1-1-2014

The Application and Challenges of RNA-Sequencing to the Study of Circadian Rhythms

Nicholas Lahens
University of Pennsylvania, nif@mail.med.upenn.edu

The Application and Challenges of RNA-Sequencing to the Study of Circadian Rhythms

Abstract

The circadian clock drives daily rhythms in behavior and physiology, often in anticipation of the coming dusk or dawn. Almost all organisms possess an internal time-keeper, as it represents an adaptation to one of the most ancient selective pressures; the day-night cycle. Mounting evidence suggests the clock plays important roles in critical metabolic and signalling pathways, the sleep/wake cycle, immune function, as well as learning and memory. Perhaps more importantly, misregulation of the clock is associated with metabolic disorders, neurodegeneration, and incidence of cancer. In an effort to unlock the connections between the circadian clock and these downstream effects, researchers have searched for genes with rhytmic transcription driven by the clock. These so-called clock-controlled genes (CCGs) mediate these observed rhythms in important biological pathways.

Over the past decade, researchers have searched for these CCGs using microarrays. However, with the growing popularity of high-throughput sequencing, and revelations about both the number and importance of non-coding RNAs (ncRNAs), investigators have begun to use RNA-seq for their circadian profiles. While RNA-seq has led to important findings about the circadian regulation of RNA editing, small RNAs, and epigenetic modifications, there is still much about its biases and limitations that we are still discovering. To this end, this thesis seeks to build upon this foundation and examine the use of RNA-seq for studying circadian transcription. I applied a hybrid RNA-seq, microarray approach to assay the circading transcriptome in liver, and eleven other mouse tissues. Notably, I saw that 1/3rd of ncRNAs conserved between human and mouse show rhythmic transcription. These rhythmic transcripts are strong candidates for future functional validation, and include important miRNA and snoRNA precursors. Additionaly, I found hundreds novel ncRNAs with rhythmic expression, which may provide novel CCGs. Lastly, I developped and applied a method for identifying the sources of bias in RNA-seq protocols. Taken together, this work extends our understanding of the circadian transcriptome, and the challenges associated with interpreting RNA-seq data.

Degree Type

Dissertation

Degree Name

Doctor of Philosophy (PhD)

Graduate Group

Genomics & Computational Biology

First Advisor

John B. Hogenesch

Keywords

Circadian, Genomics, ncRNA, RNA-seq

Subject Categories
Bioinformatics | Biology | Molecular Biology

THE APPLICATION AND CHALLENGES OF RNA-SEQUENCING TO THE STUDY OF CIRCADIAN RHYTHMS

Nicholas F. Lahens

A DISSERTATION

in

Genomics and Computational Biology

Presented to the Faculties of the University of Pennsylvania

in

Partial Fulfillment of the Requirements for the

Degree of Doctor of Philosophy

2014

Supervisor of Dissertation

____

John B. Hogenesch, Ph.D., Professor of Pharmacology

Graduate Group Chairperson

____

Li-San Wang, Ph.D., Associate Professor of Pathology and Laboratory Medicine

Dissertation Committee

Russ P. Carstens, M.D., Associate Professor of Medicine

Brian Gregory, Ph.D., Assistant Professor of Biology

Junhyong Kim, Ph.D., Professor of Biology

Michael N. Nitabach, Ph.D., Associate Professor of Cellular and Molecular Physiology and of Genetics, Yale University

THE APPLICATION AND CHALLENGES OF RNA-SEQUENCING TO THE STUDY OF CIRCADIAN RHYTHMS

ABSTRACT

THE APPLICATION AND CHALLENGES OF RNA-SEQUENCING TO THE STUDY OF CIRCADIAN RHYTHMS

Nicholas F. Lahens

John B. Hogenesch

The circadian clock drives daily rhythms in behavior and physiology, often in anticipation of the coming dusk or dawn. Almost all organisms possess an internal time-keeper, as it represents an adaptation to one of the most ancient selective pressures; the day-night cycle. Mounting evidence suggests the clock plays important roles in critical metabolic and signalling pathways, the sleep/wake cycle, immune function, as well as learning and memory. Perhaps more importantly, misregulation of the clock is associated with metabolic disorders, neurodegeneration, and incidence of cancer. In an effort to unlock the connections between the circadian clock and these downstream effects, researchers have searched for genes with rhytmic transcription driven by the clock. These so-called clock-controlled genes (CCGs) mediate these observed rhythms in important biological pathways.

Over the past decade, researchers have searched for these CCGs using microarrays. However, with the growing popularity of high-throughput sequencing, and revelations about both the number and importance of non-coding RNAs (ncRNAs), investigators have begun to use RNA-seq for their circadian profiles. While RNA-seq has led to important findings about the circadian regulation of RNA editing, small RNAs, and epigenetic modifications, there is still much about its biases and limitations that we are still discovering. To this end, this thesis seeks to build upon this foundation and examine the use of RNA-seq for studying circadian transcription. I applied a hybrid RNA-seq, microarray approach to assay the circading transcriptome in liver, and eleven other mouse tissues. Notably, I saw that 1/3rd of ncRNAs conserved between human and mouse show rhythmic transcription. These rhythmic transcripts are strong candidates for future functional validation, and include important miRNA and snoRNA precursors. Additionaly, I found hundreds novel ncRNAs with rhythmic expression, which may provide novel CCGs. Lastly, I developped and applied a method for identifying the sources of bias in RNA-seq protocols. Taken together, this work extends our understanding of the circadian transcriptome, and the challenges associated with interpreting RNA-seq data.

Table of Contents

List of tables and figures

Tables

Figures

Chapter 1: Introduction

1.1 Background

The circadian clock

The day-night cycle represents one of the most ancient environmental stimuli under which life has evolved on Earth. As a result, nearly every organism on the planet maintains an internal time-keeper, known as the circadian clock (in Latin: *circa* = around/about, and *diem* = day). This internal clock oscillates with a period of roughly 24 hours, and allows organisms to anticipate the coming of dawn and dusk. A diverse set of behaviors, biological processes, and diseases are under circadian regulation and/or affected by circadian disruption. These include sleep, body temperature, blood pressure, memory, neurodegeneration, and metabolic disorder, just to name a few [1–8].

The molecular basis for these organism-level rhythms consists of a transcriptional/translational negative feedback loop. The transcriptional activators CLOCK/NPAS2 [9, 10] and BMAL1 form a heterodimer that binds E/E'-box DNA sequence motifs in gene promoters. This heterodimer activates the transcription of the circadian repressors *Per1*, *Per2*, *Per3*, *Cry1*, and *Cry2*. Following their translation, the PER and CRY proteins form a complex, and translocate back into the nucleus. Once in the nucleus, they repress their own transcription by inhibiting the activity of CLOCK and BMAL1. As PER and CRY levels drop due to targeted degradation, CLOCK and BMAL1 activity is restored, and the cycle begins again. This cycle of activation and repression takes roughly 24 hours to complete, providing the molecular mechanism for circadian time-keeping. A second, stabilizing loop modulates *Bmal1* and *Cry1* expression. This secondary loop consists of the transcriptional activators *Rora*, *Rorb*, and *Rorc*, and the transcriptional repressors *Rev-erbα* and *Rev-erbβ* [11–14]. For a more thorough review of the core clock, please consult the following papers [15, 16].

The core oscillator is present and active in most cells throughout the body [17–19]. However, the central time-keeper, or master oscillator, in the mammalian circadian system is contained in the suprachiasmatic nucleus (SCN) of the hypothalamus. Knocking out the molecular clock in the SCN, or damaging/removing the SCN itself disrupts behavioral rhythms [20, 21]. The SCN is entrained by light input received directly from the retina through melanopsin-positive ganglion cells [22, 23]. Most other tissues maintain local or peripheral oscillators. These peripheral oscillators are capable of sustained rhythms, but all are entrained by the master oscillator present in the SCN. This cascade of light signaling through the retina, to the SCN, and on to the peripheral oscillators, keeps the internal clocks throughout an organism running in phase with the environmental day-night cycle [24–26].

Clock-controlled genes

The core, molecular oscillator is able to drive rhythms in other cellular and physiological processes by regulating the expression of clock-output or clock-controlled genes (CCGs). In some cases, this regulation is direct. For instance, *Dbp*, *Tef*, and *Hlf* are direct targets of CLOCK/BMAL1 binding [27, 28]. Many of these direct targets regulate the expression of downstream CCGs [29]. The direct CLOCK/BMAL1 target *Dbp* binds

D-box promoter elements and can drive rhythmic expression in genes downstream of the core oscillator [30]. It is through this cascade of oscillating expression in direct and indirect CCGs, that the molecular clock is able to affect a wide array of biological processes. For example, rate-limiting enzymes in cholesterol biosynthesis (*Hmgcr*), bile acid synthesis (*Cyp7a1*), catecholamine synthesis (*Th*), and catecholamine degradation (*Maoa*), all have expression affected by the clock [31–35].

Finding clock-controlled genes

Given their rhythmic expression, and their role as mediators of the clock's effect on downstream biology, the research community has expended significant effort to identify CCGs across a wide range of tissues and model organisms [34, 36–54]. Since the turn of the century, the primary workhorse for these studies has been the DNA microarray. The genome scale data provided by microarrays has allowed researchers in the clock field to not only identify individual clock genes, but explore clock regulation at the level of gene networks, and draw comparisons between different organs. These studies lead to the finding that CCGs in the SCN are enriched in pathways for peptide synthesis, secretion, and redox state, while those in the liver regulate various metabolic pathways, and those in the heart are associated with G-protein-coupled receptor signaling [34, 36]. While a CCG may oscillate in a particular set of tissues, it is still unclear whether these rhythms are driven by the local organ clock, or the master clock in the SCN. To address this question, later studies used transgenic mice with disrupted *Clock* and *Bmal1* expression in the SCN or peripheral tissues [37, 38]. They found that while the majority of CCGs appear to be driven by local clocks, there is a small subset that may receive their cues from the master oscillator. Researchers also noted that by increasing sampling resolution of their circadian tissue collections (eg. collect RNA every 2 hours, instead of every 4 hours), they decreased noise in their ability to accurately detect oscillating genes [39]. Studies that applied this philosophy were able to find greater numbers of CCGs, as well as genes displaying sub-circadian rhythms (cyclers with 12-hour and 8-hour periods) [40]. These recent studies have also brought to light the need to carefully design experiments to identify CCGs, and to select analysis algorithms that complement this experimental design [55].

While their use led to these great advances in the circadian field, microarrays are not without their disadvantages. The first disadvantage is not specific to microarrays, but arises more from assumptions made when designing these CCG-focused experiments. Most CCGs are ultimately acting at the protein level. Researchers use RNA expression as a proxy for the protein quantity. While it would be ideal to look for rhythms in protein activity at the genome-level, our ability to detect proteins and their activity remains limited with current technology. At present, nucleic acids are simply a more tractable molecule to study at scale. Secondly, microarrays are limited by the composition of their probesets. In other words, an array can only assay transcripts for which it has matching probes. Manufacturers have sought to mitigate this shortcoming by increasing the number and diversity of probesets included in successive versions of their arrays. Alternatively, one could use a tiling array covering the entire genome, but this approach is not feasible for organisms with larger genomes (like humans and mice). Lastly, microarrays do not yield nucleotide-level information or splicing information. Some researchers and manufactures have developed specialized exon and splicing arrays to partially overcome this deficit. In fact, one study used exon arrays to identify a set of CCGs in the liver with clock-regulated splicing [41]. Despite these limitations, microarrays still offer an efficient and effective means to quantify RNA expression.

RNA-sequencing

Over the past six years, high-throughput sequencing of RNA (RNA-seq) has emerged as a potential successor to microarrays for studying transcriptomics. Though there are multiple implementations of the RNA-seq paradigm, all involve reading sequences from a massive number of RNA fragments at the resolution of individual base pairs. It is this massive number of sequences that differentiate RNA-seq from traditional Sanger sequencing. In the context of this dissertation, RNA-seq refers specifically Illumina's implementation, unless otherwise stated. Briefly, RNA-seq begins with the fragmentation of RNA into short fragments (100 ~ 1000 bp), which are subsequently converted to cDNA. Since most RNA samples consist of 90-99% ribosomal RNA, these first steps are commonly preceded by some form of polyA selection or rRNA-depletion. After cDNA generation, special adapters are ligated to the cDNA fragments. This in turn is most commonly followed by a PCR step to enrich the quantity of cDNA fragments with ligated adapters. These adapters will serve as primers during the sequencing reaction. Next, the cDNA fragments are immobilized on a flow cell and amplified in place through a bridge PCR reaction. This leaves clusters of cDNA fragments that are all copies of the original that was bound to the flowcell. The sequencing-by-synthesis reaction is carried out through the incorporation of special fluorescent nucleotides with reversible terminators, which prevent the addition of more than one nucleotide at a time. Next, the flow cell is scanned with a laser to determine which nucleotides were incorporated (each nucleotide has a different fluorescent color associated with it). Finally, the 3’ terminator and flurophore are removed and the colonies are ready for the addition of the next nucleotide. These steps of nucleotide incorporation, base imaging, and terminator cleavage are repeated to yield short sequence reads ranging from 35-300 bp in length. For paired-end sequencing, the sequencing reaction is repeated from the opposite end of the cDNA fragments. This yields two reads with a range (dependent upon the fragmentation kinetics) of possible distances between them, making them easier to map to a reference genome. For further details on this procedure, consult the following papers [56–59].

Analyzing RNA-seq data

After generating these short reads, the next stage in the process is to map them back to a reference genome. The sequencing machine produces a text file containing a list of nucleotide sequences. This information is relatively useless unless we can identify which transcripts these sequence reads originated from. This is the problem solved by the alignment step. At its core, the alignment step involves sorting through all possible genomic locations for the one that provides the best match for a given sequence. The alignment step must perform this task efficiently, as it needs to repeat this millions of times (once for each read). Early aligners, like Bowtie [60], used a method based upon the Burrows-Wheeler transform to both rapidly query the reference genome for possible alignments, and to shrink the reference genome down to a size small enough to fit into the main memory of most desktop computers. While this made the aligners quick and memory efficient, these aligners suffered from decreased accuracy [61]. They tended to have difficulty handling sequencing errors, SNPs, and gapped alignments, which commonly arise from reads mapping across splice junctions. These problems compounded as the length of reads has steadily increased from 35 bp. The current generation of aligners achieve high accuracy and speed by leaving the reference genome uncompressed [62–64]. This effectively trades off memory efficiency for speed; a human sized genome requires roughly 30 GB of RAM. However, by storing the entire reference in an uncompressed state, these aligners can quickly access the reference genome and find alignments, without suffering from the inaccuracies introduced by a memory transformation function.

Comparing RNA-seq to microarrays

RNA-seq offers several advantages over microarrays. The most substantial difference being that RNA-seq is not limited to specific probe regions. In theory, RNA-seq should be able to assay any transcript originating from anywhere in the genome, regardless of prior knowledge. For quantification, RNA-seq possesses a higher dynamic range than microarrays [65, 66]. RNA-seq provides a greater depth of information in the form of nucleotide-level data. This allows for the detection of RNA-editing, as well as genetic variants [48, 67]. Also, RNA-seq data can identify the splicing state of transcripts by searching for reads that span splice junctions (ie. two thirds of the reads map to one exon, and one third map to another exon) [68–70]. This, in theory, allows RNA-seq data to differentiate between different transcript isoforms. While RNA-seq offers several advantages of microarrays, it is not without its shortcomings.

Given the relative youth of RNA-seq, when compared to microarrays, there are substantial biases and challenges present when handling and analyzing RNA-seq data. These biases/challenges come in two forms: those introduced by the molecular biology of library preparation and sequencing, and those introduced during downstream bioinformatics analyses. The molecular biology involved in library preparation and sequencing itself can introduce biases that lead to over-/under-representation of particular transcripts or genomic regions. These can arise from GC-content biases, PCR artifacts, preferential adapter ligation to particular sequences, random-priming during reverse transcription, and errors introduced during the actual sequencing reaction [71–75]. Many of these artifacts are cause by the biases inherent to the enzymes that catalyze the various steps during library preparation and sequencing. These biases are mostly likely overcome by making changes to the protocol, like using different enzymes or devising methods for skipping particularly troublesome steps (like PCR enrichment). Alignment and analysis of RNA-seq data also presents several challenges. Importantly, RNA-seq data requires significantly greater computational resources, both in terms of storage and CPU speed, than microarrays [76]. Also, while it is theoretically possible to quantify different transcript isoforms, this is an extremely difficult problem to address given current read lengths. Current methods for quantifying isoforms produce significant numbers of false positives in the form of incorrectly assembled transcripts [77]. These biases introduced by biology and informatics compound to make it extremely difficult to accurately quantify transcripts and identify differentially expression genes [78]. Microarrays offer a comparatively simple and accurate analysis pipeline. They are a mature technology and we have come to understand their associated biases [79, 80]. As a result, bias analysis in RNA-seq remains a very active area of research.

Clock-controlled genes and RNA-seq

Despite these challenges and its recent adoption, RNA-seq offers great promise in the search for CCGs. Circadian studies using RNA-seq have found evidence of oscillations in RNA editing, as well as novel and non-coding transcripts [44, 48]. Other studies have used specialized RNA-seq protocols to differentiate between accumulated RNA transcripts, and nascent transcripts in the process of being actively transcribed [49, 50]. These studies found transcripts with rhythmic RNA accumulation, but no rhythms in transcription, suggesting these transcripts may owe their oscillations to post-transcriptional processes like RNA degradation. Researchers have also leveraged ChIP-seq experiments to find loci with coordinated rhythms in clock factor binding, transcripts, and epigenetic chromatin marks [42–47, 54]. These complementary approaches have added to previous findings from the microarray studies in the liver. Not only did they confirm oscillation in these metabolic genes, but they found their transcriptional rhythms appear to be driven by both BMAL1 and REV-ERBα binding. Additionally, these studies revealed the temporal sequence of chromatin events leading to rhythmic transcription. Finally, researchers have also begun to search for CCGs among non-coding RNAs (ncRNAs) [48, 51, 53]. These include snoRNA host genes, miRNAs, and novel transcripts. Many of these transcripts were missed by previous studies, since they were not included on any microarrays. While there is little functional information about many of these non-coding transcripts, there is emerging evidence of their importance for downstream processes like sleep, and their ability to feedback into the clock [52, 81].

1.2 Motivation and thesis outline

RNA-seq has experienced explosive popularity since its introduction. When this thesis project began in August of 2009, there were 46 datasets stored in GEO that featured the keyword "RNA-seq." As of June 2014, there are 6,286. Researchers across all fields have sought to leverage this technology, and the circadian field is no exception. In the search for oscillating transcripts, recent studies have used RNA-seq to add to the strong foundation already established by microarray data. As I mentioned in the background, this work has revealed a great deal more complexity in circadian transcriptional control, especially in the area of ncRNAs. However, many of these studies focused on single tissues. Furthermore, while RNA-seq has proven extremely powerful, it is a relatively new technology, and we are still coming to grips with its limitations and biases. The work presented in this thesis demonstrates the utility and challenges of using RNA-sequencing, with a focus on the circadian system and ncRNAs.

Chapter 2 describes a hybrid circadian expression profile which uses both microarrays and RNA-seq to identify oscillating transcripts. I demonstrate the utility of this combined approach and examine core clock gene splicing and ncRNAs present in the mouse liver. In chapter 3 I apply this technique to twelve different mouse tissues in order to study the circadian non-coding transcriptome. I identify oscillating transcripts conserved between humans and mice, as well as hundreds of putative lincRNAs and antisense transcripts. Chapter 4 describes a technique to assess the sources of coverage bias in sequencing protocols. Using this technique, I identify rRNA-depletion as a significant source of bias that has been previously unappreciated. In chapter 5, I discuss the future of this work and additional experiments which will expand upon my existing findings. Finally, in chapter 6 I conclude this thesis by summarizing my work and discussing its significant findings.

Chapter 2: A combined DNA array and RNA sequencing approach to profiling circadian transcription in the mouse liver

2.1 Abstract

The circadian clock regulates biological rhythms of ~ 24 hours in most organisms. The molecular clock is comprised of transcriptional regulators that drive rhythmic expression of key mediators of physiology and behavior. Here we apply a combined approach using high resolution temporal profiling by DNA arrays with lower resolution temporal profiling by RNA sequencing to profile clock regulated gene transcription in mouse liver. This hybrid approach allows us to leverage array data to identify oscillating transcripts with a high degree of accuracy, and then explore the structure and splicing patterns of these transcripts. Analysis of this data demonstrates the importance of sampling resolution when designing experiments to identify oscillating transcripts. Furthermore, we show that more than half of core clock factors express alternatively spliced forms concurrently in mouse liver. Interestingly, we find several forms of non-coding RNAs, including microRNAs and long non-coding RNAs, exhibit high amplitude circadian rhythms. These results provide a more complete picture of circadian transcriptional output and identify new clock-controlled genes.

