43, pp. D690–D697.

[6] E. Lasda and R. Parker, "Circular RNAs: diversity of form and function," vol. 20, no. 12, pp. 1829–1842.

[7] S. P. Barrett and J. Salzman, "Circular RNAs: analysis, expression and potential functions," vol. 143, no. 11, pp. 1838–1847.

[8] W. R. Jeck and N. E. Sharpless, "Detecting and characterizing circular RNAs," vol. 32, no. 5, pp. 453–461.

[9] W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, and F. Cunningham, "The ensembl variant effect predictor," vol. 17, p. 122.

[10] D. Grün, Y.-L. Wang, D. Langenberger, K. C. Gunsalus, and N. Rajewsky, "microRNA target predictions across seven drosophila species and comparison to mammalian targets," vol. 1, no. 1, p. e13.

[11] G. Varani and W. H. McClain, "The g·u wobble base pair," vol. 1, no. 1, pp. 18–23.

[12] J. O. Westholm, P. Miura, S. Olson, S. Shenker, B. Joseph, P. Sanfilippo, S. E. Celniker, B. R. Graveley, and E. C. Lai, "Genome-wide analysis of drosophila circular RNAs reveals their structural and sequence properties and age-dependent neural accumulation," vol. 9, no. 5, pp. 1966–1980.

[13] J. E. Wilusz and P. A. Sharp, "A circuitous route to noncoding RNA," vol. 340, no. 6131, pp. 440–441.

[14] T. Hideyama and S. Kwak, "When does ALS start? ADAR2–GluA2 hypothesis for the etiology of sporadic ALS," vol. 4.

[15] L.-L. Chen, "The biogenesis and emerging roles of circular RNAs," vol. 17, no. 4, pp. 205–211.

[16] W. R. Jeck, J. A. Sorrentino, K. Wang, M. K. Slevin, C. E. Burd, J. Liu, W. F. Marzluff, and N. E. Sharpless, "Circular RNAs are abundant, conserved, and associated with ALU repeats," vol. 19, no. 2, pp. 141–157.

[17] Y. Wang and Z. Wang, "Efficient backsplicing produces translatable circular mRNAs," vol. 21, no. 2, pp. 172–179.

[18] L.-L. Chen and G. G. Carmichael, "Altered nuclear retention of mRNAs containing inverted repeats in human embryonic stem cells: Functional role of a nuclear noncoding RNA," vol. 35, no. 4, pp. 467–478.

[19] M. J. McDonald and M. Rosbash, "Microarray analysis and organization of circadian gene expression in drosophila," vol. 107, no. 5, pp. 567–578.

[20] P. Taylor and P. E. Hardin, "Rhythmic e-box binding by CLK-CYC controls daily cycles in per and tim transcription and chromatin modifications," vol. 28, no. 14, pp. 4642–4652.

[21] J. D. Buenrostro, B. Wu, U. M. Litzenburger, D. Ruff, M. L. Gonzales, M. P. Snyder, H. Y. Chang, and W. J. Greenleaf, "Single-cell chromatin accessibility reveals principles of regulatory variation," vol. 523, no. 7561, pp. 486–490.

[22] "Overview of circadian rhythms."

[23] E. C. R. Reeve and I. Black, *Encyclopedia of Genetics*. Taylor & Francis. Google-Books-ID: JjLWYKqehRsC.

[24] E. Rosato, E. Tauber, and C. P. Kyriacou, "Molecular genetics of the fruit-fly circadian clock," vol. 14, no. 6, pp. 729–738.

[25] K. C. Abruzzi, J. Rodriguez, J. S. Menet, J. Desrochers, A. Zadina, W. Luo, S. Tkachev, and M. Rosbash, "Drosophila CLOCK target gene characterization: implications for circadian tissue-specific gene expression," vol. 25, no. 22, pp. 2374–2386.

[26] J. S. Menet, S. Pescatore, and M. Rosbash, "CLOCK:BMAL1 is a pioneer-like transcription factor," vol. 28, no. 1, pp. 8–13.

[27] J. D. Buenrostro, P. G. Giresi, L. C. Zaba, H. Y. Chang, and W. J. Greenleaf, "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position," vol. 10, no. 12, pp. 1213–1218.

[28] "kundajelab/atac_dnase_pipelines."

[29] Q. Li, J. B. Brown, H. Huang, and P. J. Bickel, "Measuring reproducibility of high-throughput experiments," vol. 5, no. 3, pp. 1752–1779.

