Next-Generation Sequencing Elucidates RNA-Editing in Discrete Cell Populations of the Circadian System in *Drosophila*

Master's Thesis

Presented to

The Faculty of the Graduate School of Arts and Sciences
Brandeis University
Interdepartmental Program in Neuroscience
Michael Rosbash, Advisor

In Partial Fulfillment
of the Requirements for

Master's Degree

by

Abigail N. Zadina

May 2013
ABSTRACT

Next-Generation Sequencing Elucidates RNA-Editing in Discrete Cell Populations of the Circadian System in *Drosophila*

A thesis presented to the Interdepartmental Program in Neuroscience

Graduate School of Arts and Sciences Brandeis University Waltham, Massachusetts

By Abigail N. Zadina

The circadian system of *Drosophila* is composed of discrete populations of neurons. Previous studies purified two of these populations, the sLNvs and lLNvs, and characterized their expression using microarrays. These studies were useful in elucidating the functions of these cells and identifying new circadian genes. However, microarray data is limited and cannot provide information regarding isoform expression, alternative splicing, or RNA-editing. Therefore, using the same purification method, we sought to create RNA-sequencing libraries from these populations and use the data to identify RNA-editing sites. Overall, we were able to create 3'-biased libraries by extracting RNA from cells and amplifying the mRNA using dT-priming. We did this for PDF- and TH-expressing cells, which encompass
the sLNvs/ILNvs and dopaminergic neurons in the brain, respectively. Using the sequencing data from these cells we identified potential editing sites, which we then verified using site-specific PCR. In this manner we identified a total of 20 editing sites. 17 of these sites had been previously identified in heads, and 3 were novel and fly strain specific. Additionally, some of the edited sites found in both cell types showed variations in editing levels and cycling across two time-points. These results suggest that RNA-editing may be regulated in a tissue-specific manner and that the circadian system may control some aspects of this regulation.
# TABLE OF CONTENTS

Title Page  i  
Copyright  ii  
Abstract  iii  
List of Figures and Tables  vi  
Introduction  1  
1. RNA-Seq from Cells  7  
   1.1 Background  
   1.2 Results  
   1.3 Discussion  
   1.4 Figures  
2. RNA-Editing in PDF- and TH-cells  20  
   2.1 Background  
   2.2 Results  
   2.3 Discussion  
   2.4 Figures  
Conclusions  42  
Appendix  44  
   A.1 Supplementary Tables  
   A.2 Materials and Methods  
References  50
LIST OF FIGURES AND TABLES

Figure 1.1 Methodology for Creating Libraries from Cells: 15
1.2 RNA Amplification Protocol: 16
1.3 Quantification of RNA from Cells: 17
1.4 Validation of cDNA#1: 17
1.5 Quantification of cRNA#1: 18
1.6 3’-biased Libraries: 18
1.7 Random vs dT Primed Libraries: 19

Figure 2.1 Bioinformatic Analysis of RNA-Editing: 37
2.2 PDF-expressing Neurons: 37
2.3 Editing Levels in Cells: 38
2.4 Strain Specific Editing Sites: 38
2.5 Edited Gene Ontology: 39
2.6 Cycling Editing Levels in Cells: 39
2.7 Schematic of Insect nAChR: 40
2.8 Editing in nAcRalpha-34E: 40
2.9 Editing in Brd8: 41

Supplementary Table 1 Editing Sites Across Cell Types: 44
INTRODUCTION

Circadian rhythms are behavioral and physiological oscillations that occur over twenty-four hour periods. Driven by pacemaker cells in the mammalian suprachiasmatic nucleus and a subset of 150 neurons in the *Drosophila* brain (1, 2), these evolutionarily conserved cycles enable organisms to respond to and anticipate daily changes in the external environment and serve to coordinate in time processes at the molecular, cellular, tissue, and systems levels. These circadian influenced processes impact sleep/wake behavior, metabolism, immune response, learning and memory, reproduction, and mating (1). Moreover, the disruption of the circadian clock has been implicated in several metabolic, sleep, neurological, and neurodegenerative diseases (3, 4), further underscoring the importance of circadian regulation to organismal health and survival.

Underlying these rhythms is a molecular feedback mechanism consisting of four key proteins. In *Drosophila melanogaster*, these proteins are CLOCK (CLK) (5), CYCLE (CYC) (6), PERIOD (PER) (7-9), and TIMELESS (TIM) (10), which are homologous to the mammalian CLOCK, BMAL1 (corresponding to CYC), PERIOD1, and PERIOD2 proteins. There is no mammalian homolog of TIM, which is instead replaced by mammalian CRY1 and CRY2 in the core feedback loop (2). In this negative feedback loop, the transcription factors CLK and CYC form a heterodimer and activate the expression of *per* and *tim* (11). PER and TIM subsequently dimerize
in the cytoplasm, enter the nucleus, and bind the CLK/CYC heterodimer, effectively turning off their own expression (12).

This alternation of transcriptional activation and repression cycles with a twenty-four hour period. The CLK/CYC heterodimer drives *per* and *tim* transcription during the day, with peak levels of *per* and *tim* accumulating by late day/early night. The PER/TIM dimer then translocates to the nucleus during the night, inhibiting CLK/CYC activity (13). The repression of CLK/CYC continues until PER and TIM are degraded. The ubiquitin ligase, SLIMB, ubiquitinates phosphorylated PER, signaling PER’s degradation (14), while TIM undergoes light-induced degradation that is mediated by the photoreceptor, CRYPTOCHROME (CRY) (15, 16), and the ubiquitin ligase JETLAG (17).

Phosphorylation states not only impact the degradation of the core clock proteins, but also play a role in their functional activity. For example, PER hyperphosphorylation by DOUBLETIME (DBT, CASEIN KINASE Iε) leads to PER ubiquitination and degradation (18), and changes the repressor activity of PER (19). The phosphorylation state of PER is rhythmic due to the actions of phosphatase PP2A (20). As for TIM, phosphorylation by SHAGGY (SGG, GLYCOGEN SYNTHASE KINASE 3β) assists in the translocation of TIM to the nucleus (21). Additionally, phosphorylation of CLK by a PER-DBT dimer facilitates the cycling of CLK’s functional activity and leads to CLK degradation (18, 22).

Besides the principal feedback loop involving CLK/CYC activation of *per* and *tim* transcription, there is an additional loop involving the transcription of *Par domain protein 1 (Pdp1)* and *vrille (vri)* (23). CLK/CYC drives the expression of
Pdp1 and vri, which are transcriptional activators and repressors, respectively. PDP1 then activates clk and cry transcription at a time opposite of when per and tim expression occurs. VRI, which accumulates faster than PDP1, represses clk and cry, reinforcing the timing of circadian feedback loops.

The principal and ancillary feedback loops likely have considerable influence over downstream systems. Expression studies using fly heads have identified many genes whose transcription appears to cycle (24-26). Other studies have shown that key circadian transcription factors, such as CLK and CYC, rhythmically bind to gene targets (27). These studies are limited, however, by the heterogeneity of the fly head. For example, the genes bound by CLK change when the eyes are ablated from the head. Moreover, when subsets of neurons were sorted and their expression profiles characterized with arrays, as was done for the small and large ventral-lateral neurons (28, 29), the profiles were markedly different from pan-neuronal expression experiments. This indicates that further characterization of circadian genes and dynamics at the level of specific neuronal groups will be beneficial for identifying new clock genes and outputs.

Within the fly brain there are three groups of clock neurons, i.e., neurons expressing CLK, CYC, PER, and TIM. Identified using immunostaining techniques, these clock neurons are categorized as the lateral neurons (LNs), dorsal neurons (DNs), and lateral posterior neurons (LPNs) (30-32). The LNs are further broken down into four groups: small lateral-ventral neurons (sLNvs), large lateral-ventral neurons (ILNvs), the 5th sLNv, and the lateral-dorsal neurons (LNds). The sLNvs and ILNvs are commonly denoted as PDF-cells because they are the only neurons in the
brain that express PIGMENT DISPERSING FACTOR (PDF), a neuropeptide (33). The PDF-cells likely target areas near the Pars Intercerebralis, a brain region that influences growth and sleep (34). There are three groups of DNs, the DN1s, DN2s, and DN3s. The DN1s consist of anterior and posterior clusters (DN1ps and DN1as) that are GLASS (a transcription factor) negative and positive, respectively (35). Additionally, the DN1ps can be either CRY-positive or -negative (36, 37). The DN1s and DN3s are glutamatergic and are believed to provide input to the sLNvs (38).