2.2 Introduction

The circadian clock is a cell-autonomous molecular mechanism that drives daily rhythms in behavior, physiology, and metabolism [3–5]. Dysfunction in the clock has been linked to a wide array of diseases, including sleep and metabolic disorders [1, 2, 6]. Oscillations of the clock at the molecular level are driven by the interactions of two transcriptional/translational negative-feedback loops [82]. The main loop consists of the transcriptional activators *Clock*, *Bmal1,* and *Npas2* [9, 10, 83]. These transcriptional activators drive the transcription of, and are subsequently repressed by the repressors *Per1*, *Per2*, *Per3*, *Cry1*, and *Cry2* [16]. The secondary loop consisting of *Nr1d1*, *Nr1d2*, *Rora*, *Rorb,* and *Rorc*, further modulates the transcription of *Bmal1* and *Cry1* [11–13]. Rhythmic expression of these core clock genes in turn drives oscillations in the expression of their target genes, also known as clock controlled genes (CCGs). It is largely through these CCGs that the circadian clock is able to influence various biological pathways [27].

Since CCGs mediate the circadian clock's biological influence, researchers have spent a great deal of time and effort profiling the expression of oscillating transcripts [36–38, 84]. Over the past decade, DNA microarrays have been the primary tool used for this purpose. These arrays have proven extremely useful due to their low cost and well-established analysis methods [85]. Nevertheless, DNA microarrays can only assay a finite, defined set of loci and/or transcripts. Given the well-trodden state of this array-based approach to circadian expression profiling , circadian researchers would benefit from a new approach and a leap in technology to progress beyond what we can currently see, particularly to assess the rich diversity of ncRNAs in the transcriptome.

One such alternative is high-throughput sequencing of RNA/cDNA (RNA-seq). This technology provides several advantages over DNA microarrays, including the ability to profile virtually any transcribed region of the genome, and analyze alternative splicing by sequencing exon-exon junctions [57, 65, 68]. Researchers have successfully used RNA-seq for circadian profiling in several model organisms [44, 48, 53, 86].

Nevertheless, the monetary costs and investment of time required for RNA-Seq analyses – both for the bench top and computational work – is prohibitively expensive for many applications [76, 87, 88]. This is particularly relevant to studies of circadian transcription, which are exquisitely sensitive to sampling resolution and require large sample sizes [39]. Therefore, we have developed a hybrid approach that combines the advantages of DNA arrays with the un-paralleled sequence resolution of RNA-Seq. Using this method, we analyze circadian transcriptional rhythms in the mouse liver and demonstrate that alternative splicing generates extensive diversity among clock genes. Additionally, we sequenced liver RNA of $Clock^{\Delta 19}$ mutant mice and found that the loss of the molecular clock did not alter the splicing patterns of core clock genes. Finally, we identify many ncRNAs that oscillate with high-amplitude in the liver, and appear to be regulated by the clock. These results reveal a greater degree of diversity in the mammalian circadian liver transcriptome than previous array-based studies and broaden the list of CCGs to include new classes of ncRNAs.

2.3 Results

Sampling resolution in circadian experiments

We collected liver mRNA every six hours for two days (8 samples total) from wildtype animals, and every six hours for one day (4 samples total) from *Clock*$^{\Delta 19}$ mutant animals. This mutation results in the loss of exon 19 from the mature *Clock* transcript and produces a dominant-negative form of the CLOCK protein [89, 90]. The mutant CLOCK interacts with and inhibits the function of BMAL1, thereby eliminating oscillations of the core molecular clock and leading to behavioral arrhythmicity in constant conditions [38, 89]. We sequenced all of these mRNA samples using the Illumina GA IIx platform, yielding ~46 gigabases of sequencing data (see section 2.5 Methods). We successfully mapped 94% - 97% of the raw reads to the mouse genome (Table S2.1) using the RNA-seq Unified Mapper (RUM) [61]. Additionally, we sequenced mRNA from the livers of mice with the *Clock*$^{\Delta 19}$ mutation collected every six hours for one day (4 samples total). To identify genes with rhythmic transcription, we used quantification values for each transcript generated by RUM for the wildtype samples and analyzed them with JTK_CYCLE [91]. This analysis yielded 1166 genes with rhythmic transcription.

We compared these 1166 oscillating genes from our RNA-seq data with a previous DNA array study which identified over 3000 oscillating genes in mouse liver [40]. In practice, the RNA-seq data found at most 12% of the oscillating genes identified by the array study (Fig. 2.1A). Aside from the difference in technology (sequencing vs. arrays), the principal difference between these two experiments is the sampling resolution (every 6 hours vs. every hour). However, when we looked at *Bmal1, Per2,* and *Dbp,* three clock genes identified as oscillating by both the arrays and RNA-seq, we saw excellent agreement between the two data sets (Fig. 2.2B). These observations led us to examine the effects of sampling resolution on our ability to accurately identify oscillating genes.

To simulate different sampling resolutions we took the Hughes *et al*. data set [40] (one hour resolution), and sampled different subsets (ie. every other array = two hour resolution, every third array = three hour resolution, etc.). We then analyzed each of these smaller data sets using JTK_CYCLE, and compared the results to the one hour data set, which served as our gold standard. We saw a steep drop-off in our ability to identify oscillating genes in the gold standard as we increased the time between samplings (Fig. 2.2A; upper panel). Additionally, we saw a reciprocal increase in the number of false positives (genes incorrectly identified as oscillating when compared to the gold standard) as we increased the time between samples (Fig. 2.2A; lower panel). For example, the four hour resolution data correctly identified 26% of the gold standard at the lowest levels of statistical stringency, with a false positive rate of 24%. The statistical weakness of low sampling resolutions is particularly apparent when one day of data is analyzed rather than two (Fig. 2.2B). The four-hour resolution data from a single day correctly identifies 13% of the gold standard with a false positive rate of 72%. These data indicate that as the sampling resolution for a circadian time course experiment decreases, the ability to correctly identify oscillating genes drops drastically. Furthermore, this effect appears to be more extreme for data collected over a single day instead of two. Thusly, data spanning multiple days allow us to see the repeated expression patterns that are the hallmark of rhythmic genes. Based on these data, we recommend collecting data over two days at no less than two-hour resolution for the most accurate profile of circadian transcription.

These simulations with the Hughes *et al*. data explained the low concordance between the RNA-seq and array data sets. The six-hour resolution from the array simulation (same resolution as the RNA-seq experiment) has a 50% false positive rate and only correctly identifies 11% of the gold standard as oscillating (Fig. 2.2A). This is extremely close to the RNA-seq data, which has a false positive rate of 58% and correctly identifies 12% of the gold standard. Furthermore, comparing the overlap between the lists of cycling genes identified in the six-hour array data set and in the one-hour array data set is remarkably similar to the previous RNA-seq vs. arrays comparison (Fig. S2.1). These results suggest that the differences we saw are likely the result of sampling resolution, rather than a difference in technology.

Transcript diversity of clock genes

Next, to characterize splicing of rhythmically expressed genes we leveraged the structural data from RNA-seq with the high temporal resolution of the array data set. To this end we used the Hughes *et al*. array data, re-analyzed with JTK_CYCLE, as a first pass to identify oscillating genes. We performed this re-analysis because JTK_CYCLE has greater sensitivity and specificity than the algorithms originally used to identify oscillating genes in the array data [91]. We then used RNA-seq to examine the splicing patterns of these oscillating genes. Of the 4016 oscillating genes identified by the arrays (JTK q-value < 0.05), 2530 express more than one spliceform in our wildtype RNA-seq data (see section 2.5 Methods).

We are particularly interested in alternative splice forms of core clock genes. The majority of core clock genes (16 of 19) have multiple annotated spliceforms, with over half of those expressing multiple spliceforms in the liver (Table 2.1). Furthermore, our *Clock*$^{\Delta 19}$ RNA-seq data showed little change in the spliceform usage of core clock genes. While we detected fewer spliceforms for *Per3* and *Csnk1a1* in the *Clock*$^{\Delta 19}$ samples than in the wildtype samples, this is likely due to the loss of amplitude in cycling genes shown in a previous study [38]. The expression levels of these alternative transcripts may simply be too low to detect. This evidence for expression comes in the form of RNA-seq reads which map across exon-exon junctions. By examining which junctions are used, we determined which transcripts are expressed. We focused subsequent analyses on three key circadian genes: *Clock*, *Dbp*, and *Tef*. Each of these genes concurrently express both a principal, well-studied spliceform (Fig. 2.3A; orange gene models), and at least one minor, but less well-studied spliceform (Fig. 2.3A; purple gene models). In the case of *Clock*, we saw the spliceform skipping exon 18 (Fig. 2.3A; top panel), which was originally identified during the positional cloning of *Clock* [90]. For *Dbp*, there was a spliceform skipping exon 2 (Fig. 2.3A; middle panel). Finally, in the case of *Tef*, we saw a spliceform with an extra exon added between the first and second (Fig. 2.3A; bottom panel). We validated that spliceforms from the same gene are expressed and oscillate in phase with each other using spliceform-specific qPCR primers in wildtype samples (Fig. 2.3B). These spliceforms were also expressed in the *Clock*$^{\Delta 19}$ samples, but were arrhythmic and expressed at extremely low levels (Fig. S2.2). Taken together, these results not only confirm the concurrent expression of multiple spliceforms in clock genes, but also suggest that transcription and splicing may co-occur for these genes.

Alternative transcriptional start site in Dbp intron

In addition to alternative splicing, we also saw many genes with expression peaks located within their introns. We have seen 853 oscillating genes with intronic expression (RPKM of intron is at least 10% of those for adjacent exons) in the wildtype samples. We chose to focus on *Dbp* because of its importance as a well-characterized circadian output gene in the liver [92]. According to RNA-seq coverage plots, there is an expressed peak located in *Dbp*'s first intron (Fig. 2.4A). We did not detect this peak in our RNA-seq data from the *Clock*$^{\Delta 19}$ samples (Fig. S2.3A), which is likely due to the extreme loss of amplitude in *Dbp* expression we observed previously (Fig. S2.2B). This peak oscillates in phase with the remainder of the transcript, and has amplitude similar to that of the surrounding exons. This result was confirmed by quantitative PCR, from an independent tissue collection, using primers specific to the intronic sequence (Fig. 2.4B; lower panel), and the mature spliced transcript (Fig. 2.4B; upper panel). These results provide strong evidence that this peak is expressed, and that it oscillates with the remainder of the *Dbp* transcript.

There are three likely explanations for this peak in expression: 1) It is a novel exon spliced into the *Dbp* transcript. 2) It is a small RNA that is processed from the spliced intronic sequence, similar to *Mir132* and *Mir212* [93]. 3) It represents a novel, alternative transcriptional start site (TSS) for another *Dbp* spliceform. However, there was no evidence of splice junctions connecting this peak to any exons in the *Dbp* transcript (Fig. S2.3A), which suggests this is not a novel exon. The height of this peak in the RNA-seq data suggested that this was not simply a retained intron. To determine the size of the transcript(s) containing this intron, we performed a northern blot with probes specific to the region of the intronic peak as well as the spliced junction between exons 1 and 2 (Fig. 2.4C). For the splice junction probe, we saw bands at the appropriate lengths for *Dbp* pre-mRNA and the mature spliced transcript, as expected (Fig. 2.4C; upper panel). For the intron probe, we saw the expected pre-mRNAs (Fig. 2.4C; lower panel). However, rather than seeing a short length band corresponding to a small RNA, we saw a band of a similar length to the mature spliced *Dbp* mRNA. Furthermore, since we performed the northern blot using mouse liver RNA samples from the peak and trough of *Dbp* expression (CT34 and CT46 respectively), we saw that all of these bands oscillate together. Interestingly, a recent study has found evidence of rhythmic CLOCK/BMAL1 binding, and rhythmic changes in histone modifications directly upstream of this intronic peak in expression [44]. These histone modifications include H3K4me3, a marker for active promoters [94], as well as H3K9ac and H3K27ac, which are also associated with promoters [95, 96]. This site of rhythmic clock factor binding and chromatin modification is located within the first intron of *Dbp* and is independent of the binding/chromatin signal present at the annotated TSS. These new data indicate this intronic transcript is not a small RNA, but rather an alternative TSS. To test this hypothesis, we performed a 5' RACE experiment and did in fact find evidence of a TSS in *Dbp*'s first intron corresponding to the beginning of this peak in intronic expression (Fig. S2.3C). Taken together, all of these results suggest that an alternative form of *Dbp* is transcribed from an alternative TSS located in its first intron. This alternative, un-annotated TSS is similar to those found for *Clock* and *Timeless* in *Drosophila* [48].

A cycling miRNA cluster

In addition to robust expression within introns, we were also surprised to see RNA-seq coverage of miRNAs. Given that our RNA isolation and library construction methods were not optimized for small RNAs, those RNA-seq reads aligning to miRNA loci (taken from miRbase [97]) are likely from the larger, primary transcripts. We were very interested to examine miRNAs from a circadian perspective, since previous work has shown they play an important role in the regulation of the *Drosophila* clock [52]. None of these loci oscillated in our wildtype samples when analyzed by JTK_CYCLE. This is likely due to the low sampling resolution of this study, and the fact that the annotations we used were limited to stem-loop sequences. Through manual curation and visual inspection of the top hits near the significance threshold for oscillation, we found several members of a miRNA cluster with a primary transcripts that oscillate (Fig. 2.5A). This cluster, located on chromosome 7, consists of *Mir290, Mir291a, Mir292, Mir291b,* *Mir293*, *Mir294*, and *Mir295*. Interestingly, the annotated locations of the stem-loop structures correspond to dips in the RNA-seq coverage plot (Fig. 2.5A). Given the high degree of RNA secondary structure associated with the stem-loop region, it is possible that this drop in coverage is due to the inaccessibility of the RNA during library construction. We tested three regions by qPCR: *Mir292*, *Mir291b*, and the putative TSS for this miRNA cluster (Fig. 2.5B). Each of these regions showed consistent, low-amplitude oscillations. Additionally, we saw loss of these oscillations in our *Clock*$^{\Delta 19}$ samples (Fig. S2.4), indicating expression at these loci is regulated at least in part by the molecular clock. These miRNAs have largely been characterized in mouse ES cells [98], so little is known about their role in liver biology.

Novel circadian lincRNAs

From the wildtype RNA-seq data, we found seven oscillating junctions in intergenic regions that were not part of any known transcript, raising the possibility that our data includes novel, clock-regulated transcripts. We chose to focus on two of these loci, located on chromosomes 6 and 7 (Fig. 2.6). Interestingly, the structure of the adjacent junctions from the RNA-seq data (Fig. 2.6A and B; middle panels) forms putative exons from their boundaries (Fig. 2.6A and B; bottom panels). These putative exons correspond to peaks in the coverage plot (Fig. 2.6A and B; top panels). This is true even for the chromosome 6 locus, despite the relatively noisy coverage plot. In the case of the chromosome 7 locus, the putative exons line up almost exactly with peaks in the coverage plots. We were able to confirm that they are rhythmically expressed using custom designed qPCR primers (Fig. 2.6C). Interestingly, when we examined these same transcripts in the *Clock*$^{\Delta 19}$ samples, they continued to oscillate but at a different phase than in the wildtype livers (Fig. 2.6C). This suggests that the clock influences the expression of these novel trancripts, but is not the sole driver. Additionally, we used the PCR primers to clone and sequence their amplicons, and confirm the presence of the spliced forms of these transcripts (see Appendix B). Our RNA-seq data define the boundaries and structure of the chromosome 7 transcript most clearly, predicting a transcript length of 1843 bp. We performed Northern blots using probes specific to the same spliced regions assayed by the qPCR primers (Fig. 2.6D). These probes hybridized in bands of the same size predicted by RNA-seq. Having confirmed their expression, we examined the sequences for these putative transcripts for ORFs and found none, which indicates these transcripts likely do not have protein-coding potential. Additionally, we performed a BLASTX search using these transcript sequences to determine if proteins produced by these transcripts are similar to any known protein. BLASTX yielded no results, which provides further evidence that these transcripts are non-coding. Taken together, these data suggest these putative transcripts are long non-coding intergenic RNAs (lincRNAs).

2.4 Discussion

This study presents a combined RNA-seq and DNA microarray analysis of circadian transcriptional rhythms, revealing a high degree of diversity in oscillating transcripts. This hybrid approach has many significant advantages, notably combining the ease and affordability of DNA microarrays to conduct high-resolution studies, with the un-biased resolution of RNA-seq. Our hybrid approach will prove useful for future studies, since it allows us to leverage the vast amount of pre-existing array data to bolster the ability of RNA-seq studies to accurately identify oscillating transcripts. In an effort to make this data available to the community, we have integrated our RNA-seq data into the web interface we use for circadian microarray profiles (http://bioinf.itmat.upenn.edu/circa) [99].

We also demonstrated how a circadian profiling experiment's power to accurately identify oscillating transcripts is greatly influenced by sampling resolution. To this end, we recommend collecting a minimum of 24 samples (1-hour resolution for 1-day study, 2-hour resolution for 2-day study, etc.) to accurately identify the majority of oscillating transcripts, while limiting the number of false positives (Fig. 2.2). Given the substantial investment of time and resources required to both perform and analyze a circadian profile by RNA-seq, researchers may be limited to a lower than ideal sampling resolution. The integration of cheaper or pre-existing microarray data with RNA-seq provides one way of alleviating this problem.

Our RNA-seq data have also shown that most genes with rhythmic transcription express multiple spliceforms at the same time. This diversity in oscillating transcripts has not been previously seen by microarrays due to the limitations of the technology. While there are exceptions [41], the general trend appears to be that alternative transcripts from the same gene are regulated together and accumulate with the same phase and period. This trend is in agreement with previous RNA-seq studies performed in *Drosophila* and *Arabidopsis* [48, 100]. One of the best uses for this splicing information in future studies may be for comparing spliceform usage of rhythmic genes across different tissues.

In addition to alternative splicing, our RNA-seq data have allowed us to identify oscillating, non-coding RNAs. We used RNA-seq to characterize novel oscillating transcripts that appear to be non-coding (Fig. 2.6). Our finding that RNA-seq is capable of identifying novel circadian clock genes is in agreement with previous studies [48, 53]. Furthermore, lncRNAs are an ideal place to look for new clock components, since new lncRNAs are continually being discovered [101], and there is emerging evidence of their involvement in cancer and disease [102, 103]. Given the existence of many transcripts that are not polyadenylated [104], future studies using rRNA depletion, as well as RNA isolation and library construction techniques optimized for small RNAs may find a greater number of novel oscillating transcripts and miRNAs in the mammalian system. Additionally, since novel transcripts and miRNAs are traditionally not detectable by arrays, it is likely there are more oscillating transcripts of this kind that will require sequencing at a higher temporal resolution to accurately identify. The miRNA cluster centered around *Mir292* (Fig. 2.5) that we did identify as oscillating appears to be involved in mouse ES cell stress response [105]. Interestingly, *Casp2* and *Ei24*, two of the validated targets of this miRNA cluster involved in stress response, are themselves rhythmically transcribed [40]. Previous studies demonstrating that the clock and CCGs can be regulated by miRNAs [52, 106, 107] emphasize the need to include small RNA sequencing in future profiling experiments, and we have demonstrated that RNA-seq is capable of profiling the larger, miRNA primary transcripts. Furthermore, given that our focus has previously been restricted to protein-coding transcripts, due largely to technological limitations, it is likely that non-coding RNAs hold great promise for future studies as a source of novel CCGs and clock regulators.

2.5 Methods

Circadian tissue collection

Mouse liver tissue was collected as previously described [40]. Briefly, WT 6-week old male C57/BL6 mice were acquired from Jackson Labs; *Clock*$^{\Delta 19}$ mice were prepared as previously described [38]. Mice were entrained to a 12-h:12-h light:dark schedule for 1 week before being shifted to total darkness. Mice were supplied with food and water *ad libidum*. Starting at CT18, 3-5 mice were sacrificed in the darkness every 6 hours for 2 days. Livers samples were quickly dissected and snap-frozen in liquid nitrogen. Liver samples used for quantitative PCR were collected in the same manner, except collection began at CT24 and continued every 2 hours for 2 days. All animal experiments were performed in accordance with the approval of the Institutional Animal Care and Use Committee.

RNA-seq library preparation

Liver samples were homogenized in Trizol reagent (Invitrogen) using a Tissuelyser homogenizer (Qiagen). RNA was extracted using RNeasy columns according to the manufacturer’s protocol (Qiagen). For the full library preparation protocol, please see Appendix C. Briefly, equal quantities of total RNA were mixed from 3 animals for each time point. Total RNA was subjected to two rounds of poly(A) selection with Dynabeads (Invitrogen). The mRNA was fragmented for 5 minutes by metal-ion hydrolysis (Ambion), and then used as a template for a random-primed cDNA generation. Following a second-strand synthesis reaction, and a phenol-chloroform extraction, cDNA fragments were prepared for sequencing by an end-repair reaction. Blunted fragments were adenylated to create a single A overhang, and Illumina adapters were ligated to these sticky ends. Library fragments were size-selected by gel electrophoresis (350-550bp), and amplified by 13 PCR cycles.