[30] H. Koohy, T. A. Down, M. Spivakov, and T. Hubbard, "A comparison of peak callers used for DNase-seq data," vol. 9, no. 5, p. e96303.

[31] N. U. Rashid, P. G. Giresi, J. G. Ibrahim, W. Sun, and J. D. Lieb, "ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions," vol. 12, no. 7, p. R67.

[32] Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu, "Model-based analysis of ChIP-seq (MACS)," vol. 9, p. R137.

[33] A. M. Ackermann, Z. Wang, J. Schug, A. Naji, and K. H. Kaestner, "Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes," vol. 5, no. 3, pp. 233–244.

[34] "taoliu/MACS."

[35] Y. Liao, G. K. Smyth, and W. Shi, "featureCounts: an efficient general purpose program for assigning sequence reads to genomic features," vol. 30, no. 7, pp. 923–930.

[36] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, "edgeR: a bioconductor package for differential expression analysis of digital gene expression data," vol. 26, no. 1, pp. 139–140.

[37] J. T. Robinson, H. Thorvaldsdóttir, W. Winckler, M. Guttman, E. S. Lander, G. Getz, and J. P. Mesirov, "Integrative genomics viewer," vol. 29, no. 1, pp. 24–26.

[38] J. Rodriguez, C.-H. A. Tang, Y. L. Khodor, S. Vodala, J. S. Menet, and M. Rosbash, "Nascent-seq analysis of drosophila cycling gene expression," vol. 110, no. 4, pp. E275–E284.

[39] M. M. Bellet and P. Sassone-Corsi, "Mammalian circadian clock and metabolism – the epigenetic link," vol. 123, no. 22, pp. 3837–3848.

[40] B. Langmead and S. L. Salzberg, "Fast gapped-read alignment with bowtie 2," vol. 9, no. 4, pp. 357–359.

[41] S. Kadener, D. Stoleru, M. McDonald, P. Nawathean, and M. Rosbash, "Clockwork orange is a transcriptional repressor and a new drosophila circadian pacemaker component," vol. 21, no. 13, pp. 1675–1686.

[42] J. Zhou, W. Yu, and P. E. Hardin, "CLOCKWORK ORANGE enhances PERIOD mediated rhythms in transcriptional repression by antagonizing e-box binding by CLOCK-CYCLE," vol. 12, no. 11, p. e1006430.

[43] R. Pique-Regi, J. F. Degner, A. A. Pai, D. J. Gaffney, Y. Gilad, and J. K. Pritchard, "Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data," vol. 21, no. 3, pp. 447–455.

[44] A. Rybak-Wolf, C. Stottmeister, P. Glažar, M. Jens, N. Pino, S. Giusti, M. Hanan, M. Behm, O. Bartok, R. Ashwal-Fluss, M. Herzog, L. Schreyer, P. Papavasileiou, A. Ivanov, M. Öhman, D. Refojo, S. Kadener, and N. Rajewsky, "Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed," vol. 58, no. 5, pp. 870–885.

[45] A. Ivanov, S. Memczak, E. Wyler, F. Torti, H. T. Porath, M. R. Orejuela, M. Piechotta, E. Y. Levanon, M. Landthaler, C. Dieterich, and N. Rajewsky, "Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals," vol. 10, no. 2, pp. 170–177.

[46] Y. A. Savva, L. E. Rieder, and R. A. Reenan, "The ADAR protein family," vol. 13, no. 12, p. 252.

[47] X. Li, I. M. Overton, R. A. Baines, L. P. Keegan, and M. A. O'Connell, "The ADAR RNA editing enzyme controls neuronal excitability in drosophila melanogaster," vol. 42, no. 2, pp. 1139–1151.

[48] K. Larsen, K. K. Kristensen, J. Momeni, L. Farajzadeh, and C. Bendixen, "A-to-i RNA editing of the IGFBP7 transcript increases during aging in porcine brain tissues," vol. 479, no. 3, pp. 596–601.

[49] A. P. Holmes, S. H. Wood, B. J. Merry, and J. P. de Magalhães, "A-to-i RNA editing does not change with age in the healthy male rat brain," vol. 14, no. 4, pp. 395–400.

[50] I. L. Hofacker and P. F. Stadler, "Memory efficient folding algorithms for circular RNA secondary structures," vol. 22, no. 10, pp. 1172–1176.