Each of these neuronal groups contributes to the maintenance of circadian behavior. Flies typically show oscillations in locomotor activity, being more active during the day with peaks of activity appearing in the morning, at lights-on, and evening, at lights-off (39). Using related observations of hamster activity rhythms, Pittendrigh and Daan (40) developed the dual oscillator model to explain circadian behavior. The morning (M) oscillator is phase matched to sunrise and accelerates in response to light. The evening (E) oscillator is matched to sundown and decelerates in response to light. These proposed oscillators have been traced to specific neurons within the fly and are responsible for the observed morning and evening peaks in activity.

The sLNvs and CRY(+) DN1ps are responsible for the M peak (37, 41, 42), while the 5th-sLNv, LNds, and CRY(-) DN1ps control the E peak (36, 41, 42). When the sLNvs and ILNvs are genetically ablated the M peak disappeared. Similarly, when the LNds were ablated, flies did not display an E activity peak (42). Complementing this result, it was also shown that rescue of per in the LNvs of per01 flies, which are normally arrhythmic, restored M activity. This rescue was
dependent upon the sLNvs because, when per was rescued in only the ILNvs, the per^{01} flies remained arrhythmic (41).

The other circadian neuronal groups (e.g. ILNvs, DNs, LPNs) have been implicated in mediating light signals and entraining the clock to temperature oscillations. Light is the main environmental cue that entrains the clock to a 12-hour:12-hour, light-dark cycle. The light cues can impact the clock via the photoreceptor CRY, but also through neuronal signaling that originates from photoreceptors in the eyes, and/or Hobfauer-Buchner eyelets (43). This light-induced neuronal signaling may target the ILNvs, which show increased activity in response to light (44, 45). While the LNs appear to primarily mediate entrainment to light cues, the DNs and LPNs may be involved in relaying temperature cues.

When researchers placed flies in an environment in which the light-dark cycle was 90-degrees out of phase relative to the temperature cycle, the DNs and LPNs were found to have modified molecular oscillations. The molecular oscillations in the LNs, though, remained in phase with the light-dark cycle (46). Modifications in molecular oscillations in response to temperature are likely due to the differential splicing of per and tim transcripts, which has been shown to be temperature dependent (47).

Besides the 150 clock neurons, other neuronal populations may interact with the circadian system. The dopamine system in flies, which consists of about 200 neurons in the adult brain, is associated with learning and memory, as well as sleep and arousal levels (48). Dopaminergic neurons may be involved in the circadian system because CLK has been implicated in controlling dopamine levels by down-
regulating its synthesis at night. *Clk* mutant flies also show increased nighttime activity, which is mediated by elevated dopamine in the brain. Additionally, the increased activity due to dopamine requires the action of CRY in the ILNvs, further tying the dopamine and circadian circuitry together (49). Finally, dopamine may even impact the ability of the clock to entrain to some types of light. Flies deficient for dopamine were unable to entrain to low intensity blue light. These same flies also showed weakened circadian motor rhythms (50).

Considering that distinct cells are responsible for various aspects of circadian behavior, the work presented here focuses on further characterizing the transcriptomes of two relevant populations, PDF-expressing neurons and TH-expressing neurons, using next-generation sequencing of RNA libraries. Though gene expression in PDF-positive cells was characterized previously using microarrays (28, 29), those data are limited to gene expression levels only and cannot provide insight into isoform expression, differential splicing, or RNA-editing. Due to the limitations of microarrays, we set out to develop adequate protocols for the creation of RNA-sequencing libraries from small collections of cells and use the sequencing data obtained to analyze RNA-editing in these discrete cell populations.
CHAPTER ONE

RNA-Seq from Cells

1.1 BACKGROUND

Many studies of the Drosophilian brain rely on the use of whole heads. However, the head is made up of many tissues that are involved in discrete processes. One method to elucidate the nature of these different populations is to purify the tissue of interest and characterize its gene expression. This was done previously to study the circadian system (28, 29). In these studies, LNvs in Drosophila were selectively labeled with Green Fluorescent Protein (GFP) using a UAS/Gal4 binary system (PDF-GAL4>UAS-mcD8GFP). Brains from these modified flies were dissected and dissociated into individual cells across four time-points. The dissociated cells were then plated and GFP-positive cells were manually sorted under a fluorescent dissecting microscope, and separated into samples of lLNvs and sLNvs. Approximately one hundred cells per time-point were lysed and processed for RNA extraction. The RNA was then amplified using reverse-transcription PCR (RT-PCR), which employed a dT-primer fused to a T7 promoter sequence (dT7-RT-PCR), and in-vitro transcription (IVT), as outlined in Fig. 1.1 and detailed in Fig. 1.2. The amplified RNA, enriched for mRNA transcripts via dT-priming, was then analyzed using Drosophila genome arrays. This array data was used to look for enrichment of gene transcripts in lLNvs and sLNvs compared to pan-neuronal cell
collections and heads. While these studies were able to identify transcripts enriched uniquely in lLNs and sLNs, and identify new circadian genes, analysis beyond gross expression levels was limited by the microarray technology employed.

Microarrays are solid chips to which short oligonucleotide probes are attached. These probes are genome specific and generally designed to hybridized to the 3’-end of genes (Affymetrix GeneChip *Drosophila* Genome 2.0 Array). Therefore, array data fail to quantify the 5’-end of transcripts, data necessary for characterizing alternative start-sites, isoform expression, and alternative splicing within the transcriptome. Arrays also cannot provide information concerning SNPs and RNA-editing, which require resolution down to nucleotide levels. Finally, from a bioinformatics standpoint, arrays prove to be variable across experiments and may inadequately quantify some transcripts due to probe saturation (51).

In order to gain nucleotide-sequence resolution and obtain information regarding isoform expression, alternative splicing, and RNA-editing in specific subsets of neurons, we applied the same method of cell sorting and RNA amplification to the creation of sequencing libraries from PDF-expressing cells. Using this method (Fig. 1.2), we obtained biased libraries enriched for the 3’-end of mRNA transcripts for PDF-expressing and ELAV-expressing neurons. This bias is likely due to the inefficiency of reverse transcriptase, which generally fails to fully replicate the full mRNA strands when beginning from a 3’-dT primer. To try and overcome this bias we tried a method of amplification that employed a mix of both dT-T7 and random-T7 primers. While this method worked for diluted head RNA, it was unsuccessful for RNA samples from cells.
1.2 RESULTS

Two separate fly lines, one expressing ELAV-Gal4>UAS-eGFP (ELAV>gfp), and the other expressing PDF-Gal4>UAS-mcD8GFP (PDF>gfp), were entrained to a 12hr/12hr light:dark cycle at 22-25°C. At ZT2 and ZT14, flies from each strain were collected and their brains dissected. The brains (50-60 from PDF>gfp, and ~15 from ELAV>gfp) were then dissociated and approximately 100-120 GFP-positive cells were collected per sample. RNA was extracted and quantified on the 2100 Agilent BioAnalyzer using a PicoChip. From 100 cells, approximately 200-400 pg of RNA was obtained (Fig. 1.3). As indicated with blue arrows, the RNA had the characteristic ribosomal RNA peaks around 2000 nucleotides. Indicated with red arrows, the samples also showed peaks for <100 nucleotide sequences. The RNA samples were then amplified using dT-T7 primers in a four step amplification process (Fig. 1.2): 1) cDNA#1 synthesis, 2) IVT Round #1 (cRNA#1), 3) cDNA#2 synthesis, 4) IVT Round #2 (cRNA#2).

cDNA#1 from all cell samples, and primers for pdf and tim mRNA, were used to validate the identity of the cell collections and check that the cycling levels of circadian transcripts is preserved through the amplification process. As shown in Fig. 1.4a, pdf mRNA enrichment in PDF-cell collections is 40,000-60,000 times greater than that of ELAV-cell collections. tim mRNA is also enriched in PDF-cell collections (Fig. 1.4b) and cycles as expected between low and high abundance from ZT2 to ZT14, respectively.