Library sequencing and analysis

RNA-seq libraries were each sequenced in a single lane on the Illumina Genome Analyzer IIx, using the 100bp paired-end chemistry. Raw reads were aligned to the mouse genome (mm9/NCBI37) using the RNA-seq Unified Mapper [61]. Transcripts were quantified by RUM using the UCSC [108], Refseq [109], and Vega [110] gene annotations. Quantification values were tested for oscillations using the JTK_CYCLE [91] package in R. miRNAs from the miRbase annotations v18 [97] were also quantified and tested for oscillations. For all junction analyses, only those junctions identified by RNA-seq reads mapping at least 8 bp on each side were used. A junction was classified as novel if it did not appear in any of the following annotations: Vega, UCSC, Transcriptome, SGP, RefSeq, Other RefSeq, NSCAN, Genscan, GeneId, Esnsembl and AceView [108–116]. Each of these annotation tracks was downloaded from the UCSC Genome Browser. Genes with alternative splicing were identified by searching for splice junctions with identical start coordinates, but different end coordinates (and vice versa). We have made our raw RNA-seq data and aligned results freely available on GEO (accession numbers: GSE40190 and GSE41082).

Sampling resolution simulation

The gold standard of oscillating genes was determined by analyzing the Hughes 1-hour data set [40] with the JTK_CYCLE [91] package in R. A 2-hour sampling resolution was simulated by taking data from every second array from the Hughes data set, beginning with CT18, and using JTK_CYCLE to identify oscillating genes. A 3-hour sampling resolution was simulated by taking data from every third array, and so on, up to a 10-hour sampling resolution. A one-day data set was simulated by repeating the above process, but only using arrays from the first day of sampling (CT18-CT41). The one-day simulation used a maximum resolution of 8-hours. For the comparison, a true positive was defined as a gene identified as oscillating in both the simulated data set and the gold standard. A false positive was defined as a gene identified as oscillating by the simulated data set, but not by the gold standard. The one-day simulation with replicates used time points separated by 24 hours as replicates for each other. For example, CT42 served as a replicate for CT18.

Quantitative PCR

0.5 µg of total RNA, pooled from 3 liver samples, was used to generate cDNA with the QuantiTect Reverse Transcription kit according to manufacturer's protocol (Qiagen). Quantitative PCR reactions were performed using SYBR Green PCR Master Mix (Applied Biosystems) in combination with custom primers on 7900 HT Real-Time PCR System (Applied Biosystems). Rps18 (Mm_Rps18_1_SG) was used as endogenous control for all qPCR experiments. Information for this primer is available from the manufacturer's website (Qiagen). Primer pairs were designed using NCBI Primer-

BLAST [117] and ordered from Integrated DNA Technologies. Sequences for primer pairs are included in Table S2.2. All analysis was performed using RQ Manager v1.2.1 (Applied Biosystems).

5-prime RACE

Total RNA was isolated from mouse livers collected at the peak phase of *Dbp* expression (CT34). The RNA was used for 5-prime race using FirstChoice RLM Race Kit (Applied Biosystems) and Superscript III (Invitrogen) by manufacturer's protocols. Race products generated using forward primers included with kit and reverse primers used for *Dbp* qPCR primers (see Table S2.2). These products were visualized on a 2% agarose gel, purified using MinElute Gel Purification columns (Qiagen), and sequenced to confirm their identities.

Northern blots

Templates for Northern probes were generated from qPCR amplicons. Briefly, amplicons were run on a 2% agarose gel following qPCR reactions and purified using MinElute Gel Purification kit (Qiagen) by manufacturer's protocol. Purified amplicons were cloned using TOPO TA Cloning Kit (Invitrogen) by manufacturer's protocol. Templates verified by sequencing (sequences included in Appendix B). RNA for Northern blots was isolated from frozen mouse liver samples (collected as described above) by TRIzol according to manufacturer's protocol (Invitrogen). For the Dbp Northern blots, 40 µg of total RNA was subjected to poly(A) purification by Dynabeads according to manufacturer's protocol (Invitrogen). Northern Blots performed using

NorthernMax Kit according to manufacturer’s protocols (Ambion). Briefly, 700 ng of poly(A) RNA (Dbp) or 20 µg of total RNA (novel chr7 transcript) were run on a 1.5% denaturing agarose gel. RNA was transferred to Amersham Hybond-N+ membrane (GE Life Sciences) by downward transfer using a TurboBlotter (Whatman). RNA was crosslinked to membrane using a UV Stratalinker 2400 (Stratagene) using the Auto Cross Link setting. Oligonucleotide probes incorporating P-32 alpha dCTP were synthesized from the probe template using Prime-It II Random Primer Labeling Kit (Agilent) and unincorporated nucleotides were removed using Micro Bio-Spin Chromatography Columns (Biorad), according to manufacturer’s protocols. Probes were hybridized overnight in an Isotemp Hybridization Incubator (Fisher) at 46ºC. Hybridized probes were exposed to a Storage Phosphor Screen (GE Life Sciences) overnight and imaged using a Storm 840 Phosphorimager (Molecular Dynamics).

2.6 Tables

Table 2.1: Core clock spliceforms expressed in RNA-seq data.

Clock genes	Annotated spliceforms[1]	Detected in wild-type	Detected in *Clock*$^{\Delta 19}$
Clock	6	4	3[3]
Npas2	1	1	1
Arntl	7	3	3
Arntl2	3	1	1
Cry1	1	1	1
Cry2	2	1	1
Per1	5	3	3
Per2	2	1	1
Per3	6	3	2
Fbxl3	3	2	2
Fbxl21	4	2	2
Nr1d1	1	1	1
Nr1d2	2	1	1
Rora	4	2	2
Rorb[2]	5	0	0
Rorc	5	3	3
Csnk1a1	6	4	3
Csnk1d	4	3	3
Csnk1e	2	2	2

[1] derived from visual inspection of UCSC, Refseq, Ensemble, & NSCAN annotations

[2] not expressed in RNA-seq data set

[3] The loss of exon 19 in the mutant masks other spliceforms

2.7 Figures

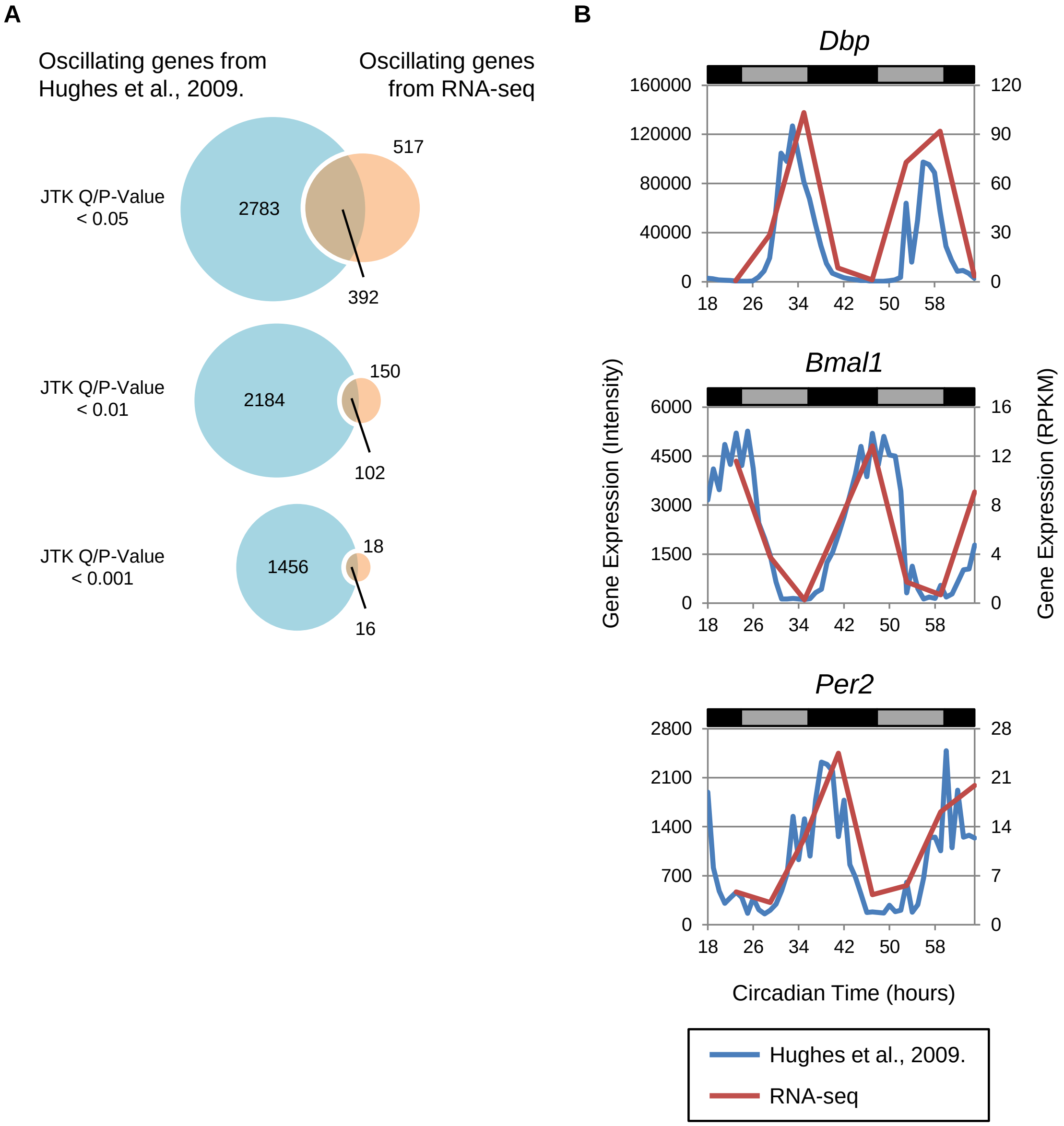
A
Oscillating genes from Hughes et al., 2009.
Oscillating genes from RNA-seq
JTK Q/P-Value < 0.05
2783
517
392
JTK Q/P-Value < 0.01
2184
150
102
JTK Q/P-Value < 0.001
1456
18
16
B
Dbp
160000
120000
80000
40000
0
120
90
60
30
0
Bmal1
6000
4500
3000
1500
0
16
12
8
4
0
Per2
2800
2100
1400
700
0
28
21
14
7
0
18
26
34
42
50
58
Gene Expression (Intensity)
Gene Expression (RPKM)
Circadian Time (hours)
Hughes et al., 2009.
RNA-seq

Fig. 2.1: Comparison between RNA-seq and previous array study.

(A) RNA-seq data and array data from Hughes et al. 2009, were analyzed with JTK_CYCLE to identify genes with oscillating transcripts. These gene lists were compared at different JTK q-value cutoffs of 0.05, 0.01, and 0.001 to determine the number of overlapping genes. For the RNA-seq data, JTK p-values were used instead of q-values due to the lower time resolution of the data set. **(B)** Expression profiles from Hughes et al. array data (blue) and wildtype RNA-seq data (red) for three genes identified as oscillating by both datasets: *Dbp*, *Bmal1*, and *Per2*. Grey and black bands identify times corresponding to subjective day and night, respectively. Array intensity plotted using the left y-axis and RNA-seq RPKM values plotted using the right y-axis.

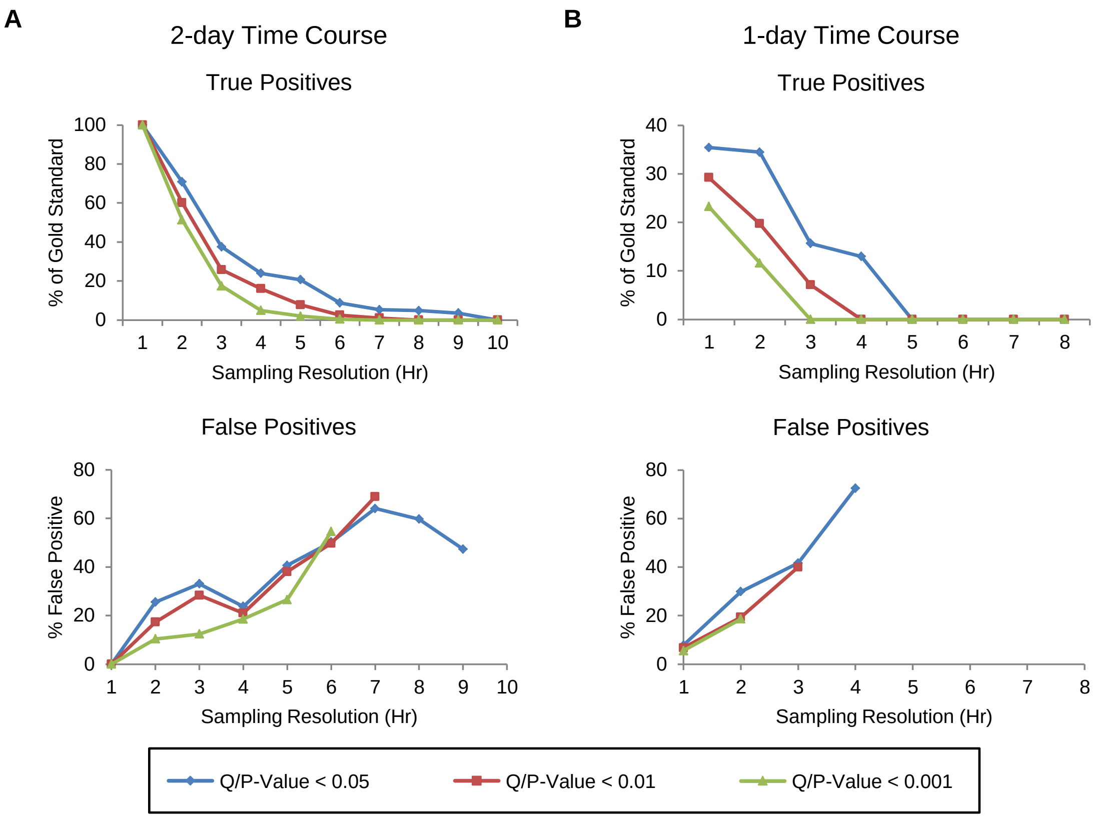
A
2-day Time Course
True Positives
% of Gold Standard
100
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10
Sampling Resolution (Hr)
False Positives
% False Positive
80
60
40
20
0
1 2 3 4 5 6 7 8 9 10
Sampling Resolution (Hr)
B
1-day Time Course
True Positives
% of Gold Standard
40
30
20
10
0
1 2 3 4 5 6 7 8
Sampling Resolution (Hr)
False Positives
% False Positive
80
60
40
20
0
1 2 3 4 5 6 7 8
Sampling Resolution (Hr)
Q/P-Value < 0.05
Q/P-Value < 0.01
Q/P-Value < 0.001

Fig 2.2: Impact of time-resolution on circadian gene expression profiles. Circadian profiling studies of different sampling resolutions were simulated by taking different subsets of the Hughes et al. 2009 array data set and analyzing them with JTK_CYCLE. These subsets used data from **(A)** 2 days of the array data or **(B)** just the first day to simulate two day and one day experiments, respectively. The list of genes with oscillating transcripts from each of these subsets was compared to those identified by the full 2 day, one-hour resolution data set, which served as a gold standard. The percentage of genes with oscillating transcripts from a given subset that overlap with those from the gold standard served as an indicator of the true-positive rate for that subset (top panels). The percentages of genes that did not overlap with the gold standard served as the false-positive rate for that subset (bottom panels). These lists of oscillating genes identified by the subsets were calculated using JTK p-value cutoffs of 0.05, 0.01, and 0.001, which the list from the gold standard always used a JTK q-value cutoff of 0.05.

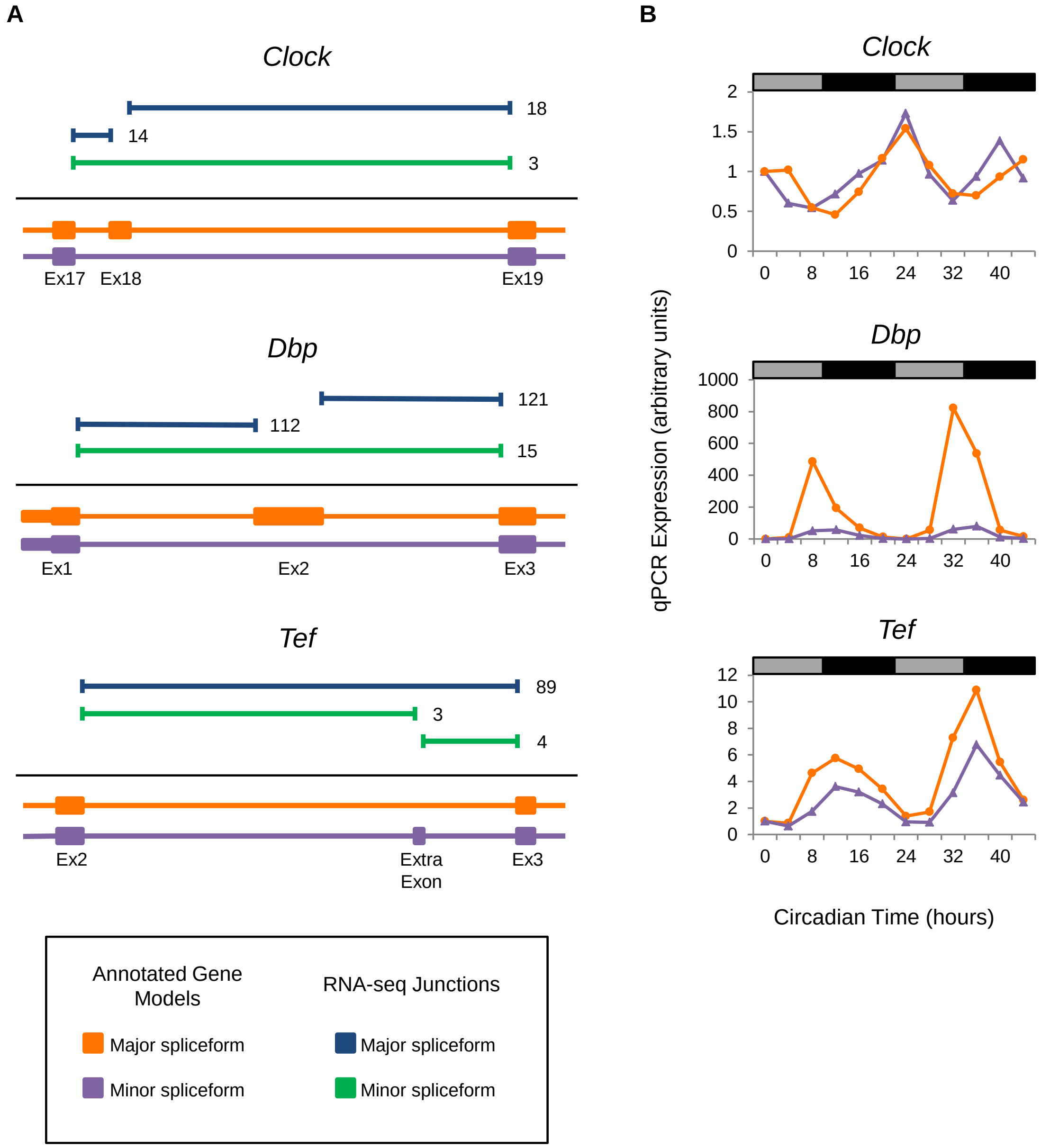
A
Clock
18
14
3
Ex17
Ex18
Ex19
Dbp
121
112
15
Ex1
Ex2
Ex3
Tef
89
3
4
Ex2
Extra Exon
Ex3
Annotated Gene Models
RNA-seq Junctions
Major spliceform
Minor spliceform
Major spliceform
Minor spliceform
B
Clock
2
1.5
1
0.5
0
0
8
16
24
32
40
Dbp
1000
800
600
400
200
0
Tef
12
10
8
6
4
2
0
qPCR Expression (arbitrary units)
Circadian Time (hours)

Fig. 2.3: Multiple spliceforms of core clock genes.

(A) Exon junctions mapped by RNA-seq data were used to identify major (blue junctions, orange gene models) and minor (green junctions, purple gene models) spliceforms for *Clock*, *Dbp*, and *Tef*. The numbers adjacent to the blue and green junctions are the number of gapped RNA-seq reads mapping to those junctions from a representative time point. **(B)** qPCR with primers specific to major (orange) and minor (purple) spliceforms shows expression of these spliceforms across two days. Grey and black bands identify times corresponding to subjective day and night, respectively. Expression of transcripts was normalized to Rps18.