[51] J. A. Cuesta and S. Manrubia, "Enumerating secondary structures and structural moieties for circular RNAs," vol. 419, pp. 375–382.

[52] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, and S. L. Salzberg, "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions," vol. 14, p. R36.

[53] Y. Yang, X. Zhou, and Y. Jin, "ADAR-mediated RNA editing in non-coding RNA sequences," vol. 56, no. 10, pp. 944–952.

[54] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," vol. 215, no. 3, pp. 403–410.

[55] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, and T. L. Madden, "BLAST+: architecture and applications," vol. 10, p. 421.

[56] P. Rice, I. Longden, and A. Bleasby, "EMBOSS: The european molecular biology open software suite," vol. 16, no. 6, pp. 276–277.

[57] M. Stapleton, J. W. Carlson, and S. E. Celniker, "RNA editing in drosophila melanogaster: New targets and functional consequences," vol. 12, no. 11, pp. 1922–1932.

[58] E. Y. Levanon, E. Eisenberg, R. Yelin, S. Nemzer, M. Hallegger, R. Shemesh, Z. Y. Fligelman, A. Shoshan, S. R. Pollock, D. Sztybel, M. Olshansky, G. Rechavi, and M. F. Jantsch, "Systematic identification of abundant a-to-i editing sites in the human transcriptome," vol. 22, no. 8, pp. 1001–1005.

[59] S. M. Rueter, T. R. Dawson, and R. B. Emeson, "Regulation of alternative splicing by RNA editing," vol. 399, no. 6731, pp. 75–80.

[60] D. P. Morse, P. J. Aruscavage, and B. L. Bass, "RNA hairpins in noncoding regions of human brain and caenorhabditis elegans mRNA are edited by adenosine deaminases that act on RNA," vol. 99, no. 12, pp. 7906–7911.

[61] S. Nainar, P. R. Marshall, C. R. Tyler, R. C. Spitale, and T. W. Bredy, "Evolving insights into RNA modifications and their functional diversity in the brain," vol. 19, no. 10, pp. 1292–1298.

[62] M. Behm and M. Öhman, "RNA editing: A contributor to neuronal dynamics in the mammalian brain," vol. 32, no. 3, pp. 165–175.

[63] J. E. Robinson, J. Paluch, D. K. Dickman, and W. J. Joiner, "ADAR-mediated RNA editing suppresses sleep by acting as a brake on glutamatergic synaptic plasticity," vol. 7, p. 10512.

[64] M. E. Hughes, G. R. Grant, C. Paquin, J. Qian, and M. N. Nitabach, "Deep sequencing the circadian and diurnal transcriptome of drosophila brain," vol. 22, no. 7, pp. 1266–1281.

[65] H. Terajima, H. Yoshitane, H. Ozaki, Y. Suzuki, S. Shimba, S. Kuroda, W. Iwasaki, and Y. Fukada, "ADARB1 catalyzes circadian a-to-i editing and regulates RNA rhythm," vol. 49, no. 1, pp. 146–151.

[66] S. Qu, X. Yang, X. Li, J. Wang, Y. Gao, R. Shang, W. Sun, K. Dou, and H. Li, "Circular RNA: A new star of noncoding RNAs," vol. 365, no. 2, pp. 141–148.

[67] R. Ashwal-Fluss, M. Meyer, N. R. Pamudurti, A. Ivanov, O. Bartok, M. Hanan, N. Evantal, S. Memczak, N. Rajewsky, and S. Kadener, "circRNA biogenesis competes with pre-mRNA splicing," vol. 56, no. 1, pp. 55–66.

[68] M. Zuker, "Mfold web server for nucleic acid folding and hybridization prediction," vol. 31, no. 13, pp. 3406–3415.

[69] A. R. Gruber, R. Lorenz, S. H. Bernhart, R. Neuböck, and I. L. Hofacker, "The vienna RNA websuite," vol. 36, pp. W70–W74.

[70] Y. Ding, C. Y. Chan, and C. E. Lawrence, "Sfold web server for statistical folding and rational design of nucleic acids," vol. 32, pp. W135–141.

[71] N. R. Markham and M. Zuker, "UNAFold: software for nucleic acid folding and hybridization," vol. 453, pp. 3–31.

[72] P. Machanick and T. L. Bailey, "MEME-ChIP: motif analysis of large DNA datasets," vol. 27, no. 12, pp. 1696–1697.