The integrity and amounts of cRNA#1 was examined using a PicoChip and the 2100 Agilent Bioanalyzer (Fig. 1.5). Roughly 90-200 ng per sample of cRNA was
obtained from IVT Round #1 of the amplification process. This cRNA included sequences 25 to >4000 nucleotides long, but showed peak abundance for sequences of length ~600 nucleotides and <200 nucleotides.

cRNA#1 was subsequently amplified with a second round of cDNA synthesis and IVT. The resulting cRNA#2 was used to make RNA-seq libraries according to Illumina’s RNA TRUseq v2 Kit. The libraries were sequenced and the reads were mapped to the Drosophila genome using TopHat. Quantification of expression levels was carried out using Cufflinks and read densities across the genome were visualized using IGV_2.2.10. As shown in Fig. 1.6a, the cell identity of each sample was maintained. PDF-gfp cell collections showed characteristic enrichment for pdf mRNA. Additionally, cycling of tim mRNA was maintained from ZT2 to ZT14 in both cell sets (Fig. 1.6b). Finally, as indicated by the arrows, the libraries are heavily biased toward the 3’-end of transcripts. There are almost no reads in the first exon of tim in any of the libraries. This bias is consistent across long genes. In the 5’-UTR of Gapdh2, there are few reads compared to the 3’-end (Fig. 1.6c).

To eliminate the 3’-bias, we modified the amplification protocol to use random-T7 primers in the first cDNA synthesis. We used two samples of 3ng of diluted RNA from fly heads and two samples of RNA that had been extracted from 100 ELAV-cells to test four cDNA synthesis conditions: 1) head RNA with dT-T7 priming, 2) head RNA that was pA-purified with oligo-dT beads and random-T7 priming, 3) cell RNA, pA-purified with random-T7 priming, and 4) cell RNA, pA-purified with a mix of random- and dT-T7 priming. The cDNA from each condition was then used in one round of IVT and the cRNA was used to make Illumina RNA-
seq libraries. The library from 3ng of diluted head RNA that had been pA-purified and random-T7 primed showed no 3'-bias (Fig. 1.7), but instead was slightly 5'-biased and only mapped 37% of reads. The 5'-bias was exacerbated when cell RNA and random-T7 priming was used. Finally, when cell RNA and a mix of random- and dT-T7 priming was used, the majority of any 3'- or 5'-bias was eliminated. However, in the library from cells using mixed priming, only 11% of sequencing reads mapped to the genome. This is a significant decrease from the cell libraries prepped using only dT-T7 priming, which generally map at levels of 50-70%.
1.3 DISCUSSION

Extracting and amplifying RNA using dT-priming resulted in the enrichment of the 3’-end of mRNA transcripts. Libraries from this amplified material were therefore biased. This bias is likely due to the failure of reverse transcriptase to reach the 5’-end of the majority of transcripts during cDNA synthesis. Testing across three different reverse transcriptases, it was found that the abundance cDNA synthesized drops off significantly around ~1000 base pairs (data not shown). In the amplification protocol, these truncated sequences then become more enriched in subsequent IVT reactions, because more copies of the shorter sequences can be made in a given time period compared to the number of copies of long sequences that can be synthesized.

Compounding the limitations of reverse transcriptase, the RNA extracted from cells is of low quality. As seen in Fig. 1.3, the cell samples all contain peaks in the 25-100 nucleotide range. This suggests that the material is partially degraded. Degradation could result from RNase contamination during cell collection. The cell collection protocol take four to six hours to complete, and the cells are at room temperature for some of this time. We tried to minimize contamination during cell collection by keeping the cells on ice and filtering all of the collection buffers used. Degradation could also be due to the limits of RNA extraction methods from such limited numbers of cells. We tried two different methods of RNA extraction, one relying upon purification of whole RNA on a column, and the other using oligo-dT beads to extract mRNA, and found little difference in the quality of the resulting RNA-seq libraries.
To overcome the problem of 3'-bias, we tried a method of amplification that used a mix of dT and random primers. The amplification of rRNA, tRNA, and intergenic sequences is a concern when starting from whole RNA and using random priming. The whole RNA in cells is generally 60-90% rRNA. Unless rRNA is selected against during library preparation, potentially more than 75% of the reads from the sequenced library will be from rRNA sequences (52). Selection against rRNA sequences might be achieved during amplification by using random hexamers that do not exactly match any portion of rRNA sequences (53). Alternatively, one can select for mRNA using oligo-dT beads. We used the latter method, and found that the amplification of rRNA was limited in our samples. Overall, our method for pA-selection and random-T7 priming worked well for diluted whole RNA from heads. Most genes had reads across all axons, and there was only a slight 5'-bias. However, for RNA from cells, the introduction of random primers significantly increased the amount of contaminating oligos in the samples and resulted in libraries that were very biased towards the 5'-ends of genes. The 5'-bias was eliminated in cells when dT and random primers were mixed, but contaminating oligos were still an issue. The libraries from cells were as much as 90% non-mRNA material, the majority of which was from primer-dimers and non-\textit{Drosophilian} sources. To address these contamination issues, future experiments will need to focus on developing "clean" methods of working and size-selecting against primer-dimer artifacts.

Overall, we were able to create biased RNA-seq libraries from specific cell populations. These libraries showed 50-70% mapping to \textit{Drosophila} genes and
were able to replicate the expected behavior of known circadian genes. Given that samples amplified with a mix of dT and random primers introduced contaminating factors and mapped poorly, we decided to move forward with dT-primed libraries to analyze RNA-editing. RNA-editing occurs primarily in the 3’- and 5’-UTRs of transcripts (54). Therefore, we reasoned that the biased libraries might provide enough coverage to at least identify editing sites in the 3’-UTR and 3’-exons of genes. Unfortunately, the bias makes quantifying isoform expression and alternative splicing prohibitively difficult because shorter isoforms will be over represented relative to longer isoforms of the same gene.
**Fig 1.1** Methodology for the creation of RNA-seq libraries from discrete cell populations. Brains from flies expressing GFP in cells of interest were dissected and dissociated. The dissociated cell suspension was plated in a culture dish and fluorescent cells were sorted under a dissecting microscope. Samples of 100 cells were lysed and their RNA extracted. The RNA was then amplified using a four step process. Step #1 used either dT-T7 priming or a mix of dT-T7 plus random-T7 primers. After amplification, cRNA#2 was used in Illumina’s standard RNA TRUseq v2 protocol to make libraries. These libraries were then sequenced on the Illumina Genome Analyzer.
Fig 1.2 RNA Amplification Four Step Protocol. Step 1) mRNA is reverse transcribed using dT-T7 priming. The mRNA strand is degraded by RNaseH and the short sequences prime second strand synthesis. Step 2) RNA Polymerase transcribes the cDNA from the T7 promoter, linearly amplifying the the original transcripts. Step 3) cDNA is synthesized from the cRNA coming out of Step 2. This cDNA synthesis using random priming for the first strand and dT-T7 priming for the second strand. Step 4) IVT is performed using the cDNA from Step 3 and the resulting cRNA is used to make RNA-seq libraries. Note: The original 5'-end of the RNA transcript is in blue and gets shortened throughout the amplification process. The main source of 5'-end truncation is alerted to in the red circle and occurs in the initial cDNA synthesis.
Fig 1.3 RNA Extraction from ~100 cells. RNA was extracted from sorted cells using Arcturus Picopure Kit. The RNA was quantified using PicoChip on the Agilent 2100 BioAnalyzer. Total amounts of RNA extracted for each sample are indicated in the figure. Blue arrows point to the ribosomal RNA peak that is characteristic of whole RNA samples. Red arrows point to peaks in the 25-100 nucleotide range, which are likely indicative of degradation of the RNA. The spike at 25 nucleotides is a control marker.

Fig 1.4 Validation of cDNA#1 Amplified from Small Amounts of RNA. A) Signal of pdf mRNA from PDF-cells relative to signal from ELAV-cells. Samples were normalized to RPL32 signal. B) Cycling signal of tim mRNA normalized to RPL32 from PDF-cells and ELAV-cells.
Fig 1.5 Quantification of cRNA1 from RNA Amplification. cRNA from STEP 2 of amplification was quantified on a PicoChip using the Agilent 2100 Bioanalyzer. Total amounts of cRNA obtained from IVT are indicated in the figure. Red arrows point to a ~100 nucleotide peak that is likely the result of over enrichment of short nucleotide sequences. Average yield for cRNA1 was 150 ng. This is equivalent to a ~500-fold amplification from the original 300 pg of RNA coming from cells.