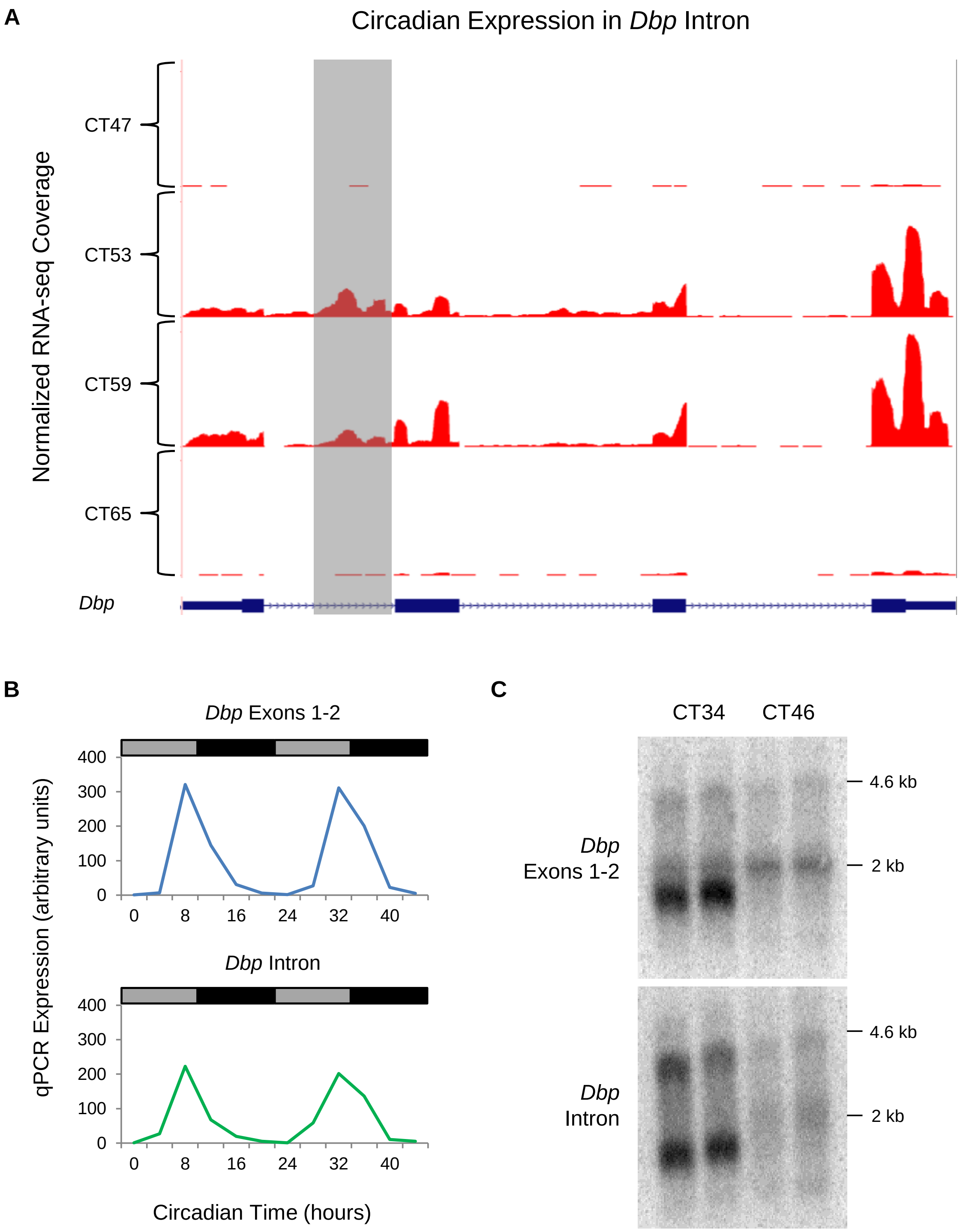
A
Circadian Expression in *Dbp* Intron
Normalized RNA-seq Coverage
CT47
CT53
CT59
CT65
Dbp
B
Dbp Exons 1-2
400
300
200
100
0
0
8
16
24
32
40
Dbp Intron
qPCR Expression (arbitrary units)
Circadian Time (hours)
C
CT34
CT46
Dbp
Exons 1-2
Dbp
Intron
4.6 kb
2 kb

Fig. 2.4: Expression peak in *Dbp* intron.

(A) Wildtype RNA-seq coverage plots from four representative time points (CT47, CT53, CT59, CT65) are displayed above the *Dbp* gene model (purple). The gray bar highlights the region in the first intron showing rhythmic expression. Coverage plots are normalized by the total number of reads in in each RNA-seq sample. **(B)** qPCR with primers specific to the mature splice junction between the first and second exon (top), and the expressed region of the first intron (bottom) was used to detect expression of these transcripts in wildtype livers across two days. Grey and black bands identify times corresponding to subjective day and night, respectively. Expression of these transcripts was normalized to Rps18. **(C)** Northern blots with probes specific to the mature splice junction between the first and second exon (top), and the expressed intronic region (bottom) were used to determine the size of the transcripts associated with each of these regions. Poly(A) RNA for the Northern was collected from two independent, wildtype mouse livers at the peak (CT34) or trough (CT46) of *Dbp* expression.

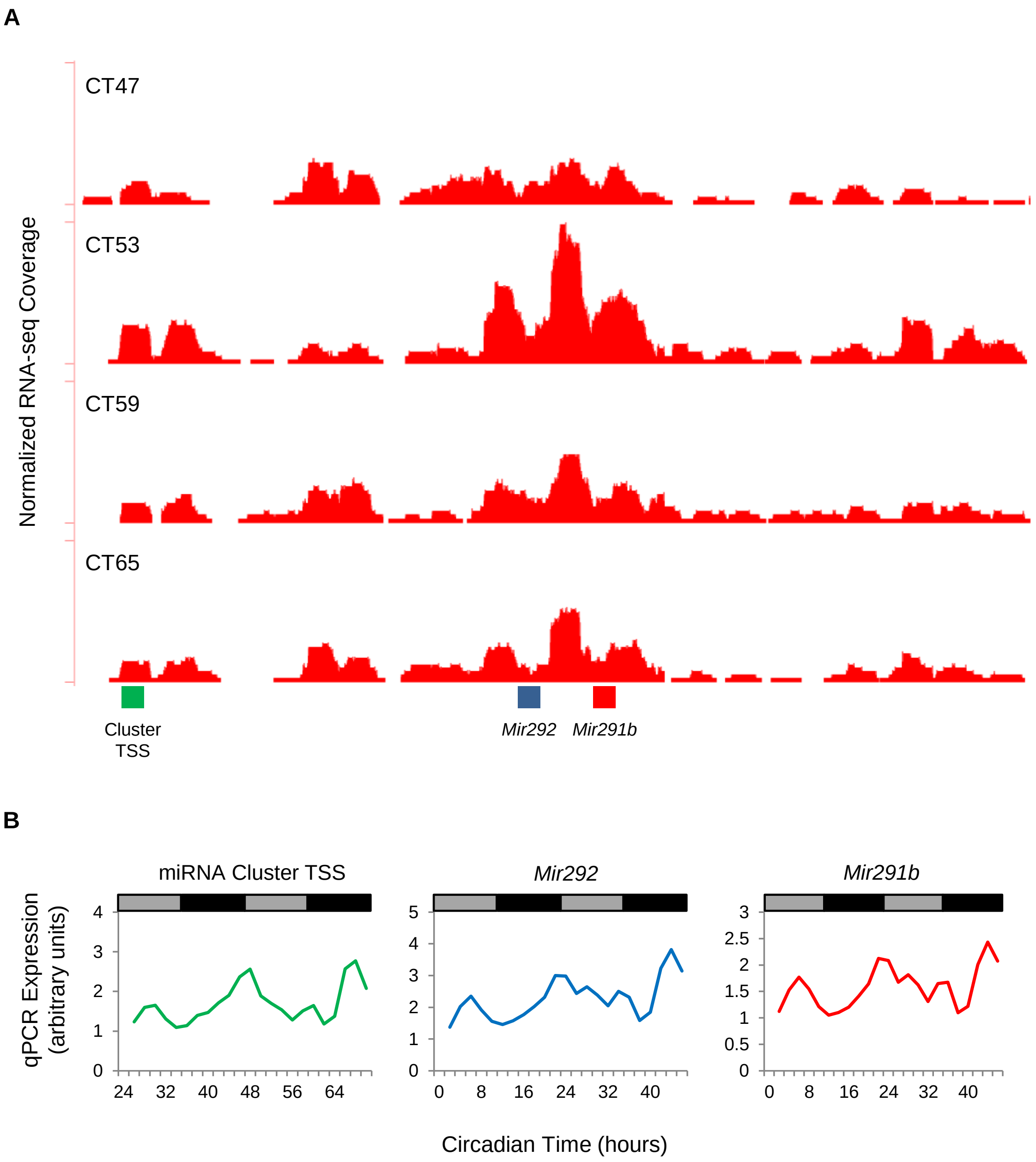
A
Normalized RNA-seq Coverage
CT47
CT53
CT59
CT65
Cluster TSS
Mir292
Mir291b
B
miRNA Cluster TSS
Mir292
Mir291b
qPCR Expression (arbitrary units)
0
1
2
3
4
24
32
40
48
56
64
0
1
2
3
4
5
0
8
16
24
32
40
0
0.5
1
1.5
2
2.5
3
0
8
16
24
32
40
Circadian Time (hours)

Fig. 2.5: Oscillating miRNA cluster.

A miRNA cluster containing *Mir290*, *Mir291a*, *Mir292*, *Mir291b*, *Mir293*, *Mir294*, and *Mir295* is located at chr7:3,217,507-3,221,276. **(A)** Wildtype coverage plots for this genomic region from four representative time points (CT47, CT53, CT59, CT65) are displayed. The annotated locations of *Mir292* (blue) and *Mir291b* (red), in addition to the putative TSS of the cluster (green) are displayed below the coverage plots. All coverage plots are normalized by the total number of reads in each sample. **(B)** qPCR with primers specific to *Mir292* (blue), *Mir291b* (red), and the putative TSS (green) were used to detect expression in wildtype livers across two days. Grey and black bands identify times corresponding to subjective day and night, respectively. Expression of these transcripts was normalized to *Rps18*.

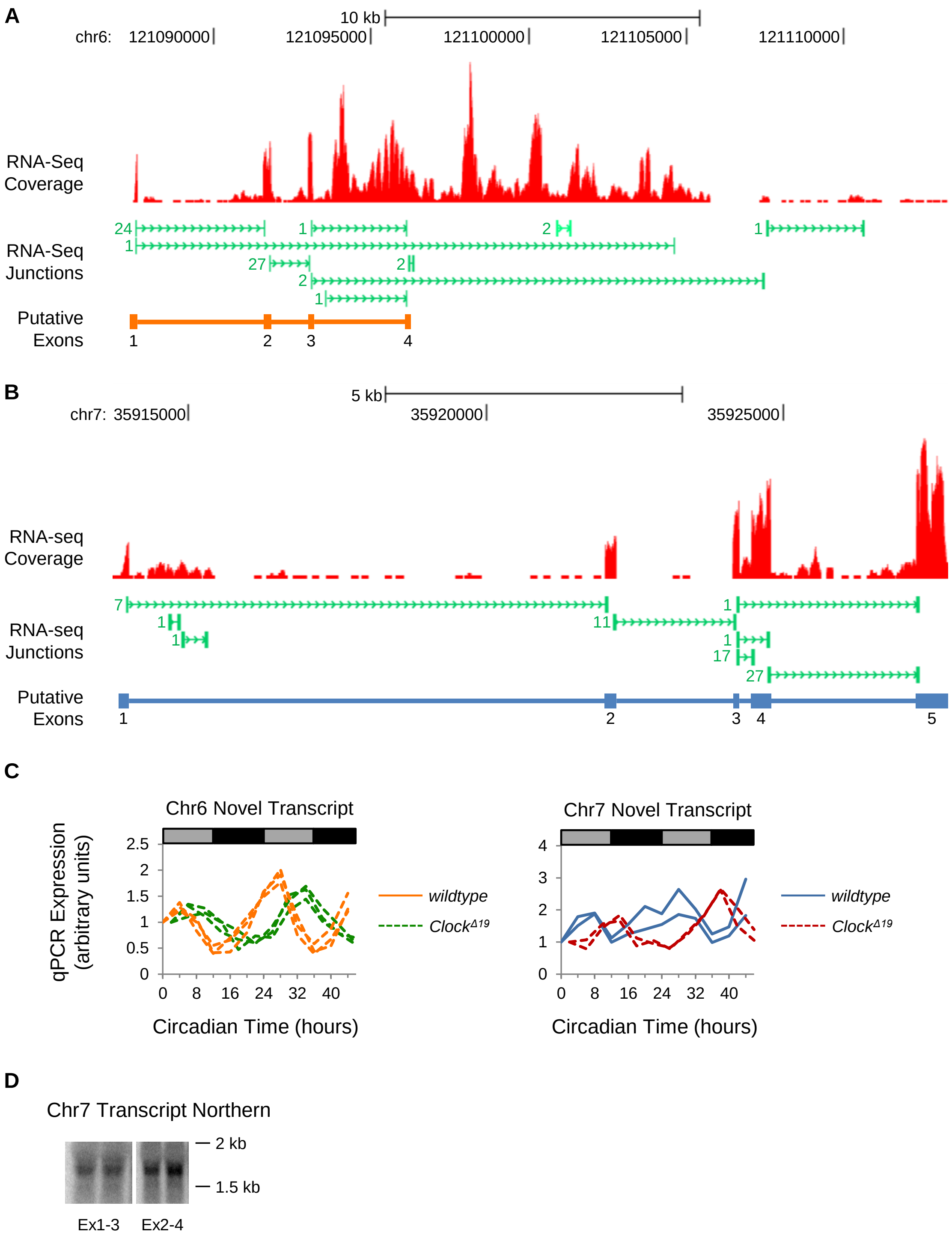
A
10 kb
chr6: 121090000 121095000 121100000 121105000 121110000
RNA-Seq Coverage
RNA-Seq Junctions
24
1
1
27
2
2
2
1
1
Putative Exons
1
2
3
4
B
5 kb
chr7: 35915000 35920000 35925000
RNA-seq Coverage
RNA-seq Junctions
7
1
1
11
1
1
17
27
Putative Exons
1
2
3
4
5
C
Chr6 Novel Transcript
qPCR Expression (arbitrary units)
2.5
2
1.5
1
0.5
0
0
8
16
24
32
40
Circadian Time (hours)
wildtype
ClockΔ19
Chr7 Novel Transcript
4
3
2
1
0
0
8
16
24
32
40
Circadian Time (hours)
wildtype
ClockΔ19
D
Chr7 Transcript Northern
2 kb
1.5 kb
Ex1-3
Ex2-4

Fig. 2.6: Novel oscillating lincRNAs.

Two putative transcripts were identified at **(A)** chr6:121,086,416-121,114,417 and **(B)** chr7:35,913,467-35,928,124. Representative RNA-seq coverage plots are displayed in red for each of these genomic regions. Numbers next to splice junctions (green) list the number of RNA-seq reads with gapped alignments identifying the corresponding junction. Locations of predicted exons displayed below junctions. **(C)** qPCR primers spanning multiple spliced exons were used to detect expression of novel transcripts across two days in wildtype (orange and blue) and *Clock*$^{\Delta 19}$ (green and red) livers. Three primer pairs spanning exons 1-2, 1-3, and 2-4 were used for chr6 transcript (left panel), while two primer pairs spanning exons 1-3 and 2-4 were used for chr7 transcript (right panel). Grey and black bands identify times corresponding to subjective day and night, respectively. Expression of these transcripts was normalized to Rps18. **(D)** Northern blots with probes specific to the same regions assayed by qPCR were used to determine the size of the chr7 transcript. Total RNA for the Northern was collected from two independent, wildtype mouse livers collected at CT28.

2.7 Supplemental tables

Table S2.1: RNA-seq alignment statistics.

Sample	Total Number of Reads	% Mapped Uniquely	% Mapped
WT CT23	13577682	82.82%	95.80%
WT CT29	12453604	84.73%	96.60%
WT CT35	27943589	81.84%	95.80%
WT CT41	12266671	84.34%	95.30%
WT CT47	13279060	84.85%	97%
WT CT53	13197401	84.07%	94.20%
WT CT59	12198646	85.62%	96.40%
WT CT65	14512233	83.43%	95.60%
Clock-Mutant CT22	28873747	82.57%	96.1%
Clock-Mutant CT28	28504730	85.63%	96.6%
Clock-Mutant CT34	25768876	85.47%	95.1%
Clock-Mutant CT40	27491918	85.95%	96%

Table S2.2: qPCR primer sequences.

Target Gene	Fw Primer (5' -> 3')	Rv Primer (5' -> 3')
Clock - Major	AGCCAGCGATGTCTCAAGCTGCA	CATCCGTGTCCGCTGCTCTAGC
Clock - Minor (Skip Ex 16)	CGCAGTCTCAGACCCTTCCTCCA	AACTGAGCTGAAAACTGAAACTGACT
Dbp - Major	CGGCCTCTGAGCGACAGGAC	CACTAACGGCCCCACTCGGG
Dbp - Minor (Skip Ex 2)	CAAAGAACCGGCCAGCTGCTTGAC	GACAGGGCGAGATCAGCGGGA
Tef - Major	GAAACCGTGTCCAGCACAGAATCG	AGGTCGGCAGGGTCAGGGTT
Tef - Minor (Extra Exon)	TCCCCTACGATGGCGAGTCCT	GGGTCCTCCTGTTCCATATGGCTG
Dbp - Exons 1-2	GGAGCGCTGCTTGGGCTGAG	GAGGGGACCCACCGCCACTA
Dbp - Intron	CCCGGGCCCCTAACCCTATCC	GCCGTAGGGCAAAGACCCAGG
mir292	AGGGCGGTTCAGTTGGGTGC	ACCTGGCGGCACTTTTCTTCCG
mir291b	CGGCTTGGCGGGAAAGTGCA	CAGCTGCAGCCGGCTTTTCA
miRNA Cluter TSS	AGCCTCCCCCACGCCTCTC	GAAGCAGCACGCCGGAGGT
Chr6 Novel Transcript - Ex1-2	GCATCAGCTCCTGCTCCAGGTTC	GCTTTCTACCCCACGGGGTCTCT
Chr6 Novel Transcript - Ex1-3	CAGCCTCTGCATCAGCTCCTGC	GGTTCCTGGGACGCACTGGA
Chr6 Novel Transcript - Ex2-4	CCGTGGGGTAGAAAGCAGGAAGA	TGGAGTGAGCGAACGAGCGTC
Chr7 Novel Transcript - Ex1-3	AAGGCAGCTCTTGGGCCTCACT	GCAGTCTGTGGGACATGTGCC
Chr7 Novel Transcript - Ex2-4	CAAATGGTGACCCCTGCGCCTG	TGCTTAGCTGGCCCCCAGTTG

2.8 Supplemental figures

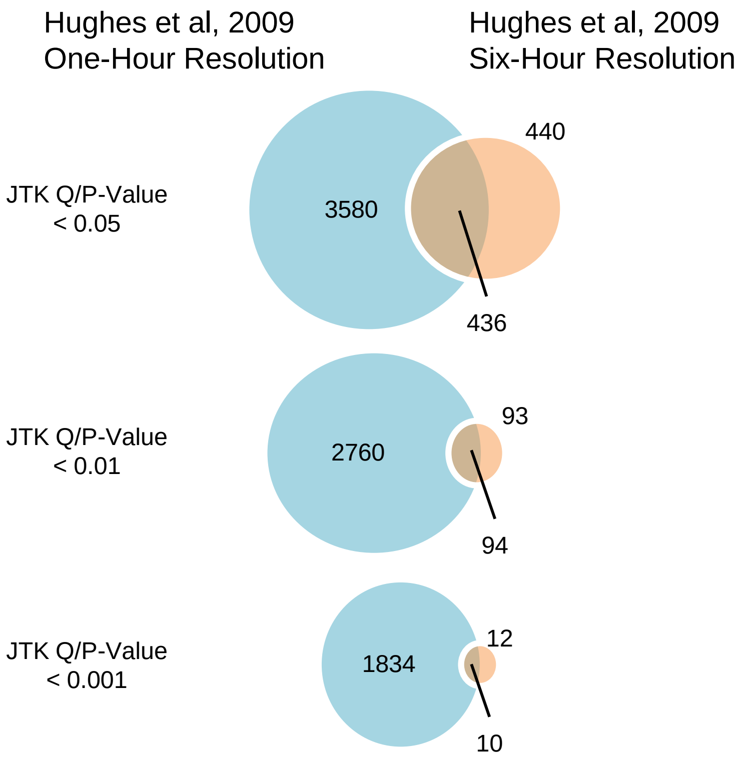

Fig. S2.1: Overlap between array data at 1-hour and 6-hour resolution.

Hughes et al. 2009 array data at 1-hour resolution and a simulated 6-hour data set were analyzed with JTK_CYCLE to identify genes with oscillating transcripts. These gene lists were compared at different JTK q-value cutoffs of 0.05, 0.01, and 0.001 to determine the number of overlapping genes. For the 6-hour data, JTK p-values were used instead of q-values due to the lower time resolution of the data set.

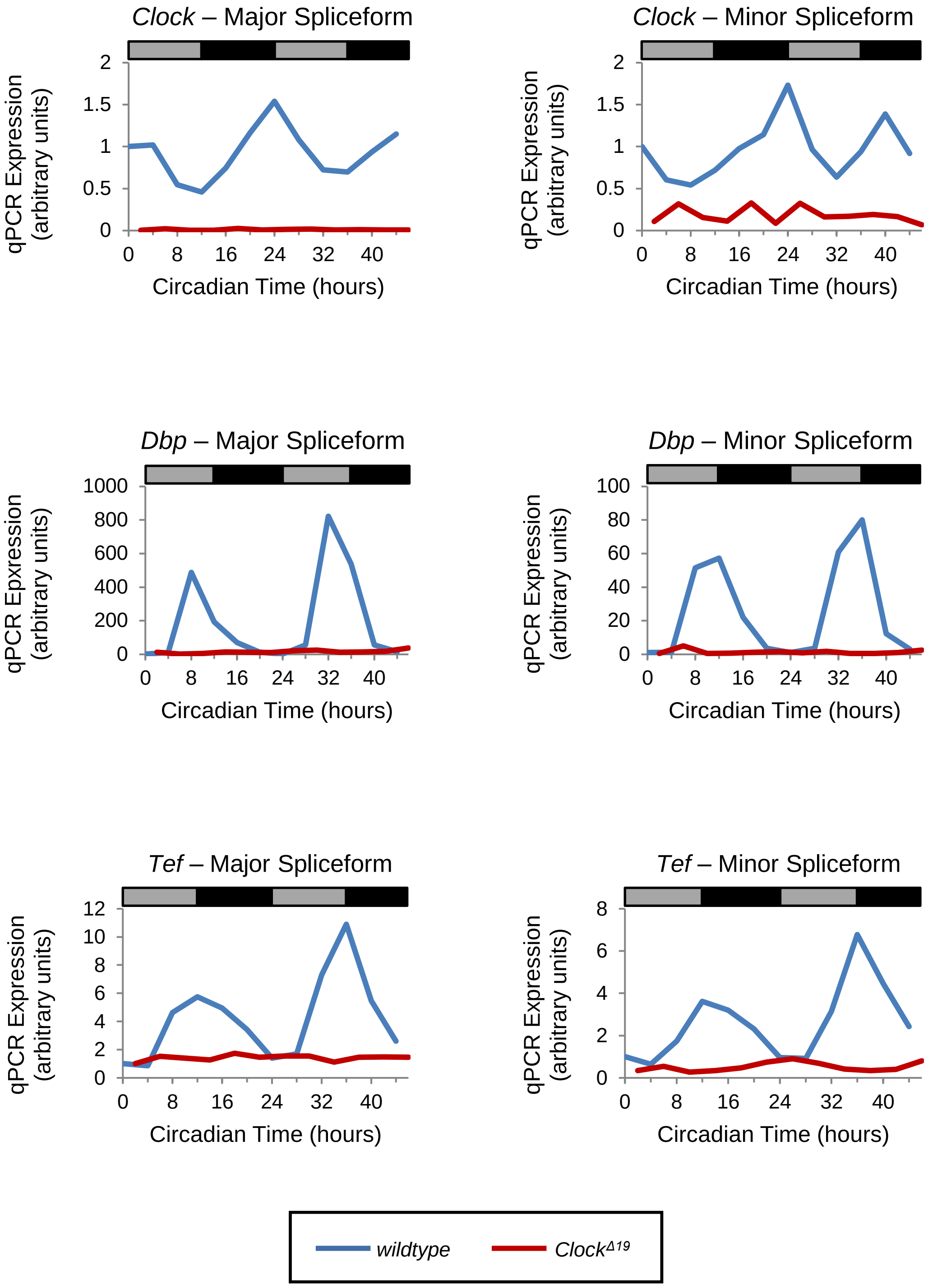
Clock – Major Spliceform
Clock – Minor Spliceform
qPCR Expression (arbitrary units)
2
1.5
1
0.5
0
0
8
16
24
32
40
Circadian Time (hours)
Dbp – Major Spliceform
Dbp – Minor Spliceform
qPCR Epxression (arbitrary units)
qPCR Expression (arbitrary units)
1000
800
600
400
200
100
80
60
40
20
Tef – Major Spliceform
Tef – Minor Spliceform
12
10
8
6
4
2
wildtype
Clock^Δ19

Fig. S2.2: *Clock*$^{\Delta 19}$ disrupts oscillations in *Clock*, *Dbp* and *Tef* expression.

Major (left) and minor (right) spliceforms for *Clock*, *Dbp*, and *Tef* were used to assay wildtype (red) and *Clock*$^{\Delta 19}$ (blue) liver RNA. These are the same qPCR primers used in Fig. 2.3B. Both spliceforms showed reduced expression in *Clock*$^{\Delta 19}$ liver relative to wildtype liver. Grey and black bands identify times corresponding to subjective day and night, respectively. Expression of transcripts was normalized to *Rps18*.

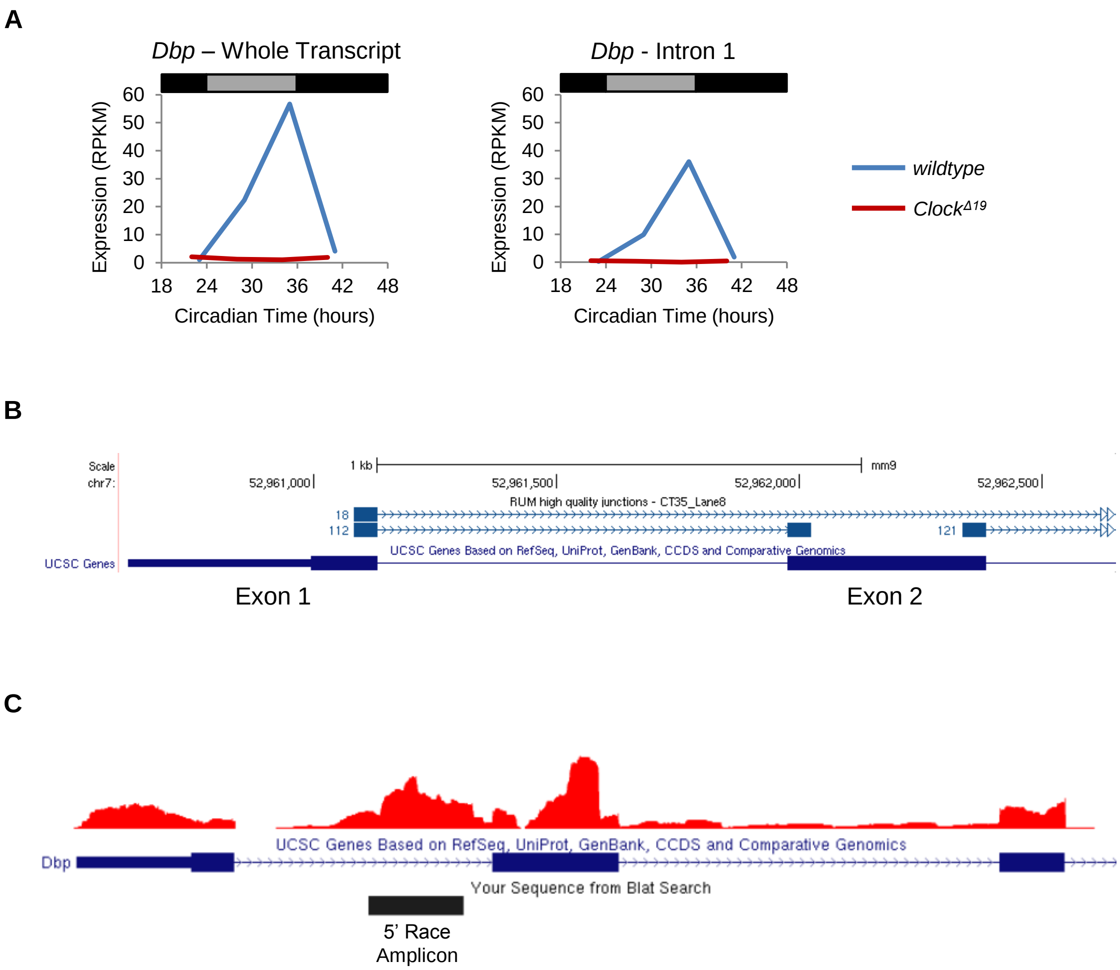
A
Dbp – Whole Transcript
Dbp - Intron 1
Expression (RPKM)
60
50
40
30
20
10
0
18
24
30
36
42
48
Circadian Time (hours)
wildtype
Clock$^{\Delta 19}$
B
Scale
chr7:
1 kb
mm9
52,961,000
52,961,500
52,962,000
52,962,500
RUM high quality junctions - CT35_Lane8
18
112
121
UCSC Genes
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Exon 1
Exon 2
C
Dbp
UCSC Genes Based on RefSeq, UniProt, GenBank, CCDS and Comparative Genomics
Your Sequence from Blat Search
5' Race
Amplicon

Fig. S2.3: Exon-exon junctions and 5' RACE suggest expression in Dbp intron corresponds to a TSS.

(A) RNA-seq expression profiles for the entire *Dbp* transcript (left) and the first intron (right) are plotted for wildtype (blue) and *Clock*$^{\Delta 19}$ (red) livers. RPKM expression values are plotted for a single day. Grey and black bands identify times corresponding to subjective day and night, respectively. **(B)** Exon-exon junctions (light blue traces above gene model) were identified by RNA-seq reads with gapped alignments mapping at least 8 bp on each side. Number listed adjacent to each junction corresponds to the number of RNA-seq reads mapped to that junction. **(C)** Amplicon resulting from 5' RACE using *Dbp* primers was sequenced and aligned to the genome using BLAT (black). RNA-seq coverage plot (red) and *Dbp* gene model (blue) are included for reference.

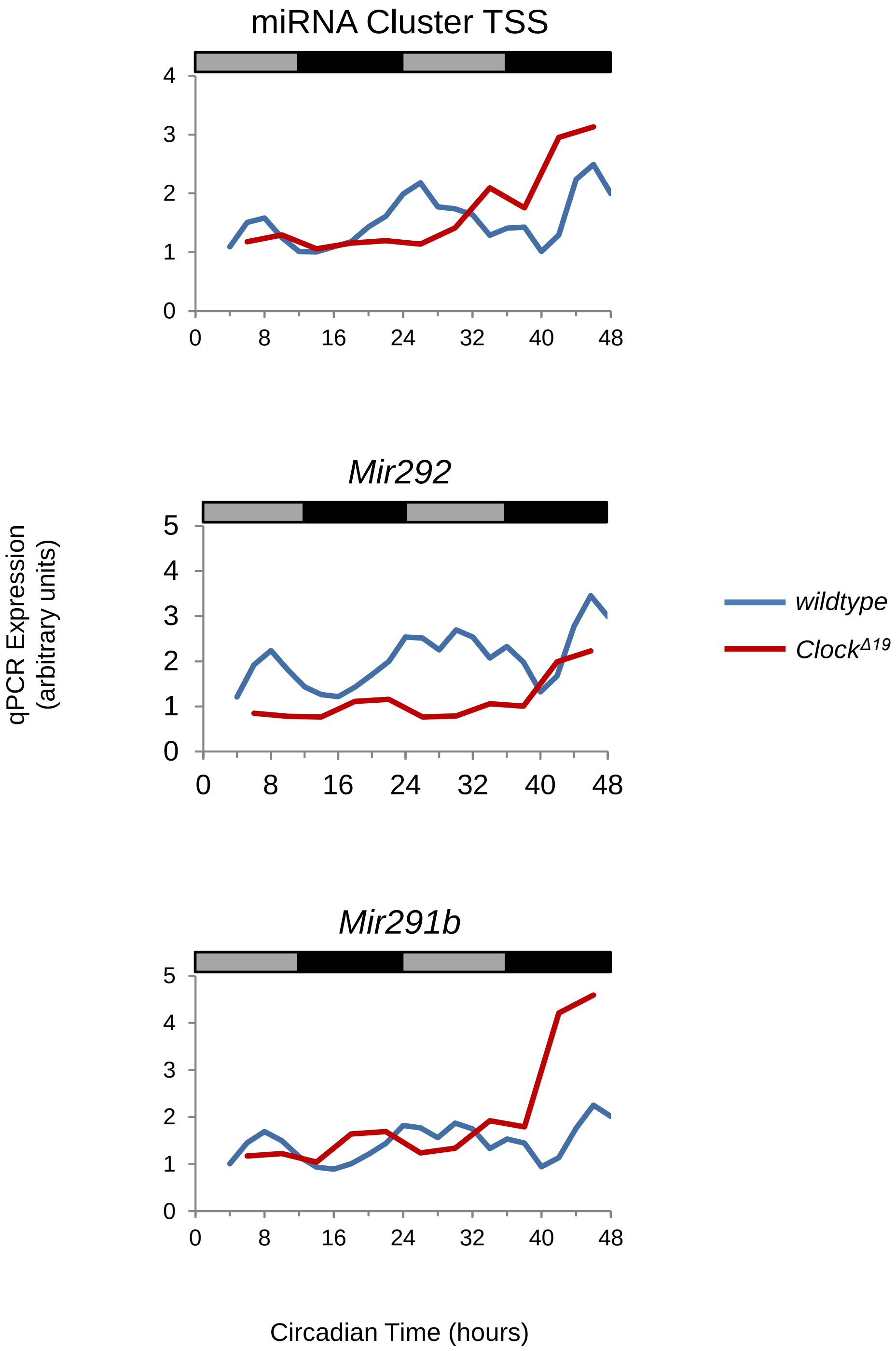
miRNA Cluster TSS
4
3
2
1
0
0
8
16
24
32
40
48
Mir292
5
4
3
2
1
0
0
8
16
24
32
40
48
Mir291b
5
4
3
2
1
0
0
8
16
24
32
40
48
qPCR Expression
(arbitrary units)
wildtype
$Clock^{\Delta19}$
Circadian Time (hours)

Fig. S2.4: *Clock$^{\Delta 19}$* disrupts oscillations in miRNA cluster.

qPCR with primers specific to *Mir292* (middle), *Mir291b* (bottom), and the putative TSS (top) were used to detect expression in wildtype and *Clock$^{\Delta 19}$* livers across two days. These are the same primers used in Fig. 2.5B. Grey and black bands identify times corresponding to subjective day and night, respectively. Expression of these transcripts was normalized to Rps18.

Chapter 3: The circadian non-coding transcriptome across twelve mouse tissues

3.1 Abstract

The circadian clock is a cell-autonomous, molecular oscillator that drives daily rhythms in behavior and physiology. Many studies have sought to identify oscillating transcripts in an effort to connect the molecular clock to these downstream rhythms in biology. The majority of this previous work has focused on protein-coding transcripts oscillating in one or two tissues. To characterize the role of the circadian clock in mouse physiology and behavior, we used RNA-seq and DNA arrays to quantify the non-coding transcriptomes of twelve mouse organs over time. We found that the expression of more than one thousand known and novel ncRNAs oscillate and are timed throughout the day. Supporting their potential role in mediating clock function, those ncRNAs conserved between mouse and human oscillate at the same rate as protein encoding genes. Furthermore, these conserved ncRNAs cover a broad range of functional groups, including potential regulators of rRNA biogenesis and cardiac health. The highly tissue-specific nature of the oscillations in these ncRNAs indicates they are involved in highly-specialized functions. Lastly, the broad scale of this data will provide an excellent resource for those investigators interesting in studying ncRNA expression at the system-level.

3.2 Introduction

Circadian rhythms are endogenous 24-hour oscillations in behavior and biological processes found in all kingdoms of life [118]. This internal clock allows an organism to adapt its physiology in anticipation of transitions between night and day. The circadian clock drives oscillations in a diverse set of biological processes, including sleep, locomotor activity, blood pressure, body temperature, and cellular redox state [2, 3, 7, 119]. Disruption of normal circadian rhythms leads to clinically relevant disorders including neurodegeneration and metabolic disorders [6, 120]. In mammals, the molecular basis for these physiological rhythms arises from the interactions between two transcriptional/translational feedback loops [15, 16]. Many members of the core clock regulate the expression of other transcripts. These clock-controlled genes mediate the molecular clock's effect on downstream rhythms in physiology.

In mammals, the core oscillator is located in the suprachiasmatic (SCN) nucleus of the hypothalamus. Other tissues contain peripheral oscillators capable of driving local rhythms in CCG transcription [1]. These peripheral oscillators receive input from the SCN in order to remain coordinated across the entire body and entrained to the correct time of day [24].

In an effort to identify new clock factors, and to study circadian physiology in different organisms, researchers have devoted significant time and effort to studying transcriptional rhythms in the SCN and the periphery [34, 40, 44, 49, 51, 121]. While these studies have traditionally been performed with microarrays, recent studies have begun to adopt RNA-seq as the primary method for transcriptional analysis. This has the advantage of allowing researchers to identify novel circadian transcripts in the mammalian system, including long-intergenic non-coding (linc) RNAs, miRNAs, and antisense transcripts [44, 49, 51, 53]. While extremely informative, most circadian studies of this nature have analyzed one or two organs. This is particularly important for the characterization of circadian non-coding transcripts (ncRNAs), given the highly tissue-specific nature of their expression [122–124]. Given the emerging roles of ncRNAs in gene regulation, disease, as drug targets, and as molecular markers [125–129], it is important that we design studies capable of identifying ncRNAs and their patterns of expression. To address this, we used strand-specific RNA-sequencing (RNA-seq) and DNA arrays to profile the transcriptomes of twelve different mouse organs: adrenal gland, aorta, brainstem, brown fat, cerebellum, heart, hypothalamus, kidney, liver, lung, skeletal muscle, and white fat. We sampled organs every 6 hours by RNA-seq and every 2 hours by arrays.

3.3 Results

Constructing the list of conserved, non-coding transcripts

To identify conserved transcripts, we started with NONCODE v3 [130] annotations for mouse and human non-coding transcripts (33,801 human transcripts; 36,991 mouse transcripts). To prevent overlapping ncRNAs from confounding the analysis (many of these appeared to be alternative spliceforms of the same ncRNAs), we merged all overlapping ncRNAs on the same strand. Next, we aligned the human and mouse transcripts against each other using BLAST [131]. Since ncRNAs have previously been shown to have relaxed constraints on sequence conservation [124], we ran *blastn* using the more permissive dc-megablast algorithm and a minimum e-value cutoff of 1E-10. We then mined these BLAST results for pairs of human and mouse ncRNAs that were each other's top BLAST hit (termed "reciprocal best hits"). Filtering for these reciprocal best hits left us with 1601 human and mouse transcript pairs, we termed conserved ncRNAs. We are confident in our ability to identify conserved ncRNAs using these relaxed BLAST parameters as we successfully found well-known, conserved ncRNAs like *Xist*, *Tsix*, *Hotair*, *H19*, and *Gas5*.

To assign genes names and annotation data to these conserved transcripts, we aligned them to human and mouse RefSeq transcripts. At this point, we found that 585 of these conserved transcripts aligned to in the sense orientation to protein-coding genes in both humans and mice. Upon visual inspection of these ncRNAs, we found that many of these mapped along the entire length of the protein-coding transcripts. While some ncRNAs in this list might represent non-coding isoforms of these protein-coding transcripts, we chose to take a conservative approach and removed these from further analysis. Following the removal of these transcripts, we were left with a final list of 1016 conserved ncRNAs. Once again, our conservative approaches to generating this list of ncRNAs still retained well-known, conserved ncRNAs. This process is summarized in Fig. 3.1.

Assign functional groups to conserved ncRNAs

In order to assign basic functional categories to this list of conserved ncRNAs, we assigned biotypes (defined by GENCODE [132] and Ensembl [115]) to these transcripts using both the Ensembl annotation and manual curation (for full details, please see section 3.5 Methods). Briefly, we mapped the Ensembl biotypes to our list of conserved ncRNAs. Next, we identified miRNA host genes by checking for overlap between the genomic coordinates of the conserved ncRNAs, and the latest miRNA annotation from miRBase [97]. Following this, there were cases where a small number of ncRNAs mapped to related functional groups, like different categories of pseudo gene. In these instances, we collapsed the related functional groups together (eg. collapsing "snRNA" and "snoRNA" into "snoRNA_host"). Lastly, we assigned any ncRNAs from the "protein_coding" functional group to the "non-coding_isoform" group. We previously filtered out all transcripts aligning in sense orientation to protein-coding transcripts during construction of the list of conserved ncRNAs. Thus, we consider it most likely that any remaining genes from the "protein_coding" functional group represent non-coding isoforms of protein-coding genes. The breakdown of conserved ncRNAs by functional group is presented in Fig. 3.2.