Fig 1.6 3’-Biased, dT-Primed and Amplified RNA-seq Libraries. Sequencing reads from libraries made using amplified cell RNA was visualized using IGV. A) Visualization of reads across pdf gene. B) tim gene. C) gapdh2 gene. Identity of the cell samples and time-points are indicated in the figure. Red arrows indicate the 3’ bias across long genes.
Fig 1.7 Random vs dT Primed Libraries. 3 ng of RNA from heads was amplified using dT-T7 priming and made into an RNA-seq library. Additionally, an RNA-seq library was made from 3 ng of RNA from heads that was pA-selected using dT-beads and amplified using random-T7 priming. Finally, libraries from ELAV-cell RNA were made using 1) pA-selection and random-T7 amplification and 2) pA-selection and a mix of random- and dT-T7 amplification. A) library visualization in IGV for tim. B) visualization of gapdh2. Red arrows indicate 3’-bias. Green arrows indicate 5’-bias.
CHAPTER TWO
RNA-Editing in PDF- and TH-cells

2.1 BACKGROUND

RNA-editing involves changing the sequence of RNA transcripts through the insertion, deletion, or conversion of nucleotide bases (54, 55). The first discovery of an editing event was in 1986 when researchers found four additional uridine bases within the cox2 transcript of trypanosomes (56, 57). Since then, editing events have been found in many other organisms and is believed to greatly contribute to the diversity of protein products that can be obtained from the genome. In fact, it has been noted that for many genomes, the number of genes encoded is much less than the number of unique proteins. RNA-editing, along with other mechanisms such as alternative splicing, is one way for biology to create these additional gene products. Editing is known to achieve this through affecting splice sites, modifying stop codons, and altering codons to specify different amino acids (54, 55).

RNA-editing is important for organismal health. Underediting and functional mutations in the editing proteins, ADARs, have been implicated in dyschromatosis symmetrica hereditaria, sporadic amyotrophic lateral sclerosis, and psychiatric disorders such as schizophrenia (54). Additionally, ADAR1 null mice die as embryos, while ADAR2 null mice show signs of seizures and die soon after birth. ADAR is also important in Drosophila. Mutant flies made null or hypomorphic for
ADAR display a lack of coordination, seizures, disrupted circadian activity patterns, and temperature sensitive paralysis (54, 55, 58).

In humans, two types of editing events have been described. The first, discovered in 1987, involved a cytidine to uridine (C to U) deamination event in the apolipoproteinB mRNA. This editing event is specific to liver and intestinal tissue, and results in the creation of an early stop codon, causing a truncated protein product (56, 59). APOBEC1 is the cytidine deaminase responsible for this C to U conversion. This enzyme is currently the only cytidine deaminase known to use RNA as a substrate (60). The other type of editing event is the deamination of adenosine to inosine (A to I), which is recognized as a guanosine (G) by ribosomal translational machinery. A to I editing is far more widespread and is carried out by ADAR (Adenosine Deaminase Acting on RNA) proteins (54, 55).

ADARs are evolutionarily conserved proteins. They contain a catalytic deaminase domain and at least one dsRNA binding domain. In mammals there are three ADARs (ADAR1, ADAR2, and ADAR3), while in insects there is only one, dADAR, which is most similar to the mammalian ADAR2 (54, 55). ADAR editing can either be promiscuous or specific. In the promiscuous case, long stretches of perfectly matched dsRNA can be edited at as many as 50% of adenosine sites. In specific editing, ADAR acts on imperfectly matched stretches of short dsRNA, editing only a few sites. The latter type of editing is seen more often in the coding regions of neuronal transcripts (55).

In Drosophila, more than 50 specific editing sites have been found in transcripts relevant to neuronal function. These sites are part of the more than 900
A to I editing sites that have been identified in annotated exons, in addition to more than 700 within introns (61, 62). A number of C to U conversions have also been discovered (63). The sites within neuronal transcripts include sites in ligand- and voltage-gated ion channels, as well as proteins involved in synaptic vesicle docking and release. These editing events are believed to impact the electrophysiological properties of neurons. For instance, the Shaker potassium channel is edited within a channel-lining transmembrane helix. The edited location with the pore likely impacts the ability of the inactivation gate to interact effectively with the channel, resulting in a K+ channel that cannot be inactivated (64). Editing sites in other channels may affect ion permeability and neurotransmitter affinity as well (55).

While editing appears to most specifically impact neural substrates, the affects across neuronal populations may be differential. When considering four A to I editing sites in the Shaker gene, researchers found that the combinations of how these sites were edited was different across Drosophila heads, thorax, eyes, antenna, and wings (65). Additionally, it has been reported that expression of dADAR is variable across neurons in the brain and that the efficiency of synaptogamin-1 (syt-1) editing at two different sites is varied across neuronal populations. For example, one particular syt-1 editing site had an editing level of ~30% (percent of transcripts with an I instead of A) in PDF-expressing neurons, and a ~80% level in GMR-expressing neurons (58).

Given the circadian phenotypes of hypomorphic dADAR mutant flies, and the differential dADAR expression and editing levels across neurons in the brain, we sought to analyze editing in two neuronal populations across two circadian time-
points using RNA-seq. We sequenced amplified RNA from PDF-expressing and dopaminergic neurons, and compared these sequences to genomic DNA. Though we were unable to identify any novel editing sites (not present in head data) specific to these cell populations due to low genome coverage in the RNA-seq samples, we were able to identify sites that potentially cycle in a circadian manner and display differential editing levels between populations. Moreover, we found different and novel editing sites between the two fly lines (strains used in cell collection), indicating that there is an advantage to having an I at those particular sites and that some strains accomplish this through hard coding in the genome, while others use RNA editing to achieve the advantageous nucleotide.
2.2 RESULTS

We sorted PDF-cells (Fig. 2.2) from a PDF-Gal4>UAS-mcD8GFP fly line at two time-points, ZT2 and ZT14. RNA from these cells was then extracted and dT-amplified as described in Chapter One. Libraries were made from the amplified RNA using the Illumina RNA TruSeq Kit v2. Using a previously published editing analysis pipeline (62), this sequencing data was aligned to a reference genome using the software program, TopHat, and then compared to the gDNA of a similar fly strain to identify A to I editing sites (Fig. 2.1). Sites where the RNA and gDNA differed were filtered against known splice junctions and checked for read quality. The quality filter requires that an edited site occur within the middle two quadrants of a 50 bp sequencing read at least once. This protects against the possibility of false-positives due to sequencing error, which tends to be greater at the ends of sequencing reads. From this analysis, 324 potential A to I editing sites were identified. However, because we did not have gDNA sequencing data available for the exact fly strain from which PDF-cells were sorted, we filtered the 324 potential sites against a database of annotated SNPs. After filtering for known SNPs in the literature, there were 30 potential editing sites remaining. To more accurately gauge the level of editing at these edited sites, and verify that they were indeed bona fide editing events and not SNPs, we made 23 PCR primers spanning a region containing the potential sites. Of these 23 sites, 18 had previously been identified as edited in heads (62). Additionally we made PCR primers across 8 of the annotated SNP sites to confirm if they were SNPs or editing sites in our fly strains.
Using RNA collected from two time-points of PDF-cells, as well as dopaminergic cells collected from a TH-Gal4>UAS-eGFP fly line, we made cDNA from dT-amplified RNA (cDNA#2 from amplification described in Chapter One). This cDNA, along with gDNA collected from each relevant strain, was used in PCR reactions for the 31 sites. The amplicons from each cell sample (PDF ZT2, PDF ZT14, TH ZT2, TH ZT14) and gDNA (PDF-Gal4>UAS-mcD8GFP, TH-Gal4>UAS-eGFP) were made into libraries and sequenced. The sequencing data was aligned to the genome using TopHat and then analyzed for specific sites using in-house scripts (see Materials and Methods). To be considered for quantification, a potential editing site in a sample must have been covered by at least 9 reads. Most sites were covered by at least 100 or more reads. However, of the 31 PCR reactions each performed in samples TH ZT2 and TH ZT14, 3 did not produce sufficient reads to be included in the analysis. Additionally, to be considered for cycling analysis, both time-points from a cell line were required to have adequate coverage. For sites that did not have sufficient coverage in a particular sample, the quantification of editing is considered null and left blank in figures and tables. For gDNA amplicons, the reads over the 31 sites of interest were quantified and a base call was made for each site. If the proportion of a nucleotide base at a particular site was >99%, the identity of that site was determined to be that base. If no one base at a site of interest was represented in the gDNA >99%, then the site was considered a SNP.