Rhythmic, conserved, non-coding RNAs

We sought to leverage our stranded RNA-seq data to find oscillating ncRNAs. The multi-organ nature of our data makes it particularly well-suited for this purpose, due to the high degree of tissue-specificity in ncRNA expression [122–124]. To identify oscillating ncRNAs, we quantified both the conserved and non-conserved ncRNAs in our RNA-seq data across all twelve mouse organs, and analyzed this expression data using JTK_CYCLE [91]. For comparison purposes, we also quantified and analyzed protein-coding transcripts across the entire RNA-seq dataset. We found that a higher percentage of conserved ncRNAs showed circadian expression, when compared to non-conserved ncRNAs (Fig. 3.3A). Furthermore, the percentage of circadian, conserved ncRNAs was very much in line with the protein-coding transcripts, suggesting that these ncRNAs are functionally relevant, and that they may be regulated by the clock. We also found that individual ncRNAs oscillate in, at most, five organs. This is likely a function of two factors: 1) lower sampling resolution (6-hour) in the RNA-seq data limited our power to identify all circadian transcripts (this is also the reason why we see protein-coding transcripts oscillating across fewer organs in the RNA-seq data, relative to the array data) and 2) ncRNAs are highly tissue-specific in their expression [122–124].

The conserved, circadian ncRNAs covered a diverse set of functional groups (Fig. 3.3B). There were 38 conserved ncRNAs antisense to coding genes, half of which were antisense to coding genes that oscillated themselves. For example, *Galt*, a critical enzyme in the galactose-metabolism pathway, had an antisense transcript that oscillated in phase with its sense transcript, in liver (Fig. 3.4A, B). While *Galt's* antisense transcript oscillated in phase, there did not appear to be any relationship between the phases of sense and antisense expression, across all of these antisense ncRNAs. We also find four circadian snoRNA host genes: *Cbwd1*, *Snhg7*, *Snhg11*, and *Snhg12*. Taking this last one as an example, *Snhg12* oscillated in both brown adipose and hypothalamus (Fig. 3.4C, D). Previous work has already found that host genes for U snoRNAs show light-driven oscillations in *Drosophila* brains [48]. Given that our collection was performed in DD conditions, our data provides further evidence of the clock's potential to influence ribosome biogenesis.

In addition to snoRNA host genes, we also found 30 circadian, miRNA host genes, each of which provides a possible avenue for the clock to regulate downstream physiological processes. For example, M*ir22* is predicted to target *Ptgs1/Cox1* (prediction by TargetScan [133, 134]), an NSAID target implicated in the reduced incidence of myocardial infarction due to low-dose aspirin treatment [135]. The host transcript for M*ir22* oscillates antiphase to *Ptgs1* in both heart and lung (Fig. 3.5). Given that incidence of myocardial infarction shows a circadian rhythm [2, 136], it could be possible this effect is regulated in part by oscillations in *Mir22hg* and *Ptgs1*. Taken together, these cases provide only a few examples of the different biochemical pathways the clock could manipulate through regulation of non-coding transcripts. Annotation data and the peak phase in expression for each conserved ncRNA are listed in Supplementary digital file S1 (see Appendix A for details).

Novel antisense transcripts

In addition to the antisense transcripts we identified from the list of conserved ncRNAs, we also identified novel antisense loci. Briefly, we divided each gene into 1 KB tiles. We quantified read counts on both the plus and minus strand for each of these tiles. Next, we used expression data from the sense orientation to calculate background expression levels for each locus. For example, given a plus-strand transcript, we used all plus-strand reads mapping to intronic tiles to calculate the background level of expression. We identified all antisense tiles exceeding 10x this background level of expression. In order to focus on novel antisense transcription events, we filtered out all tiles overlapping known transcripts on the same strand. Lastly, we required all antisense transcripts consist of at least 3 adjacent tiles with expression above background. This procedure yielded a final list of 1,979 genes with un-annotated antisense transcripts, 187 of which showed sense and antisense oscillations in the same organ. Of these, 43 antisense transcripts oscillated at least eight hours out of phase with their sense transcripts. These antiphase oscillators may be the most likely antisense loci to have functional consequences. Genes with antiphase, antisense oscillators included *Arntl* and *Per2* (Fig. 3.6). This previously-identified *Per2* antisense transcript [44, 51] oscillated in five organs, the most of any antisense transcript, providing further evidence of its functional relevance. While it is currently unclear what functional role these novel antisense transcripts play, they may provide the circadian clock with the ability to fine-tune transcription of their sense transcripts. Annotation information and phase differences between sense and antisense transcripts are listed in Supplementary data file S2 (see Appendix A for details).

Identifying putative lincRNAs

Given that RNA-seq data is not limited to a specific gene annotation, we sought to characterize novel lincRNAs. We began by collecting all reads that mapped across splice junctions (ie. reads with large gaps in their alignments). While this will cause us to miss single-exon transcripts, we have greater confidence that the data comes from a real, expressed transcripts if we see evidence of RNA splicing. To reduce the impact of spurious reads and noise, we required that splice junctions be mapped by a minimum of 16 reads across our entire dataset (this corresponds to 2 reads per time point in a single organ). We chose a fairly low threshold so as not to remove junctions present in only a single organ, and those circadian transcripts expressed in a bursting patterns (like *Dbp*). Next, we filtered out any junction mapping within 1KB of any known gene, or overlapping any NONCODE transcript. All of these steps left us with 10,452 junctions from putative transcripts. We merged all junctions within 500 bp of each other to form 5,154 putative, ncRNA transcript regions.

Of these 5,154 spliced, transcribed, putative ncRNAs, 712 oscillated in at least one organ. The percentages of these putative ncRNAs oscillating in multiple organs were much closer to those of the non-conserved ncRNAs, than to the conserved ncRNAs and protein-coding genes (Fig. 3.3A). This is likely because the list of putative ncRNAs contains both conserved and non-conserved transcripts. Taking two of these novel ncRNAs as an example (originally characterized in chapter 2), we saw from the RNA-seq coverage data that they were clearly expressed (Fig. 3.7). The transcript from chr7 even appeared to have discrete coverage in exon blocks (Fig. 3.7A). Both of these transcripts showed oscillations across multiple tissues. The chr7 transcript oscillated in phase in both liver and brown adipose (Fig. 3.7B). Interestingly, the chr6 transcript oscillated antiphase in adrenal gland and liver (Fig. 3.7D). While neither of these transcripts have known functions, or ORFs, they are clearly spliced and showed rhythmic expression. These putative transcripts may represent completely novel genes and CCGs, and provide the circadian clock with as yet unknown means for controlling downstream processes and pathways. The full list of novel, oscillating transctips, their genominc coordinates, and the tissues in which they oscillated are listed in Supplementary digital file S3 (see Appendix A for details).

3.4 Discussion

In this study we used RNA sequencing and DNA microarrays to characterize circadian oscillations in non-coding transcript expression across twelve mouse organs. We found a functionally diverse set of ncRNAs with rhythmic oscillations. The rhythms of these conserved ncRNAs tended to be highly organ-specific, which likely stems from the known expression patterns of most ncRNAs [122–124]. Those ncRNAs conserved between human and mouse oscillated with the same frequency as protein-coding genes, suggesting their functional importance. While some of these rhythmic ncRNAs have recognized functions, like snoRNA and miRNA host genes, little is known about the majority. As annotation efforts move forward, these oscillating ncRNAs may provide the most likely candidates for functional relevance. Furthermore, the oscillations of these ncRNAs may prove advantageous for functional studies e.g. by linking a cycling miRNA to its predicted target genes by comparing their cycles.

It is important that we continue to identify and characterize these ncRNAs, as mounting evidence suggests that they serve important functional roles in many biological and disease pathways [125–129]. Furthermore, the clock itself appears to be no exception for regulation by miRNAs and lincRNAs [52, 106, 107]. Not only does this functional importance suggest some ncRNAs are mediators of the clock's control of physiology, but it makes them potential candidates as drug targets. For example, the aspirin target, *Ptgs1*, shows rhythmic expression in the heart and lungs. We found the host gene for M*ir22* oscillates antiphase to *Ptgs1*, its predicted target, in these same tissues. Furthermore, mice with a *Mir22* deletion show increased sensititivity to cardiac stress [137]. Not only does this suggest *Mir22* may regulated *Ptgs1* expression, but it also makes *Mir22hg* itself a potential drug target. This is important as *Ptgs1* is the direct aspirin target implicated as the cause for the cardioprotective properties of low-dose aspirin therapy [138]. Given that *Mir22* targets *Ptgs1*, without targeting *Ptgs2*/*Cox2*, it may provide a means to fine-tune expression between these two important genes.

In addition to examining known ncRNAs, we have expanded the list of potential clock regulators and outputs genes by identifying several hundred novel, circadian lincRNAs and antisense transcripts. These include a *Bmal1* antisense transcript, as well as the *Per2* antisense transcript indentified by previous studies [44, 51]. Taken together, we hope this work will provide an excellent resource for investigators in the clock and ncRNA fields. This also serves to highlight the vast, relatively untapped potential of ncRNAs as a source for clock output genes.

3.5 Methods

Circadian tissue collection

Mice were prepared for tissue collection as previously described [40]. Briefly, 6-week old male C57/BL6 mice were acquired from Jackson Labs, entrained to a 12h:12h light:dark schedule for one week, then released into constant darkness. Starting at CT18 post-release, three mice were sacrificed in the darkness every 2h, for 48 hours. Specimens from the following organs were quickly excised and snap-frozen in liquid nitrogen: aorta, adrenal gland, brainstem, brown adipose (anterior dorsum), cerebellum, heart, hypothalamus, kidney, liver, lung, skeletal muscle (gastrocnemius) and white adipose (epididymal). Food and water were supplied *ad libidum* at all stages prior to sacrifice. All procedures were approved by the Institutional Animal Care and Use Committee.

Microarray analysis

Tissue samples were homogenized in Trizol reagent (Invitrogen) using a Tissuelyser homogenizer (Qiagen). RNA was extracted using RNeasy columns according to the manufacturer's protocol (Qiagen). RNA abundances were quantified using Affymetrix MoGene 1.0 ST arrays and normalized using Affymetrix Expression Console software (GC-RMA). Probesets on the Affymetrix MoGene 1.0 ST array were cross-referenced to best-matching gene symbols using Ensembl BioMart software, then filtered for known protein-coding status. The resulting 19,788 genes formed the protein-coding background set. These protein-coding genes were tested for oscillations with a period of 24-hours using the JTK_CYCLE [91] package in R.

RNA-seq library preparation and sequencing

Pooled RNA samples across all tissues from CT22, CT28, CT34, CT40, CT46, CT52, CT58, and CT64 (96 samples total) were converted into Illumina sequencing libraries using the Illumina TruSeq Stranded mRNA HT Sample Preparation Kit, according to manufacturer's protocol. Briefly, 1 ug of total RNA was polyA-selected, fragmented by metal-ion hydrolysis, and converted into double-stranded cDNA using Superscript II (Invitrogen). Next, the cDNA fragments were subjected to end-repair, adenylation, ligation of Illumina sequencing adapters, and PCR amplification. Libraries were pooled together in groups of six and sequenced in a single lane of an Illumina HiSeq 2000 using the 100 bp paired-end chemistry (for a total of 16 lanes).

RNA-seq analysis

Fastq files containing raw RNA-seq reads were aligned to the mouse genome (mm9/NCBI37) using STAR [63], with default parameters. All RNA-seq quantification was performed using HTSeq (http://www-huber.embl.de/users/anders/HTSeq), run in stranded mode with default parameters. Protein-coding genes were quantified using the Ensembl gene annotation [115]. All quantification values were normalized using DESeq2 [139], and tested for oscillations with a period of 24-hours using the JTK_CYCLE [91] package in R.

Identifying conserved ncRNAs from NONCODE database

BED files listing genomics coordinates for human (33,801 transcripts) and mouse (36,991 transcripts) ncRNAs were downloaded from the NONCODE v3 database [130].

Overlapping ncRNAs from the same strand were merged using the BEDTools suite [140]. This merge step reduced the number of ncRNA to 20,042 human and 27,286 mouse transcripts. Using the coordinates for these merged transcripts and the UCSC Genome Browser [141], nucleotide sequences corresponding to each of these ncRNAs were downloaded in FASTA format. Next, separate human and mouse BLAST libraries were constructed from these ncRNA sequences by running the *makeblastdb* command with default parameters. Following this, the mouse ncRNA sequences were aligned against the human ncRNA BLAST library, and vice-versa. For this alignment, BLAST [131] was run with the following arguments: -evalue 1E-10 -max_target_seqs 1 -num_threads 5 -task dc-megablast -strand plus. Only ncRNAs with reciprocal best hits were retained for further analysis.

Sequences from these conserved ncRNAs, were aligned to human and mouse RefSeq [109] transcripts. For this alignment, BLAST was run with the following arguments: -evalue 1E-10 -max_target_seqs 1 -num_threads 5 -outfmt 6 -task dc-megablast. The best hit from this RefSeq alignment was used to assign a gene name and RefSeq ID to each conserved ncRNA. All ncRNAs with an assigned RefSeq ID beginning with NM or XM (ie. from a protein-coding gene) in both humans and mice were excluded from further analysis.

Assigning functional group to conserved transcripts

The GENCODE [132] and ENSEMBL [115] annotations define functional groups, or biotypes, for each transcript. Ensembl's Biomart interface (http://www.ensembl.org/biomart/) was used to generate files that mapped Ensembl gene and transcript biotypes to corresponding RefSeq transcript IDs. These files were generated for both human (GRCh37.p13) and mouse (GRCm38.p2) transcripts. Using these files, functional groups were mapped to each conserved ncRNA, giving preference to the biotype listed for the mouse genes, when conflicts arose.

To identify miRNA host genes, annotations for human (last updated 6/24/2013) and mouse (last updated 6/24/2013) miRNAs were downloaded from miRBase [97]. The BEDTools suite [140] was used to identify conserved ncRNAs that overlap the miRBase miRNAs. Biotypes for these ncRNAs were changed to "miRNA_host."

Finally, several biotypes were only present for a few ncRNAs. If there were related biotypes present, they were collapsed into single categories: 1) "processed_transcript" and "misc_RNA," were collapsed into "lincRNA." 2) "snoRNA" and "snRNA," were collapsed into "snoRNA_host." Also, the biotypes for known snoRNA host genes *Snhg11*, *Snhg12*, and *Snhg7*, were changed to "snoRNA_host." 3) All biotypes including the word "pseudogene" were collapsed into a single "pseudogene" biotype. 4) "nonsense_mediated_decay" and "protein_coding" were collapsed into "non-coding_isoform."

Identify antisense transcription

A list of gene regions was created by taking the start and stop coordinates for each gene in the Ensembl annotation. All gene regions sharing the same gene name were merged into single loci representing all spliceforms for a given gene. This yielded 37,310 gene regions. A tiled annotation was created by adding 5 KB to the beginning and end of each gene region (to represent the promoter and 3' trailing regions), and then dividing the resulting region into 1 KB tiles. The operation started at the 5' end of each gene region, and continued sequentially until the end of the transcript. If the number of nucleotides in the gene region was not evenly divisible by 1000, the last tile's length was less than 1 KB. Reads mapping to the plus- and minus-strands for each of these tiles were quantified separately, using HTSeq (stranded mode with default parameters).

For each tile the background level of expression was calculate from the 10 closest tiles not overlapping exons (ie. intronic tiles). The background expression was calculated by taken the mean read count from these background tiles in the sense orientation, adding 1 (to prevent divide-by-zero errors in subsequent calculations), and taking the floor of this value. Next, only antisense tiles exceeding 10x their background level of expression were kept for further analysis. This yielded 85,111 tiles across 16,374 genes. To focus on novel antisense transcripts, all tiles overlapping a known Ensembl or NONCODE gene were marked for exclusion from future analysis. This reduced the number of tiles to 23,943 across 7,056 genes. Antisense transcripts were assembled from spans of three or more adjacent tiles above background expression. Any putative transcript containing at least one tile marked as overlapping a known gene were excluded from further analysis. This analysis yielded 2,643 putative, antisense transcripts covering 2,291 genes. Reads mapping to these novel transcripts were quantified using HTSeq (stranded mode, with default parameters), and tested for 24-hour oscillations using JTK_CYCLE. The phases for these antisense transcripts were compared to those of their overlapping, sense transcripts, derived from the microarray data. Some of the overlapping transcripts did not have corresponding array probes. This yielded a final list of 1,979 genes with novel antisense transcription events.

Identify putative lincRNAs

The STAR output files with the *SJ.out.tab* extension store splice junctions identified by reads with gapped alignments. All junctions with < 16 mapped reads across all tissues and samples were filtered out. Next, any junctions mapping within 1KB of any Ensembl or Refseq [109] transcript, or overlapping with any NONCODE transcript were removed using the BEDTools suite. Lastly, all junctions within 500 bp of each other were merged to form the final list of putative non-coding loci. Expression values within these transcript regions were calculated, normalized, and checked for oscillations as described above.

Data access

We deposited all sequencing data in the NCBI Gene Expression Omnibus (GEO) under accession number GSE54652. We have also added our data to the web interface we use for high-throughput circadian profiles (http://bioinf.itmat.upenn.edu/circa).

3.6 Figures

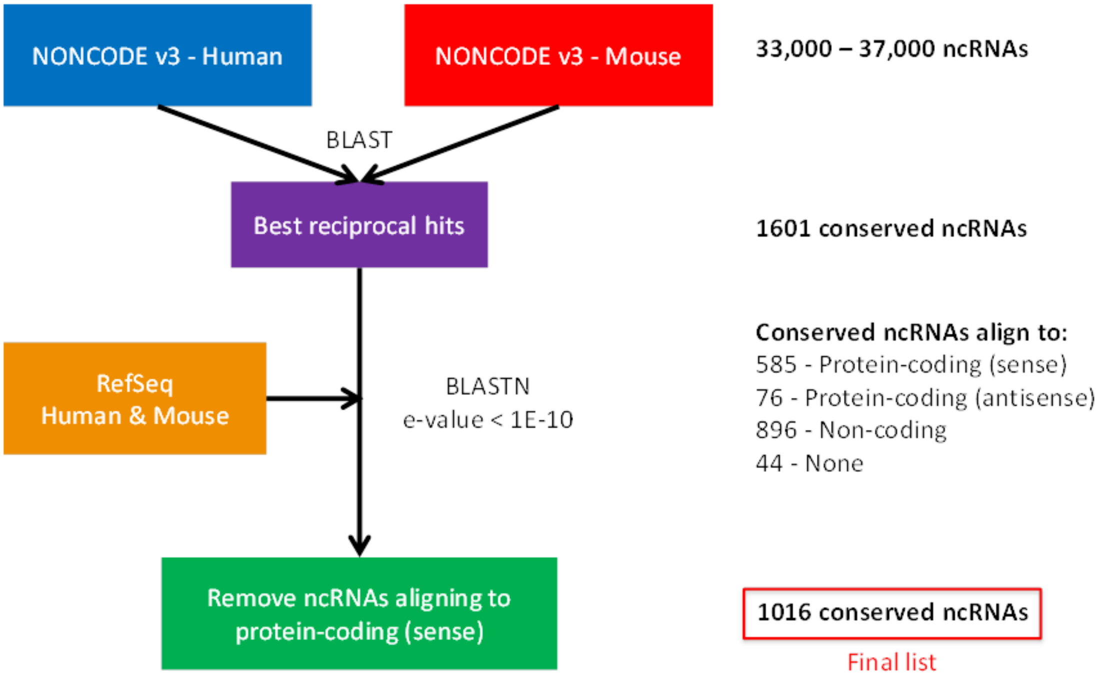

Fig. 3.1: Method overview for identifying conserved ncRNAs

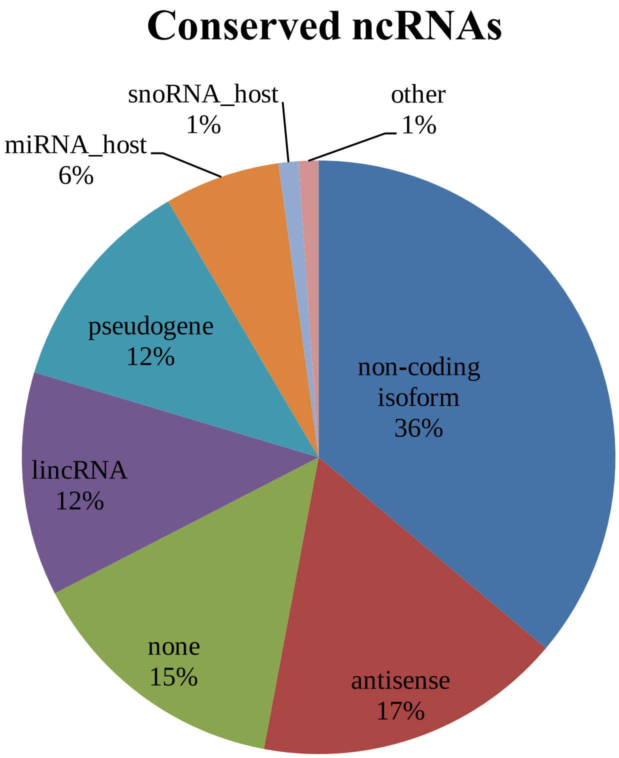

Fig 3.2: Functional groups for conserved ncRNAs

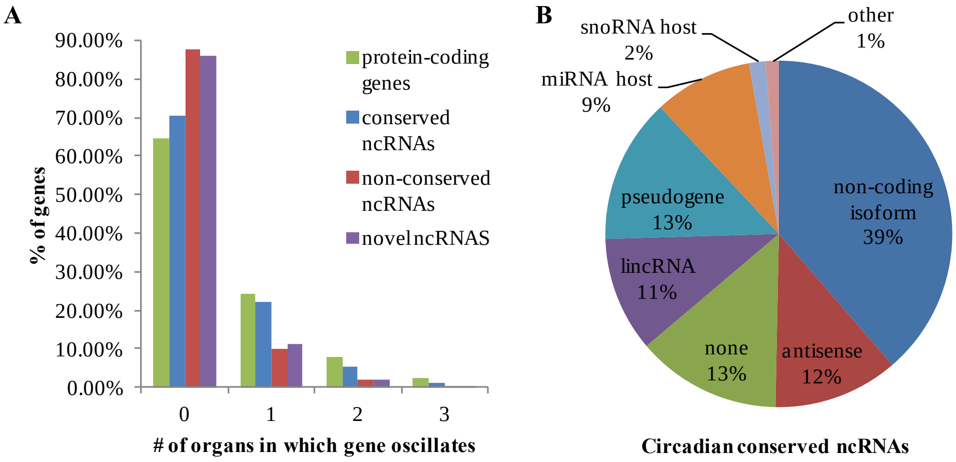

Fig. 3.3: Characteristics of rhythmic ncRNAs.