Overall, of the 31 potential sites for which PCR verification was done, 18 had been previously characterized in fly heads (62). This head data was analyzed from two replicates of six time-points. Because cycling editing was not found in heads,
the samples were pooled across time. There was sufficient data to characterize all 18 of these sites in PDF-cells, and 15 of the sites in TH-cells. A site was considered edited if it was coded as an A in the genome and showed levels of I >5% in the RNA. Surprisingly, one of these 18 sites, *Frq1* 18064159 (genome coordinate position), was not found to be edited in either cell group, though it was previously identified as being edited 77% of the time in heads. Of the remaining sites that were edited in at least one cell group, the editing levels were quite varied across cell types (Fig. 2.3 and supplementary Table1). Considering differences in editing levels, there were 6 sites with at least a 1.75 fold difference in editing level between PDF-cells and heads (Fig. 2.3, green bracket), 9 sites when comparing TH-cells and heads (Fig. 2.3, blue bracket), and 5 sites when comparing PDF- and TH-cells (Fig. 2.3, red bracket).

Additionally, one site, nAcRalpha-34E 14089236, was only edited in PDF-cells (editing level 41%) and not TH-cells (1%).

Of the remaining sites, 5 were potentially novel editing sites and 8 were annotated SNPs. In PDF-cells, all 5 of the potential novel sites and 8 annotated SNP sites turned out to be encoded as G or A/G SNPs in the genome. Therefore, these are not edited sites in PDF-cells. In TH-cells, the 5 potentially novel sites were either unedited or encoded as guanines/SNPs in the genome. However, in TH-cells, 3 of the 8 annotated SNP sites were actually encoded as A in the genome and were highly edited to I in the RNA (Fig. 2.4). The 3 sites have not been previously described as edited (61, 62). In the PDF-cell data set, these 3 sites were hard-coded in the genome as guanines, suggesting there may be some evolutionary advantage to having a G/I at these sites.
Because other editing analyses have shown that editing is abundant in the nervous system, we decided to investigate the functions of our edited genes to see if they were involved in processes specific to neurons. We analyzed the ontology of the 14 edited genes, encompassing the 20 editing sites investigated, using the online Functional Annotation Tool from the DAVID Bioinformatics Resource. As expected, the majority of genes were involved in functions specific to the nervous system, such as ion channel activity and synaptic transmission (Fig. 2.5). Additionally, there were 3 genes involved in cytoskeleton organization, and 2 genes involved in the regulation of gene expression.

We were not only interested in identifying cell specific editing sites and changes, but also site that might be edited under circadian regulation. To identify these sites we looked for cycling editing levels between ZT2 and ZT14 in the PCR data. To qualify as cycling, there must have been sufficient data (>9 reads) at both time-points, and the fold change between time-points must have been at least 1.75. Using these thresholds, 4 sites were found to be cycling specifically in PDF-cells (Fig. 2.6a), 10 in TH-cells (Fig. 2.6b), and 2 in both cell sets (Fig. 2.6c). Of the total of 6 cycling sites in PDF-cells, 5 had peak editing at ZT2. For TH-cells, 9 of the 12 cycling sites were most edited at ZT14. The 2 sites that cycled in both sets of cells, cycled in opposite phase. The presence of cycling at these sites indicates that the circadian system might regulate editing in some fashion. However, additional experiments across more time-points should be done to verify cycling and rule out the possibility that the cycling is due to daily environmental cues.
Finally, we analyzed the enrichment in expression for our edited genes, wanting to further elucidate how the two cell populations regulate these genes and their mRNA transcripts. Differential gene expression is one way for cell populations to regulate the activity of gene products and could be used in concert with RNA-editing to achieve very specific effects. Using RNA-seq data from the two cell types and from heads (RNA extracted and dT-amplified as was done for cells), the gene enrichment levels for the 14 genes, encompassing the 20 edited sites, were determined using Cufflinks (supplementary Table 1). Enrichment levels were determined by pooling the samples across time (PDF-cells: 2 replicates of ZT2 and ZT14. TH-cells: ZT2, and two replicates of ZT14. Heads: 1 replicate of ZT2 and ZT14), and dividing the gene signal from cells by the signal from heads. None of the genes were specifically enriched in PDF-cells, aside from Brd8, which had nearly 80-fold enrichment. In fact, only 7 of the genes had expression levels >5 FPKM (Fragments Per Kilobase transcript per Million mapped reads) in PDF-cells. Brd8 was also enriched in TH-cells at a level of 13-fold. In TH-cells, a total of 8 genes were enriched 2-fold or higher compared to whole heads. Of these 8, nAcRalpha-34E, a nicotinic acetylcholine receptor subunit, was enriched 25-fold. Interestingly, this is the same gene for which PDF- and TH- cells had significant differences in editing levels at 3 of the 5 sites characterized by PCR. At those 3 sites in nAcRalpha-34E, there was almost no editing in TH-cells. This suggests that there is not a correlation between high expression and high levels of editing. The examination of nAcRalpha-34E would actually suggest the inverse relationship. One explanation for this could be that highly edited transcripts are more likely to be targeted for
degradation. Or, it could be that PDF-cells utilize RNA-editing to create a more responsive receptor as a means to make up for low expression.
2.3 DISCUSSION

Previous studies have identified numerous editing sites within *Drosophila* (61, 62). However, the nature of editing in specific cell types has never been extensively studied. While earlier research has indicated that there are tissue-specific differences in editing levels for some sites (58, 65), a large-scale analysis of numerous sites has not been completed. Therefore, we sought to analyze editing and identify novel sites in two cell populations, PDF- and TH-cells, using RNA-seq and subsequent PCR verification. Additionally, because ADAR mutants have an altered circadian phenotype (58), we analyzed edited sites for cycling. Experiments using whole fly heads, did not identify cycling editing, but tissue heterogeneity could be masking circadian cycling in discrete neuronal subtypes (62). Altogether we used site-specific PCR to investigate 20 editing sites in PDF- and TH-cells. While we were unable to identify any novel sites specific to the cell types, we found 3 sites that were novel and specific to the strain from which TH-cells were collected. In addition, we found differential editing levels across the cell populations and cycling over time for some sites.

Many of the edited sites showed differential levels of editing across PDF- and TH-cells, as well as heads. It is possible that this is due to differences in expression of ADAR in these cells. It was previously found that levels of ADAR are varied across the brain (58). Expression of ADAR is ~10-fold higher in TH-cells versus PDF-cells according to RNA-seq data. However, this increased expression in TH-cells did not result in increased editing across sites. In fact, many sites were edited at lower
levels in TH-cells, suggesting that varying levels of ADAR in the two cell types is not sufficient to explain the editing differences.

The varied levels of editing are not surprising considering that these cells serve different functions within the brain. PDF-cells (sLNvs and lLNvs) are part of the circadian system, contributing to the morning peak in activity rhythms and mediating responses to light cues (41, 42, 44, 45). TH-cells are involved in arousal, sleep, memory, and may be indirectly involved in the circadian system (48-50). One can imagine that these cell populations gain their functional properties through the differential regulation of gene and protein dynamics. This can be accomplished by expressing particular genes, changing expression with regard to time of day, and/or differentially altering gene transcripts through alternative splicing and RNA-editing.

In particular, \textit{nAcRalpha-34E}, a nicotinic-acetylcholine receptor subunit, was edited differently between cell types at 3 of the 5 sites investigated. At these 3 sites, editing levels were higher in PDF-cells. One site, \textit{nAcRalpha-34E} 14089236 was not edited above threshold levels in TH-cells. Additionally, the expression levels of this gene are different across the cell types. \textit{nAcRalpha-34E} is hardly expressed in PDF-cells, while in TH-cells, it is expressed at low levels, but has a 25-fold enrichment over gene signal from heads. These expression and editing differences suggest that this particular receptor has varied properties between these cell populations.

\textit{nAcRalpha-34E}, also known as Dalpha5, is part of the nicotinic acetylcholine receptor (nAChR) family. Acetylcholine is the primary excitatory neurotransmitter within the fly CNS and nAChRs are responsible for mediating this activity (66).
nAChRs are ligand-gated cation channels that are assembled from 5 subunits (Fig. 2.7), which can be either alpha or beta. These subunits all have an extracellular ligand-binding domain, 4 transmembrane helices (M1-M4), and an intracellular loop between M3 and M4 (66, 67). The M2 from each subunit lines the channel pore, while M1 and M4 are located near subunit interfaces, and M3 adjacent to the cell membrane. The intracellular loop is believed to be involved in receptor assembly and contains several phosphorylation sites that are important for receptor regulation (66).

Within nAcRalpha-34E, 10 editing sites have been identified. Of the 5 sites we investigated, 2 were located within M3 and the remaining were in M4 (Fig. 2.8). Altogether, 3 of the edited sites, 1 in M3 and 2 in M4, result in amino acid changes. The two sites resulting in amino acid changes in M4 (Thr->Ala, Ile->Val) were edited at comparable levels in both cell types, while the site in M3 (Ile->Val) was 18-fold more edited in PDF-cells. Considering that M4 is located near the interface between subunits, editing in this segment could impact subunit assembly. It could also influence the formation of secondary structure. Additionally, M4 has been implicated in channel gating (68). As for the M3 site that results in an amino acid change, it may impact secondary structure as well. Substitution mutations in Torpedo californica nAChR M3 have also been shown to increase the level of current response of the channel (69). In light of this, it could be possible that to make up for low expression of this gene, PDF-cells utilize the editing site in M3 to increase the total responsiveness of the receptor.
Because PDF- and TH-cells are tied to the circadian system and ADAR deficient flies show alterations in circadian behavior (58), we analyzed the levels of editing in the two cells types across time at ZT2 and ZT14. Previous work analyzing editing in heads across 6 timepoints did not identify any sites that cycled (62). However, it is possible that the tissue heterogeneity of the head masked cycling that occurs in neuronal populations where circadian time is relevant. In our data, a number of sites were found to cycle in either cell type and 2 sites cycled in both populations. Though there was cycling in editing at these sites, none of the genes within which the sites reside cycle in expression in either cell type, aside from sk, which cycled in PDF cells (15-fold higher at ZT14). In the absence of cycling gene expression, cycling editing could provide the means to achieve temporal regulation over protein dynamics. Interestingly, the editing site in sk, sk5290552, was found to have cycling editing levels in TH-cells (24-fold higher at ZT14). The editing at this site occurs in an exon and results a Tyr->Cys change in amino acid. This amino acid change could be relevant to the temporal function of the SK cation channel. It appears that TH-cells regulate this SK functionality with cycling editing, while PDF-cells accomplish the same task by editing this transcript at constant levels and instead changing the expression level of the gene. Considering the two cycling sites found in both populations, the editing levels cycled in opposite phase in PDF- and TH-cells. This difference of phase could be due to differences in temporal function and activity of these two cell types.

It is not surprising that editing would cycle in PDF-cells, which express all of the core clock proteins, are essential for proper circadian motor rhythms, and have
previously been shown to have numerous genes that cycle in expression (28).
However, it is not immediately clear how TH-cells, which are not considered a part
of the ~150 clock neurons, achieve cycling gene expression or cycling editing.
According to our RNA-seq data, TH-cells do express some clock genes. TH-cells
express cyc and pdp1 at low levels, and show cycling levels of tim, per, and vri
eexpression. Though TH-cells do not show clk expression at detectable levels, it may
still be present in TH-cells. It has been shown that clk may indirectly regulate
dopamine levels and control the cycling expression of TH (49). Even in the absence
of clk, it is possible that cycling tim and per are enough to achieve cycling expression
of other genes, and that the clock in dopaminergic cells is maintained through
mechanisms other than the classical circadian transcriptional feedback loop. For
instance, it was shown that glial cells expressing cycling per, and known to mediate
circadian motor rhythms through Ebony, were anatomically close to dopaminergic
neurons. It was proposed that these glia might regulate dopaminergic signalling in a
circadian fashion through circuit mechanisms (70). Thus, it seems possible that TH-
cells could achieve full clock functionality through cycling tim and per, as well as
input from other clock circuitry. However, it is not certain that the rhythmic editing
in either of the cell types we investigated is a result of circadian regulation. It could
be due to environmental cues, like light, that occur cyclically over 24-hours. Further
experiments involving additional time-points and the use of clock mutants will need
to be done to identify the source of cycling at these edited sites.

In addition to the cycling and editing level differences that we discovered in
the PDF- and TH-cell data, we also found 3 novel editing sites. These 3 sites were
located in the \textit{nct}, \textit{Pkc98E}, and \textit{Brd8} gene transcripts. The \textit{nct} and \textit{Brd8} sites are within exons and result in Thr->Ala amino acid changes, while the \textit{Pkc98E} site is within the 3’-UTR. The sites represent strain differences between the flies used for PDF- and TH-cell collection because the gDNA nucleotides encoding these sites were guanines in the PDF-Gal4>UAS-mcD8GFP fly line and adenines in the TH-Gal4>UAS-eGFP line. The fact that these sites were encoded as G in one fly strain, and edited from A to I (which is translated as G) in another fly indicate that there is some evolutionary advantage to having a G, and/or the resulting amino acid, at these sites. The evolutionary pressure for a site to be either hard coded or edited has also been identified for editing sites in other organisms. For example, in the GABA\textsubscript{A} receptor subunit alpha3 in mouse brain, an editing site resulting in a Ile->Met amino acid change was found to be hard-coded as a methione in frogs and pufferfish (71).

For the \textit{nct} and \textit{Brd8} sites found in the TH-cell data, the editing may modify protein function because they are nonsynonymous substitutions that change a polar amino acid to a hydrophobic one. The \textit{Brd8} site in particular might be important for neuron function as indicated by the 80- and 13-fold enrichment over head signal found for PDF- and TH-cells, respectively. \textit{Brd8} function remains unknown, but is suspected to be involved in the negative regulation of gene expression and contains a bromodomain within its protein structure (FlyBase: FBgn0039654). Bromodomains are found in many DNA-binding proteins and may mediate transcriptional activity and protein-protein interactions (72). The edited site lies outside the bromodomain (Fig. 2.9) near the C-terminal of the protein. Thus, it is difficult to say what functional consequence it could have.
Overall, we determined the editing levels of 20 sites within the mRNA of PDF- and TH-cells. Many of these sites showed differential levels of editing and cycling within the specific cell populations. This suggests that editing is a method by which tissues specify and regulate their functions to achieve their biological purposes within living systems. Additionally, the identification of sites that potentially cycle indicates that the circadian system may employ RNA-editing to further modify and maintain biological rhythms. Finally, we identified 3 novel editing sites that were encoded differently across fly strains. These sites indicate that biology can often achieve the same task in multiple ways, hard-coding the identity of proteins in the genome or modifying them at the mRNA level.
2.4 FIGURES

Fig 2.1 Bioinformatic Pipeline for Editing Analysis.

Fig 2.2 PDF-Expressing Neurons. A-B) PDF-expressing neurons in adult brain in blue. Arrows point to projections (A) and cell bodies (B). Images courtesy of R. Veggeberg, Rosbash Lab.
Fig 2.3 Editing Levels in Cells of Previously Identified Sites. Editing levels were determined by using site-specific PCR and are reported as the proportion of sequencing reads that were inosines. Sites within the red bracket showed at least a 1.75 fold difference in editing level between PDF- and TH-cells. Blue bracket, sites that were at least 1.75 fold different between TH-cells and heads (nascent or mature mRNA). Green bracket, sites that were at least 1.75 fold different between PDF-cells and heads. Editing levels in heads was published in Rodriguez et al. Mol Cell. 2012.