(A) Percentage of genes oscillating in the given number of organs. Data is displayed for protein-coding genes (green), conserved ncRNAs (blue), non-conserved ncRNAs (red), and novel, putative ncRNAs characterized in this study (purple). Note, this graph is cut off at a maximum of 3 organs. While there is data for genes oscillating in 4 and 5 organs, their numbers are so small that they are not readily visible on this graph. **(B)** Breakdown of functional groups for conserved ncRNAs with circadian expression.

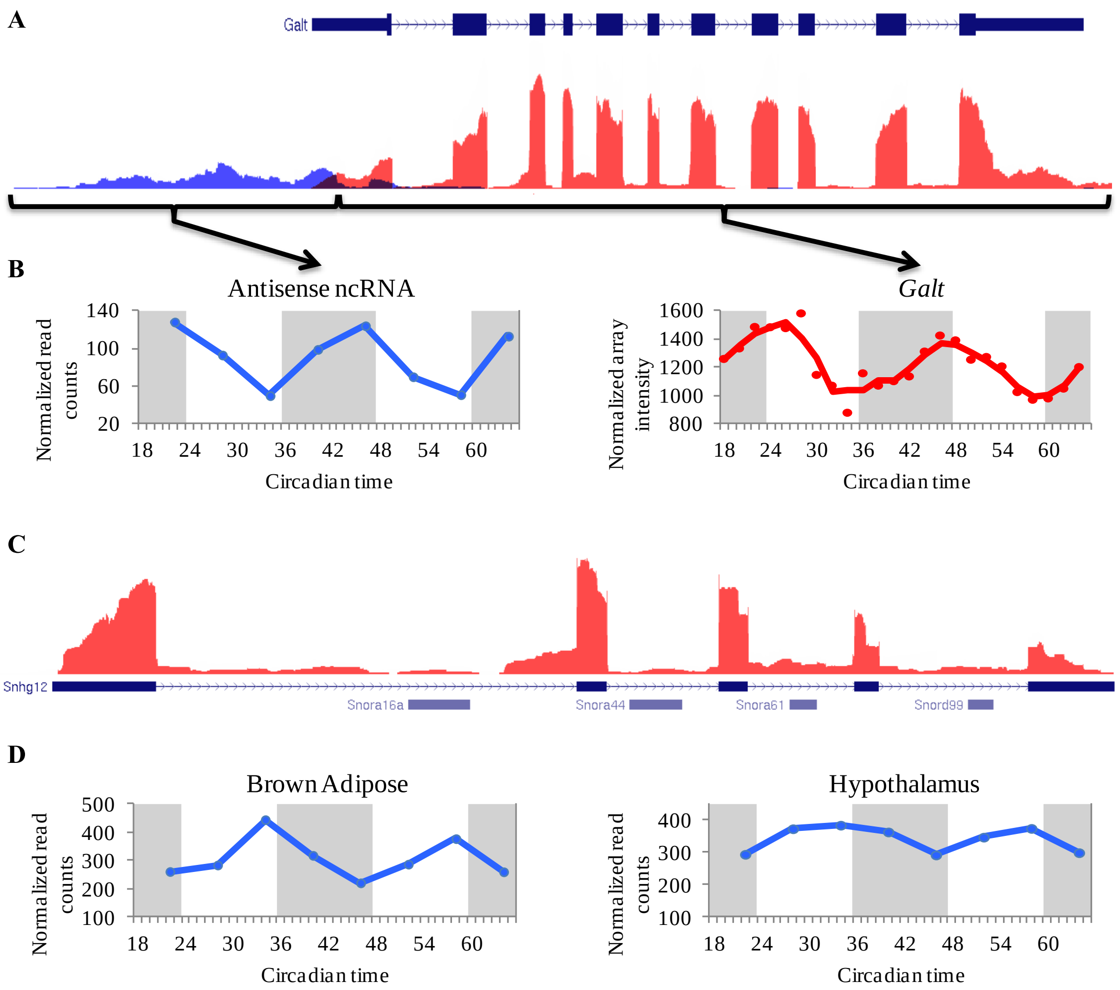
A
Galt
B
Antisense ncRNA
Normalized read counts
140
100
60
20
18 24 30 36 42 48 54 60
Circadian time
Galt
Normalized array intensity
1600
1400
1200
1000
800
18 24 30 36 42 48 54 60
Circadian time
C
Snhg12
Snora16a
Snora44
Snora61
Snord99
D
Brown Adipose
Normalized read counts
500
400
300
200
100
18 24 30 36 42 48 54 60
Circadian time
Hypothalamus
Normalized read counts
400
300
200
100
18 24 30 36 42 48 54 60
Circadian time

Fig. 3.4: Representative examples of conserved, oscillating ncRNAs.

(A) RNA-seq coverage plot for *Galt* (red) and its antisense transcript (blue). The gene model for *Galt* is displayed above the coverage plots. **(B)** Expression profiles for *Galt* (red; data from microarrays) and the antisense transcripts (blue; data from RNA-seq). Gray regions indicate subjective night. **(C)** RNA-seq coverage plot for *Snhg12*. The gene model is displayed below the coverage plot. Note the locations of the mature snoRNA sequences located in the introns of *Snhg12*. **(D)** RNA-seq expression profiles for *Snhg12* in brown adipose and hypothalamus.

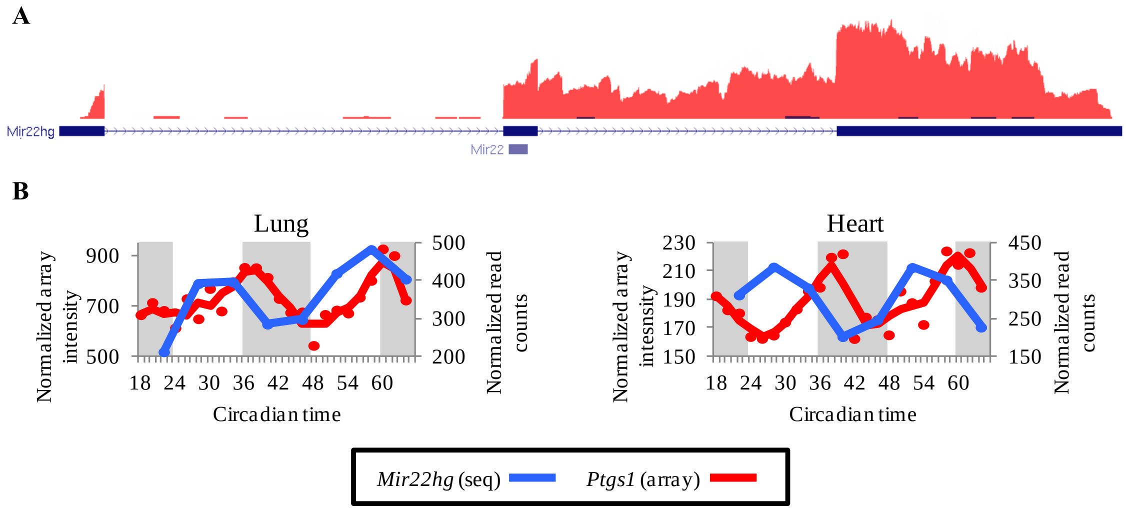

Fig. 3.5: *Mir22hg* expression is antiphase to its target *Ptgs1*.

(A) RNA-seq coverage plot for *Mir22hg*. The gene model is displayed below the coverage plot. Note the location of the mature *Mir22* sequence in the second exon of *Mir22hg*. **(B)** Expression profiles for *Mir22hg* (blue; data from RNA-seq) and its predicted target *Ptgs1* (red; data from microarrays), from lung (left) and heart (right) samples. The blue traces use the y-axes on the right, and the red traces use the y-axes on the left.

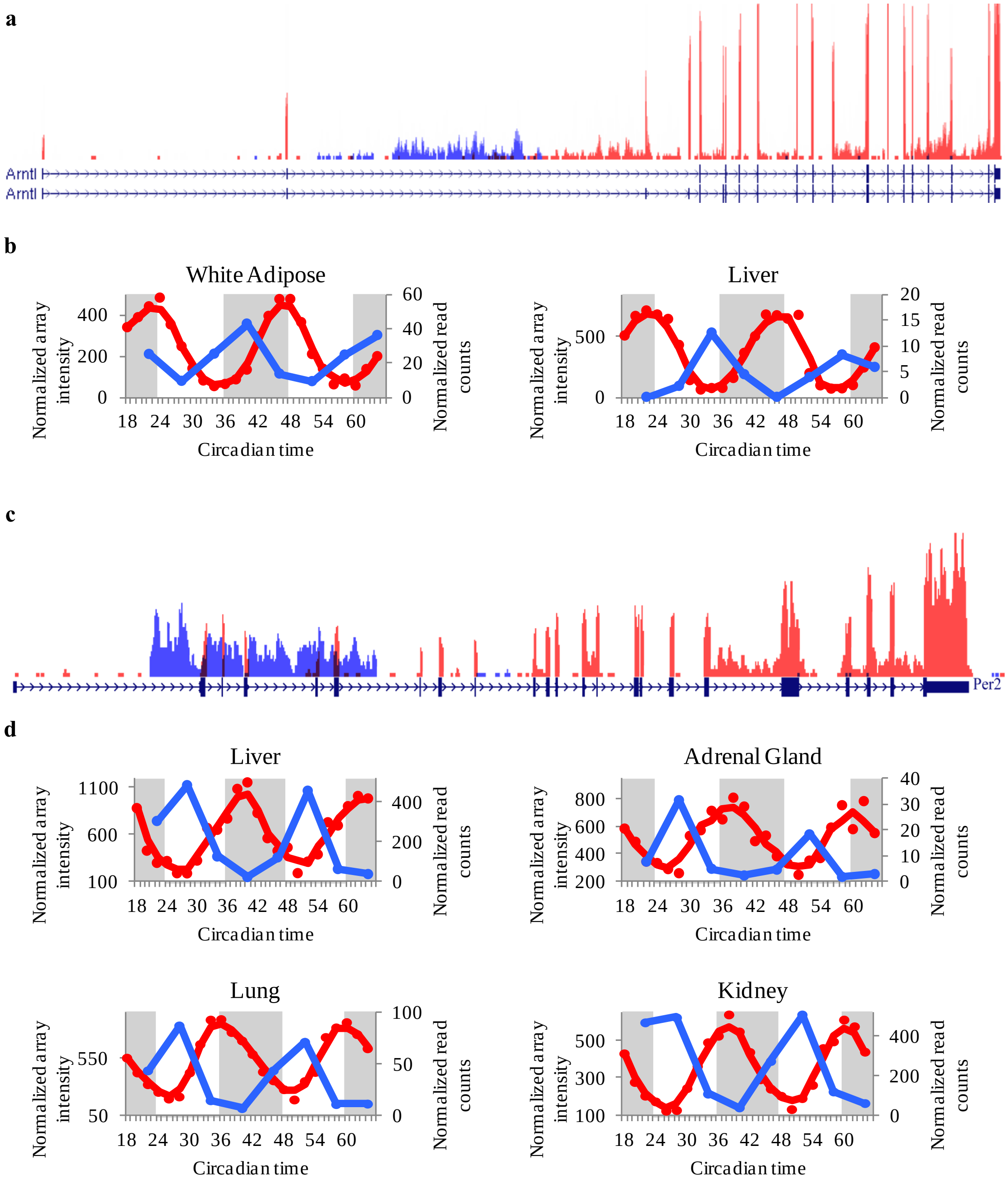
a
Arntl
Arntl
b
White Adipose
Normalized array intensity
400
200
0
60
40
20
0
Normalized read counts
18 24 30 36 42 48 54 60
Circadian time
Liver
Normalized array intensity
500
0
20
15
10
5
0
Normalized read counts
18 24 30 36 42 48 54 60
Circadian time
c
Per2
d
Liver
Normalized array intensity
1100
600
100
400
200
0
Normalized read counts
18 24 30 36 42 48 54 60
Circadian time
Adrenal Gland
Normalized array intensity
800
600
400
200
40
30
20
10
0
Normalized read counts
18 24 30 36 42 48 54 60
Circadian time
Lung
Normalized array intensity
550
50
100
50
0
Normalized read counts
18 24 30 36 42 48 54 60
Circadian time
Kidney
Normalized array intensity
500
300
100
400
200
0
Normalized read counts
18 24 30 36 42 48 54 60
Circadian time

Fig. 3.6: Antiphase, antisense transcripts of *Arntl* and *Per2*.

(A) RNA-seq coverage plot for *Arntl* (red) and its antisense transcript (blue), from white adipose. The gene model for *Arntl* is displayed below the coverage plots. **(B)** Expression profiles for *Arntl* (red; data from microarrays) and the antisense transcripts (blue; data from RNA-seq), from white adipose and liver. Gray regions indicate subjective night. **(C)** RNA-seq coverage plot for *Per2* (red) and its antisense transcript (blue), from liver. The gene model for *Per2* is displayed below the coverage plots. **(D)** Expression profiles for *Per2* (red) and the antisense transcript (blue) from liver, adrenal gland, lung, and kidney.

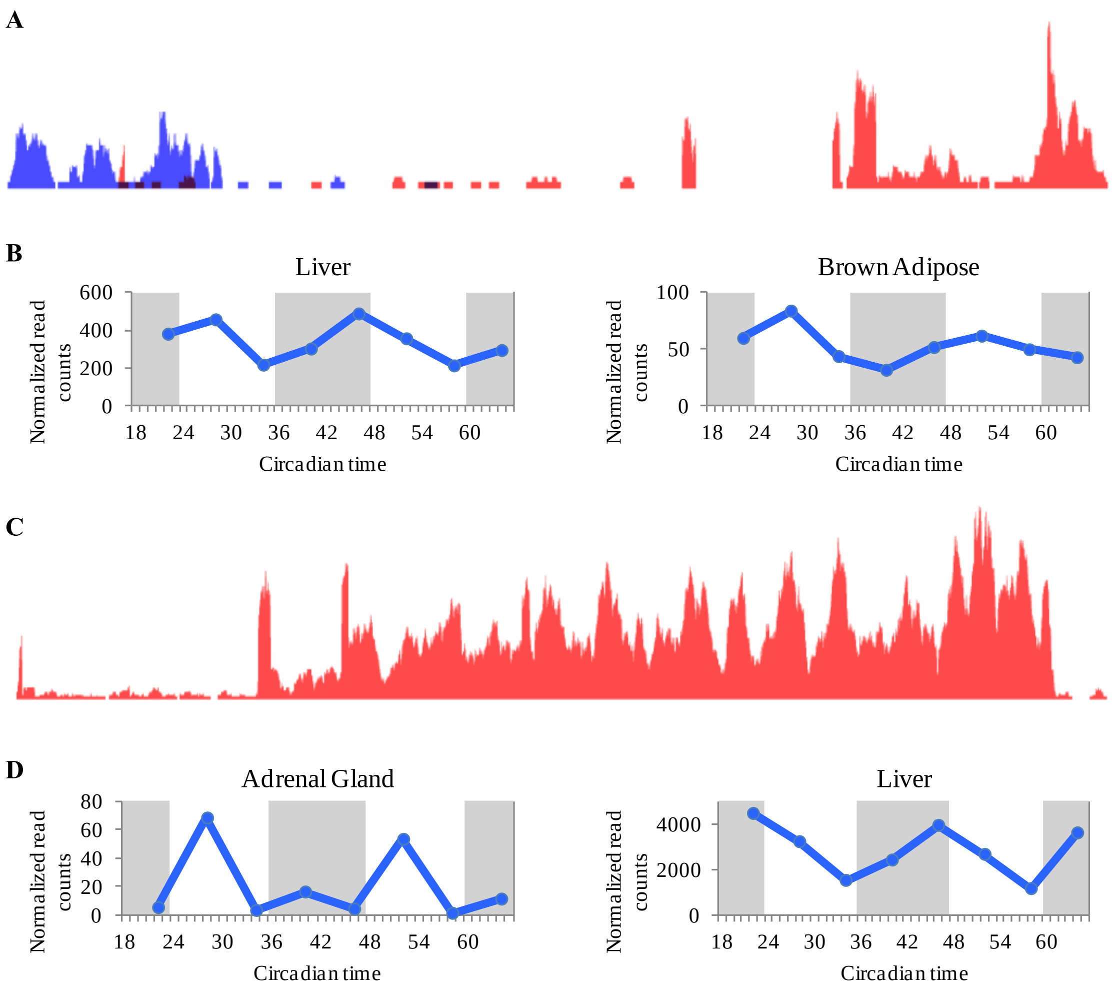

Fig. 3.7: Novel circadian ncRNAs.

(A) RNA-seq coverage plot for novel transcript located on chr7:35,913,467-35,928,124 (red). Note, there is some antisense transcription (blue) around the 5' end of this transcript. **(B)** RNA-seq expression profiles for novel chr7 transcript in liver and brown adipose. Gray regions indicate subjective night. **(C)** RNA-seq coverage plot for novel transcript located on chr6:121,086,416-121,114,417 (red). **(D)** RNA-seq expression profiles for novel chr6 transcript in adrenal gland and liver.

Chapter 4: IVT-seq reveals extreme bias in RNA-sequencing

4.1 Abstract

RNA sequencing (RNA-seq) is a powerful technique for identifying and quantifying transcription and splicing events, both known and novel. However, given its recent development and the proliferation of library construction methods, understanding the bias it introduces is incomplete but critical to realizing its value. Here we present a method, in vitro transcription sequencing (IVT-seq), for identifying and assessing the technical biases in RNA-seq library generation and sequencing at scale. We created a pool of > 1000 *in vitro* transcribed (IVT) RNAs from a full-length human cDNA library and sequenced them with poly-A and total RNA-seq, the most common protocols. Because each cDNA is full length and we show IVT is incredibly processive, each base in each transcript should be equivalently represented. However, with common RNA-seq applications and platforms, we find ~50% of transcripts have > 2-fold and ~10% have > 10-fold differences in within-transcript sequence coverage. Strikingly, we also find > 6% of transcripts have regions of high, unpredictable sequencing coverage, where the same transcript varies dramatically in coverage *between* samples, confounding accurate determination of their expression. To get at causal factors, we used a combination of experimental and computational approaches to show that rRNA depletion is responsible for the most significant variability in coverage and that several sequence determinants also strongly influence representation. In sum, these results show the utility of IVT-seq in promoting better understanding of bias introduced by RNA-seq and suggest caution in its

interpretation. Furthermore, we find that rRNA-depletion is responsible for substantial, unappreciated biases in coverage. Perhaps most importantly, these coverage biases introduced during library preparation suggest exon level expression analysis may be inadvisable.