Fig 2.4 Strain Specific Editing. These three sites were edited A to I in the strain from which TH-cells were collected.
Ontology of Edited Genes

- **Ion Channel Activity**
  - Ca-alpha1D (voltage-gated Ca2+ channel)
  - nAcRalpha534E (ligand-gated cation channel)
  - nAcRbeta64B (ligand-gated cation channel)
  - slo (voltage-gated/Ca-activated K+ channel)
  - SK (Ca-activated cation channel)
  - stj (voltage-gated Ca2+ channel)

- **Synaptic Transmission**
  - Syn (neurotransmitter transport)
  - shi (synaptic vesicle endocytosis)

- **Cytoskeleton Organization**
  - CG42540 (structural component)
  - Msp5300 (actin binding)

- **Regulation of Gene Expression**
  - Brd8 (negative regulation)
  - pan (transcription activator/repressor)

- **Protein Phosphorylation**
  - Pkc98E (Ca-dependent kinase)

- **Signal Transduction**
  - nct (Notch signaling pathway)

Fig 2.5 Edited Gene Ontology. Gene ontology was analyzed for edited genes using the Functional Annotation Tool from DAVID Bioinformatics Resources 6.7, NIAID/NIH.

Fig 2.6 Cycling Editing Levels in Cells. A) Editing sites that showed cycling between ZT2 and ZT14 in only PDF cells. B) Sites that cycled in only TH-cells. C) Sites that cycled in both cell types. Note: Sites were considered cycling if the fold change between time-points was at least 1.75 fold.
Fig 2.7 Schematic of Insect nAChR. A) Bird's eye view of assembled receptor with 5 subunits. Transmembrane regions, M1-M4, and ligand binding domains are indicated. B) Subunit domain structure: N-terminal extracellular ligand-binding domain, 4 transmembrane domains, and one intracellular loop. Note: Figure adapted from Dupuis et al. Neuroscience and Biobehavioral Reviews. 2012.

Test sequence and protein translation in E. coli.

Fig 2.8 Editing nAChR alpha034E. The coding and protein sequence is shown for residues 451-808. Editing sites in the coding sequence are highlighted in yellow and the changed adenosine is marked in red. If the edit resulted in an amino acid change, the amino acid is colored gray in the protein sequence. Blue indicates an extracellular domain. Purple corresponds to the transmembrane domains (M1-M4). Green indicates an intracellular domain. Coding sequence information was obtained from the NCBI and aligned to the protein translation using the translator tool at f33.net (Pawłowski, N.). Protein domain information was identified using the TMHMM Prediction Server from the CBS, Technical University of Denmark.
Fig 2.9 Editing in Brd8. Coding and protein sequence are shown for residues 661-874. Editing site are highlighted in yellow in the coding sequence with the changed nucleotide in red. The bromodomain is indicated in purple in the protein sequence. Coding sequence information was obtained from the NCBI and aligned to the protein translation using the translator tool a fr33.net (Pawlowski, N.). Protein sequence domains were identified using InterPro from the European Bioinformatics Institute.
CONCLUSIONS

We set out to optimize methods for creating RNA-seq libraries from small populations of cells, so as to be able to better characterize the neurons involved in the maintenance of circadian rhythms in Drosophila. Using dT-amplification we were able to create 3’-biased libraries from cells and use the sequencing data to analyze RNA-editing in two populations of cells, PDF- and TH-cells. In this manner we were able to identify 20 editing sites, 17 previously found in heads and 3 that were novel. Additionally, we found that many of the sites were edited at varying levels and cycled in the two cell populations. This suggests that tissue-specific regulation of editing does occur, and that some of this regulation may be mediated by the circadian system.

Our data was limited, though, by the 3’-bias of our libraries. Due to this bias, we were only able to identify a limited number of editing sites because of low genome coverage. Moreover, the bias confounds analyses of gene enrichment and isoform expression because the current methods used to quantify sequencing data normalize gene signal to the size of the gene. Thus, long genes will possibly be underrepresented because the limited signal coming from the 3’-end in the biased libraries will be divided by a larger gene size.

To reduce the 3’-bias that results from dT-amplification, we tried methods that used a mix of both random- and dT-primers. Initial experiments using this
method were unsuccessful in cells. However, lab members continued to experiment with different amplification conditions and library preparations, and were just recently successful in creating unbiased libraries from cells that mapped well to the genome. We used this working method to sequence RNA from six time-points of PDF-cells and are in the process of analyzing these data. With these data, we hope to better elucidate cycling gene and isoform expression, as well as identify tissue-specific editing events.

Future work will apply these amplification and RNA-seq methods to additional cell populations. Recently, we were able to successfully purify evening cells (LNds) from fly brains and create RNA-seq libraries at one time-point. The expression profile of these cells has never been characterized. Therefore, RNA-seq data could be very useful in illuminating how these cells functions in the circadian circuit, e.g. elucidate neuropeptides and receptors used in signaling, and identify novel circadian genes.
**Appendix**

### A.1 Supplementary Tables

<table>
<thead>
<tr>
<th>symbol</th>
<th>Chr</th>
<th>Coordinate</th>
<th>PDF GFP Strain gDNA</th>
<th>TH GFP Strain gDNA</th>
<th>Pooled PDF-Cell Editing</th>
<th>Pooled TH Cell Editing</th>
<th>Nascent Editing in Heads</th>
<th>Editing in Heads</th>
<th>PDF-Cell Gene Enrichment</th>
<th>TH-Cell Gene Enrichment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Msp-300</td>
<td>chr2L</td>
<td>5164694</td>
<td>A</td>
<td>A</td>
<td>0.13</td>
<td>0.12</td>
<td>0.29</td>
<td>0.62</td>
<td>0.67</td>
<td></td>
</tr>
<tr>
<td>nAcRalpha-34E</td>
<td>chr2L</td>
<td>14083516</td>
<td>A</td>
<td>A</td>
<td>0.87</td>
<td>0.08</td>
<td>0.84</td>
<td>0.73</td>
<td>0.00</td>
<td>25.88</td>
</tr>
<tr>
<td>nAcRalpha-34E</td>
<td>chr2L</td>
<td>14083517</td>
<td>A</td>
<td>A</td>
<td>0.88</td>
<td>0.05</td>
<td>0.84</td>
<td>0.81</td>
<td>0.00</td>
<td>25.88</td>
</tr>
<tr>
<td>nAcRalpha-34E</td>
<td>chr2L</td>
<td>14089204</td>
<td>A</td>
<td>A</td>
<td>0.65</td>
<td>0.96</td>
<td>0.83</td>
<td>0.88</td>
<td>0.00</td>
<td>25.88</td>
</tr>
<tr>
<td>nAcRalpha-34E</td>
<td>chr2L</td>
<td>14089236</td>
<td>A</td>
<td>A</td>
<td>0.41</td>
<td>0.01</td>
<td>0.46</td>
<td>0.00</td>
<td>0.00</td>
<td>25.88</td>
</tr>
<tr>
<td>Ca-alpha1D</td>
<td>chr2L</td>
<td>16174153</td>
<td>A</td>
<td>A</td>
<td>0.83</td>
<td>0.58</td>
<td>0.75</td>
<td>1.00</td>
<td>0.12</td>
<td>2.87</td>
</tr>
<tr>
<td>Ca-alpha1D</td>
<td>chr2L</td>
<td>16174194</td>
<td>A</td>
<td>A</td>
<td>0.51</td>
<td>0.95</td>
<td>1.00</td>
<td>0.12</td>
<td>1.2</td>
<td>2.87</td>
</tr>
<tr>
<td>stj</td>
<td>chr2R</td>
<td>9697052</td>
<td>A</td>
<td>A</td>
<td>0.29</td>
<td>0.42</td>
<td>0.68</td>
<td>0.84</td>
<td>0.09</td>
<td>2.12</td>
</tr>
<tr>
<td>nAcRbeta-64B</td>
<td>chr3L</td>
<td>4429963</td>
<td>A</td>
<td>A</td>
<td>0.96</td>
<td>0.81</td>
<td>0.90</td>
<td>0.14</td>
<td>3.95</td>
<td></td>
</tr>
<tr>
<td>CG42540</td>
<td>chr3L</td>
<td>4590709</td>
<td>A</td>
<td>A</td>
<td>0.15</td>
<td>0.15</td>
<td>0.44</td>
<td>0.06</td>
<td>2.63</td>
<td></td>
</tr>
<tr>
<td>Syn</td>
<td>chr3R</td>
<td>6021148</td>
<td>A</td>
<td>A</td>
<td>0.54</td>
<td>0.95</td>
<td>0.94</td>
<td>0.04</td>
<td>2.38</td>
<td></td>
</tr>
<tr>
<td>slo</td>
<td>chr3R</td>
<td>20501385</td>
<td>A</td>
<td>A</td>
<td>0.83</td>
<td>0.71</td>
<td>0.83</td>
<td>0.71</td>
<td>0.21</td>
<td>0.98</td>
</tr>
<tr>
<td>nct</td>
<td>chr3R</td>
<td>20679633</td>
<td>G</td>
<td>A</td>
<td>0.48</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pkc98E</td>
<td>chr3R</td>
<td>24872398</td>
<td>G</td>
<td>A</td>
<td>0.50</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Brd8</td>
<td>chr3R</td>
<td>25051833</td>
<td>G</td>
<td>A</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pan</td>
<td>chr4</td>
<td>120661</td>
<td>A</td>
<td>A</td>
<td>0.15</td>
<td>0.25</td>
<td>0.45</td>
<td>0.05</td>
<td>1.48</td>
<td></td>
</tr>
<tr>
<td>pan</td>
<td>chr4</td>
<td>120807</td>
<td>A</td>
<td>A</td>
<td>0.77</td>
<td>0.43</td>
<td>0.76</td>
<td>1.00</td>
<td>0.05</td>
<td>1.48</td>
</tr>
<tr>
<td>SK</td>
<td>chrX</td>
<td>5290552</td>
<td>A</td>
<td>A</td>
<td>0.41</td>
<td>0.12</td>
<td>0.32</td>
<td>0.38</td>
<td>0.32</td>
<td>1.56</td>
</tr>
<tr>
<td>shi</td>
<td>chrX</td>
<td>15791327</td>
<td>A</td>
<td>A</td>
<td>0.28</td>
<td>0.21</td>
<td>0.39</td>
<td>0.40</td>
<td>0.25</td>
<td>3.07</td>
</tr>
</tbody>
</table>

*Table 1* Editing Sites Across Cell Types. Editing sites and levels were validated using site-specific PCR. Enrichment levels were determined using RNA-seq data from both cell types and whole heads. Enrichment was calculated by dividing the gene signal from cells by the signal from heads. Data from heads was previously published in Rodriguez et al. Mol Cell. 2012.
A.2 METHODS AND MATERIALS

• FLY STRAINS

The following GFP-expressing lines were used for cell collections. 1) yw, UAS-mcD8GFP; Pdf-Gal4. 2) w; UAS-eGFP; Elav-Gal4. 3) w-; Th-Gal4/UAS-eGFP.

• CELL SORTING

Flies expressing GFP in the cells of interest were entrained to a 24-hour light/dark cycle in incubators at 22-25 degrees Celsius for 3-4 days. At ZT2 and ZT14, flies were collected for cell sorting. Cell sorting was modified from previous methods (28, 29). Briefly, brains were dissected in a cold saline solution and then transferred to a 1.5 ml tube of cell medium on ice. For PDF-cell, ELAV-cell, and TH-cell collections, ~50-60, 10-15, and 20-30 brains were dissected, respectively. The brains were then spun down and washed once with saline. Papain was added to each sample and the brains were dissociated for 25-30 minutes at room temperature. Approximately 3 volumes of cold cell medium was added to stop the dissociation. The samples were then spun down and washed once with cell medium. After resuspending in 300 ul of medium, the brains with triturated with flame-rounded 1 ml pipette tips. The resulting cell suspension was then plated on sylgard plates and allowed to settle for 20 minutes. GFP-positive cells were then sorted a total of 3 rounds (transferring to new plates) using a fluorescent microscope. Cells were then lysed with either XB lysis buffer (Arcturus PicoPure Kit
#0204) at 42 degrees, or with DynaBead lysis/binding buffer (DynaBead mRNA Direct MicroKit, Ambion 61021). The cell lysates were stored at -80 degrees.

**RNA EXTRACTION AND AMPLIFICATION FROM CELLS**

RNA from cell lysates was extracted with either the Arcturus PicoPure Kit (whole RNA) or DynaBead mRNA MicroKit (oligo-dT selection). Extractions were performed according to recommended kit protocols. RNA amplification was then carried out according to methods modified from (28, 29). dT-T7 primers were added to isolated RNA and the samples were evaporated to a volume of 2 ul. The first-strand synthesis of cDNA#1 was then carried out using reverse transcriptase, Superscript III (invitrogen). Second strand synthesis using DNA polymerase then followed. The double-stranded cDNA#1 was then ethanol precipitated at -80 degrees for at least 3 hours. cDNA#1 was pelleted, washed, and resuspended in 4 ul of water and immediately used in the first-round IVT reaction (carried out according to Ambion MEGAscript Kit protocol). The IVT reaction was incubated at 37 degrees for 16-20 hours. The resulting cRNA was isolated using Qiagen's RNAeasy minElute Kit. The 14 ul of eluted cRNA was then used in a second round of cDNA synthesis (cDNA#2).Briefly, random primer was added to the cRNA and the samples were evaporated to 5 ul. The samples were then reverse transcribed to cDNA using Superscript III for first-strand synthesis, and DNA polymerase for second-strand synthesis. cDNA#2 was ethanol precipitated as before, and then used in a second round of IVT (cRNA#2). cRNA#2 was used to make RNA-seq libraries.
• CREATION OF RNA-SEQ LIBRARIES

RNA-seq libraries from cells were created using cRNA#2 from the amplification protocol. Libraries were created according to standard protocols for the Illumina RNA TruSeq Kit v2, entering the protocol at the fragmentation stage.

• RNA EXTRACTION FROM FLY HEADS

Whole RNA from the yw, UAS-mcD8GFP; Pdf-Gal4 strain was collected at ZT2 and ZT14 from heads using Trizol reagent from Invitrogen and the manufacturer's recommended protocol. The whole RNA was then diluted to levels comparable to those of extracted cells and amplified according the protocol described above. Libraries were then made from the amplified head RNA.

• GENOMIC DNA EXTRACTION AND LIBRARY CREATION

The gDNA from the Pdf-Gal4>UAS-mcD8GFP and Th-Gal4>UAS-eGFP fly lines was extracted according to the Qiagen DNeasy Blood and Tissue Kit. The gDNA was then sheared by sonication using the Diagenode Biorupter. Shearing was done at the medium setting for 40 rounds of 30 second on/off cycles. Libraries were then made from the sheared gDNA. Library preparation was done according standard Illumina DNA TruSeq protocols.

• QUANTITATIVE PCR
Q-PCR was performed to check the levels of amplification products, cDNA#1 and cDNA#2, from cells. PCR was carried out using Syber Master Mix (Qiagen), using protocols described previously (27).

• SITE-SPECIFIC PCR

PCR reactions for the 31 RNA-editing sites of interest were carried out using the gDNA from the PDF- and TH-cell lines, as well as cDNA#2 from both time-points of each cell type. Reactions were done using Platinum Taq Polymerase Hi Fidelity (Invitrogen) according to manufacturer's recommended protocol. The reactions for each sample template were pooled and cleaned using Qiagen's PCR minElute Kit. The amplicons were then made into sequencing libraries according to Illumina's DNA TruSeq Kit.

• LIBRARY AND EDITING DATA ANALYSIS

Sequencing was done using Illumina's Genome Analyzer. The RNA-sequencing reads and PCR amplicons were mapped to the Drosophila genome using TopHat and a gene annotation file. The following options were used with TopHat: -m 1 -F 0 -p 6 -g 1 --microexon-search --no-closure-search -G --solexa-quals -l 50000. After mapping the expression signals were quantified using Cufflinks and the following options: -G -b --max-bundle-frags 10000000. Additionally, matrix and visualization files which quantify the number and types of reads for each base in the genome were made using previously published methods (62). For the gDNA
libraries, mapping was carried out with Bowtie2 using the options: --solexa-quals --sensitive -x.

Editing analysis of RNA-seq data was done using previously published methods (62). Analysis of the PCR amplicon sequencing for the 31 sites of interest was done by making matrix files and then using the unix GREP command to pull out the base reads covering the coordinate position of each site.
REFERENCES