4.2 Introduction

High-throughput sequencing of RNA (RNA-seq) is a powerful suite of techniques to understand transcriptional regulation. Using RNA-seq, not only can we perform traditional differential gene expression analysis with better resolution, we can now comprehensively study alternative splicing, RNA editing, allele specific expression, and identify novel transcripts, both coding and non-coding RNAs [58, 142, 143]. In contrast to the more established microarray based RNA expression analysis, the flexibility of RNA-seq has allowed for the development of many different protocols aimed at different goals (e.g. gene expression of poly adenylated transcripts, small RNA sequencing, total RNA sequencing, etc.). However, this same flexibility has the potential for complex technical bias, as different methods are routinely employed in RNA isolation, size selection, fragmentation, conversion to cDNA, amplification, and finally, sequencing [144–147]. While progress has been made in generating and analyzing RNA-seq data, we understand comparatively little about the technical biases the various protocols introduce. Understanding these biases is critical to differential analysis, to avoiding experimental artifacts (e.g. in characterizing RNA editing), and to realizing the full potential of this powerful technology.

Previous efforts at understanding bias identified several contributing sources, including GC-content and PCR enrichment [71, 72], priming of reverse transcription by random hexamers [73], read errors introduced during the sequencing-by-synthesis reaction [74], and bias introduced by various methods of ribosomal RNA (rRNA) subtraction [147]. Studies that revealed these sources of bias typically use computational methods on existing sequencing data to assess the performance of various sequencing technologies and library protocols. One downside to this approach is that it can be difficult to know whether anomalies in coverage are natural, or are due to technical artifacts. For example, nearly every RNA-seq study has differences in intra-exonal coverage, which could arise from naturally occurring splice variants sharing part of an exon, or could be due to technical error in library construction or sequencing.

Given that researchers are continually developing new sequencing methodologies and library generation protocols [148], we need a means for assessing the technical biases introduced by each new iteration in technology. One attractive alternative is to generate libraries from RNA that has been *in vitro* transcribed (IVT) from cDNA clones, where the nucleotide sequence at every base is known, the splicing pattern established and inviolate, and the expression level is known to be uniform across the transcript. Thus, any observed biases in coverage or expression must be technical rather than biological. This is the experimental equivalent of simulated data that computational researchers commonly use to develop and assess alignment algorithms [61, 149, 150]. Jiang and colleagues used a similar approach with 96 synthetic sequences derived from *Bacillus subtilis* or the deep-sea vent microbe *Methanocaldococcus jannaschii* genomes [151], organisms that do not have RNA splicing or poly adenylation. The focus of that work, though, was creating a useful set of standards that could be used in downstream analysis, not exploring library construction bias in a comprehensive set of complex mammalian samples.

Here we present and apply IVT-seq at scale to better understand bias introduced by RNA-seq. In brief, individual plasmids were produced, pooled, and subjected to in vitro transcription. Next, this RNA was mixed with complex mouse total RNA at various concentrations, and sequenced using the two most common RNA-seq protocols, polyA seq or total RNA seq, on the Illumina platform. We find coverage bias in most IVT transcripts, with over 50% showing > 2-fold changes in within-transcript coverage and 10% having > 10 fold differences attributable to library preparation and sequencing. Additionally, we find > 6% of IVT transcripts contain regions of high, unpredictable sequencing coverage, which vary significantly between samples. These biases are highly reproducible between replicates and suggest that exon-level quantification may be inadvisable. Furthermore, we created sequencing libraries from the original plasmid templates and using several different RNA selection methods (rRNA depletion, polyA selection, and no selection). We find that both rRNA depletion and polyA selection are responsible for a significant portion of this coverage bias, and computational analysis shows that poorly represented regions of transcripts are associated with low complexity sequences. Taken together, these results show the utility of the IVT-seq method for characterizing and identifying the sources of coverage bias in sequencing technologies.

4.3 Results and discussion

IVT-seq library preparation and sequencing

To generate IVT-seq libraries (for full details, please see the section 4.5 Methods), individual glycerol stocks each harboring a single, human, fully sequenced plasmid from the Mammalian Gene Collection [152] were produced and plasmid DNA was extracted and plated at 50 ng per well in 384-well plates. The contents of three 384-well plates containing a total of 1062 cDNA clones (Appendix D) were mixed, transformed into bacteria, and plated as single colonies. These plates were scraped, amplified for a few hours in liquid culture, and purified as a pool (Fig. 4.1A). Next, plasmids were linearized, purified, and SP6 polymerase was used to drive *in vitro* transcription of the cloned cDNA sequences (Fig. 4.1B). Following a DNase I treatment to remove the DNA template and RNA purification, a pool of 1062 different human RNAs derived from fully sequenced plasmids was produced.

To approximate what happens in a total RNA sequencing reaction, we subjected this IVT RNA to rRNA-depletion and then prepared libraries using the Illumina TruSeq protocol (Fig. 4.1C, IVT only). To account for possible carrier effects, we also mixed the IVT RNA with various amounts of mouse total RNA derived from liver. The addition of the mouse RNA gave these samples greater diversity (transcripts from ~10k genes vs. 1062) and more closely resembled a real biological sample. Also, by adding background RNA from a different species (mouse) than the IVT RNA (human), we make it easier to differentiate between the IVT transcripts and mouse sequences during downstream analysis. Since the IVT RNA does not contain rRNA sequences while the mouse RNA does, the quantity of mouse RNA will be significantly reduced by the rRNA depletion step. In order to account for this we mixed IVT and mouse RNA such that following rRNA depletion we would have final pools with IVT:mouse ratios of 1:1, 1:2, and 1:10. Finally, to account for mouse RNAs potentially mapping to the human reference genome and our IVT sequences, we prepared a pool consisting of mouse RNA alone. We pooled the resulting six libraries and sequenced them using an Illumina HiSeq 2000. We performed this entire process in duplicate.

Mapping and coverage of IVT-seq data

Following sequencing and de-multiplexing, we aligned all of the data to the human reference genome (hg19) using the RNA-seq Unified Mapper (RUM) [61]. For all analyses, we only used data from reads uniquely mapped to the reference, excluding all multi-mappers (data contained in RUM_Unique and RUM_Unique.cov files). Of the 1062 original IVT transcripts, we found 11 aligned to multiple genomic loci, while 88 aligned to overlapping loci. To avoid any confounding effects in our analyses, we filtered those transcripts from all analyses, leaving us with 963, non-overlapping, uniquely-aligned IVT transcripts. We saw excellent correlation in expression levels between replicates (transcript-level R^2 between replicates > 0.95; Fig. S4.1A). Secondly, at least 90% of the 963 IVT transcripts are expressed with an FPKM ≥ 5 in all IVT-seq datasets except mouse only (Table 4.1). In the IVT-only samples, over 80% of the IVT sequences are expressed above 100 FPKM (Fig. S4.1B). Since we prepared the MGC plasmids and

IVT transcripts as pools, it is likely that the IVT transcripts showing low or zero coverage were initially present at low plasmid concentrations prior to the transformation and IVT steps. Using the IVT-seq technique, we are able to specifically detect the vast majority of the human IVT transcripts with high coverage in both the absence and presence of the background mouse RNA.

While we do see reads aligned to the human IVT transcripts in the mouse only data, these transcripts collectively represent ~2% of reads (Table 4.1). Those transcripts with higher coverage are likely the result of mouse reads aligning to highly similar human sequences. We excluded these sequences from our analyses.

Within-transcript variation in RNA-seq coverage of IVT transcripts

Consider first the IVT only data. Given that these transcripts were generated from an IVT reaction using cDNA sequences, this data is unaffected by splicing or other post-transcriptional regulation. Thus, most regions of transcripts should be "expressed" and present at similar levels. The exceptions would be repetitive sequences that map to multiple genome locations and may be poorly represented, and the ends of the cDNAs, which are subject to fragmentation bias. To account for this we created a simulated dataset which models the fragmentation process and which deviates from uniform data only by the randomness incurred by fragmentation. We generated two such datasets using the Benchmarker for Evaluating the Effectiveness of RNA-Seq Software (BEERS) [61]. The first dataset contains all of the IVT transcripts expressed at roughly the same level of expression (~500 FPKM). For the second, we used FPKM values from the IVT-only samples as a seed, creating a simulated dataset with expression levels closely matching real data (Fig. S4.2). These datasets are referred to as simulated and QM-simulated (Quantity Matched), respectively. The simulated data provides an ideal result, while the QM data allows us to control for any artifacts arising from expression level (eg. transcripts with lower expression may show more variability). Next, we aligned both simulated datasets using RUM, with the same parameters as for the biological data. Thus, both simulated datasets also serve as a controls for any artifacts introduced by the alignment (eg. low coverage in repeat regions). For full details on the creation of simulated data, see the section 4.5 Methods.

Using IVT data derived from the BC015891 transcript as a representative example, the ideal, theoretical coverage plot from the simulated data shows near-uniform coverage across the transcript's entire length, with none of the extreme peaks and valleys characteristic of biological datasets (Fig. 4.2A). However, our observed data shows a high degree of variability, with peaks and valleys within an exon (Fig. 4.2B). Furthermore, these patterns are reproducible across our replicates (Fig. S4.3). We see many other cases of extreme changes in coverage; over 50% of the IVT transcripts show > 2-fold changes in within-transcript coverage attributable to library preparation and sequencing (Table 4.2 and Fig. S4.4). For example, BC009037 shows sudden dips to extremely low levels of expression in both of its exons (Fig. 4.2C). Both simulated datasets show no such patterns, which indicates this coverage variability is not the result of alignment artifacts. Furthermore, the absence of this pattern in the QM-simulated data indicates these fold-change differences in coverage are not due to sampling noise introduced by transcripts with low or high coverage. In the case of BC016283, the peaks and valleys in coverage lead to greater than five-fold differences in expression levels between exons (Fig. 4.2D). Once again, these patterns are reproducible across replicates (Fig. S4.3). The SP6 polymerase cannot fall off and then re-attach at a later point in the transcript, leaving a region un-transcribed. Therefore, given that these patterns show troughs followed by peaks, they cannot be the result of artifacts from *in vitro* transcription. Furthermore, we sequenced the IVT products directly and found the vast majority were transcribed with little to no bias. Taken together, these data suggest that these coverage patterns are primarily the result of technical biases introduced during library construction, rather than biology. These results are consistent with a previous study that uses *in vitro* transcribed RNA as standards in RNA-seq experiments [151], suggesting that our IVT-seq methodology is suitable for identifying technical variability in sequencing data.

Between-sample variation in RNA-seq coverage of IVT transcripts

In addition to this variability within transcripts, we also find many transcript regions showing extreme variability in coverage across samples (Fig. 4.3). For example, the sixth exon of BC003355 varies wildly relative to the remainder of the transcript across all IVT:mouse dilutions. Interestingly, the overall pattern of variation relative to the rest of the transcript across the dilutions is maintained between the replicates. Almost no reads in the mouse-only sample map to this transcript, which eliminates the possibility that this variability is due to incorrect alignment of mouse RNA.

Including BC003355, we find 86 regions of high, unpredictable coverage (hunc) spread across 65 transcripts (Appendix E). Therefore, over 6% of the 963 IVT transcripts contain regions showing wild but reproducible variations in RNA-seq coverage between samples. While identifying these hunc regions, we used a two-stage filter to eliminate variable regions resulting from mouse reads mapped to highly similar human sequences. First, we eliminated all hunc regions coming from transcripts with FPKM >= 5 in either mouse-only dataset. Next, to account for localized misalignment of mouse reads, we filtered out all hunc regions with an average coverage >= 10 in either mouse-only dataset. We also removed those hunc regions with mouse-only coverage >= 10 in the flanking 100bp on either side. Given the stringent criteria we used to identify these hunc regions (for full details see section 4.5 Methods), it is likely that this is an underestimate. To address the possibility that mouse RNAs may interact with homologous human RNAs and interfere with them in *trans*, we assayed the sequences surrounding these regions using the MEME Suite [153], but we found no sequence motifs these regions have in common. Furthermore, the depth of coverage at these regions does not follow a linear relationship with the increasing mouse RNA, which suggests it is not simply a direct interaction with the background RNA. There is no clear cause for these hunc regions, particularly since we prepared all samples from the same pool of IVT RNA and the only difference between samples is the relative ratios of IVT RNA to mouse liver RNA. We also searched for hunc regions that were divergent between the two replicates, but found none. If such regions do exist, they could be identified and overcome by creating libraries in duplicate. The hunc regions we identified above with expression patterns maintained between replicates present a greater challenge, as they could not be identified and filtered out by creating duplicate libraries. This is particularly problematic for using exon-level expression values to identify alternative splicing events or differential expression. The within-transcript and between-sample variation we see in our IVT-seq data suggests that library generation introduces strong technical biases, which could confound attempts to study the underlying biology.

Sources of variability in RNA-seq coverage

There are three potential sources for technical bias in library preparation: RNA-specific molecular biology (i.e. RNA fragmentation, reverse-transcription), RNA selection method (i.e. rRNA-depletion, polyA selection), and sequencing-specific molecular biology (i.e. adapter ligation, library enrichment, bridge PCR). To identify biases introduced solely by sequencing-specific molecular biology, we created a DNA-seq library from the same MGC plasmids used as templates for the IVT-seq libraries (Fig. S4.5). In doing this, we skip the steps specific to the IVT or RNA molecular biology. We also prepared two additional IVT-seq libraries using polyA selection or no selection, instead of rRNA depletion. By comparing our plasmid library data and the IVT-seq data using various selection methods, we can identify which coverage patterns are the result of

RNA-specific molecular biology, the RNA selection method, or of some common aspect of the library generation protocol.

We sequenced the plasmid library using an Illumina MiSeq and aligned the resulting data to the human reference genome using the same method as the IVT-seq libraries. In this plasmid data, we see 924 of the cDNA clone sequences with FPKM values ≥ 5, compared to ~870 in both of the IVT only samples (Table 4.1). This small drop in coverage is likely because the IVT RNA goes through more pooling steps during library construction than the plasmids. Furthermore, the plasmids are not affected by transcription and reverse transcription efficiencies. Additionally, the plasmid data maps to the cDNA sequences with an average, normalized coverage of 42.08, which is within the range of coverage values we see for the IVT-seq samples. We sequenced the no selection and polyA selection libraries on a HiSeq 2500. This data also shows cDNA clone coverage values similar to the other IVT-seq libraries.

The plasmid data represents the "input" into the IVT reaction and the no selection data represents the closest measure of its direct output. By measuring the 3'/5' ratio in depth of coverage for each IVT transcript, we can assess the processivity of the SP6 polymerase. In a perfectly processive reaction, this 3'/5' ratio would be 1, indicating the polymerase did not fall off the cDNA template and lead to the formation of truncated products. The median 3'/5' ratios for the plasmid and no selection data were 1 and 0.98, respectively, indicating premature termination of the IVT reaction was not a factor in our analyses.

Effect of different RNA selection methods on coverage patterns

Our analysis is illustrated by an examination of the coverage plots for BC003355 across all of our different datasets. The high degree of variability we noted in this gene's coverage plot from our rRNA-depleted data is absent in the no selection and plasmid data (Fig. 4.4A). While the polyA data also shows fewer peaks and valleys than the rRNA depleted total RNA-seq data, it is marked by the well-documented 3’ bias. This data suggests that the rRNA-depletion step is likely responsible for a large quantity of the observed coverage biases.

To quantify the variability for each selection method, we calculated the coefficient of variation at the single base level in coverage for all IVT transcripts across each of these datasets (Fig. 4.4B). Using a Wilcoxon rank-sum test (plasmid n = 924, no selection n = 870, rRNA-depleted n = 869), we find the rRNA-depleted data has significantly higher variability than the no selection and plasmid data ($p < 2.2e\text{-}16$). Furthermore, the rRNA-depleted and polyA libraries are > 60% more variable on average than the plasmid library (Fig. 4.4C). This suggests that a significant portion of the observed variability in coverage across transcripts in the IVT-seq data is the result of RNA-specific molecular biology, specifically the RNA selection step. Furthermore, after accounting for bias introduced by the sequences themselves (plasmid data) and bias introduced by the IVT reaction (‘no selection’ data), we find that 50% of transcripts have 2-fold and 10% have 10-fold variation in within transcript expression (Table 4.2 and Fig. S4.4). While it is well appreciated that polyA selection introduces bias, we found that rRNA-depleted data introduced just as much if not more. Neither simulated dataset showed transcripts with a 2-fold or higher change in within transcript expression. Again, this suggests that the observed within transcript variations are not the result of alignment artifacts or sampling due to low/high expression. One commonly acknowledged source of bias arises from random priming during library preparation [10]. When we examined the different libraries, we saw that fragments from all of the RNA-seq data showed nucleotide frequencies characteristic of random priming bias (Fig. S4.6). As expected, the plasmid data showed no such bias, since it was derived directly from DNA and required cDNA generation step. However, the significant differences between all RNA libraries suggest that bias from random priming is not the only factor. The plasmid and no selection data still contain a fair amount of variability when viewed alongside the simulated data (Fig. 4.4A; black). When we examine the entire dataset, both the plasmid and no selection data have significantly higher variation than either simulated dataset (Wilcoxon rank-sum test; simulated data n = 963, QM-simulated data n = 869, plasmid n = 924, no selection n = 870; $p < 2.2e-16$). This data suggests that sequencing-specific molecular biology common to all libraries we prepared (adapter ligation, library amplification via PCR), is also responsible for a portion of the observed coverage variability and sequencing bias.

Biases associated with sequence features are dependent on RNA selection method

Given these significant differences in coverage variability, we sought to identify sequence features that might contribute to this bias. We considered three quantifiable sequence characteristics: hexamer entropy, GC-content, and sequence similarity to rRNA (see *Materials and methods* for a detailed description of these metrics). For each sequencing strategy (plasmid, no selection, rRNA-depleted, polyA), we tested if any of the three sequence characteristics has a significant effect on variability in sequencing coverage, as measured by the coefficient of variation. While we are primarily focused on coverage variability as an indicator of sequencing bias, we also looked at depth of coverage, as measured by FPKM.

For each sequencing strategy we sorted the transcripts by coverage variability or depth. Next, we selected the 100 most and 100 least extreme transcripts from each list. We compared the values of the sequence characteristics between the 100 most and 100 least extreme transcripts using a Wilcoxon rank-sum test. Significant *p*-values indicate a significant association of the sequence characteristic with coverage variability and/or depth. The results of our analysis are displayed as box-plots (Fig. 4.5 and Fig. S4.8). To check for any confounding effects between coverage depth and variability, we tested the least and most expressed transcripts for any correlations with variability in coverage (Fig. S4.7). The polyA library showed a significant correlation (p < 2.2e-16) between coverage variability and depth, which indicates sequence features could be affecting coverage through variability (or vice versa).The rRNA-depleted data showed a slight, significant correlation (p = 0.04933). It is possible some feature of RNA selection affects both variability and coverage, given that we saw no significant correlations for the two remaining samples.. This indicates that coverage variability and depth are independent for the plasmid and no selection data.

All three sequence characteristics have a significant association with variability and depth-of-coverage in at least one of the sequencing strategies. In particular, lower hexamer entropy, a measure of sequence complexity [154–156], is strongly associated with higher variance in all of the RNA libraries (no selection p = 4.712e-05; rRNA-depletion p = 3.956e-11; polyA p = 0.003921; Fig. 4.5A). This suggests that bias associated with hexamer entropy is due partially to RNA-specific procedures in library preparation. Furthermore, an association with lower hexamer entropy indicates there are more repeat sequences in the transcripts with higher variability. This could be indicative of complex RNA secondary structures, as repeated motifs could facilitate hairpin formation. Furthermore, the absence of this association from the plasmid data suggests that this observation is not due to mapping artifacts. The plasmid data contains the same sequences as the RNA-seq data, and would be subject to the same biases introduced by our exclusion of multi-mapped reads.

Higher GC-content is strongly associated with lower coverage variability in the no selection and polyA data (p = 5.627e-13; p = 4.914e-05; Fig. 4.5B), suggesting that the effects of GC-bias on within-transcript variability could arise, in part due to some RNA-specific aspects of library preparation. Also, it appears that GC-bias is not a significant contributing factor to either depth of coverage, or the extreme variability in the rRNA-depleted data. Meanwhile, lower GC-content is associated with higher coverage in the plasmid data (p = 3.776e-05), and lower coverage depth in the no selection and polyA libraries (no selection p = 8.531e-05; polyA p = 0.0009675; Fig. S4.8B). Given that this trend switches directions between the plasmid library and the RNA libraries, this also suggests that some RNA-specific aspect of library preparation is introducing GC-bias distinct from the high GC-bias associated with Illumina sequencing [157].

Interestingly, higher rRNA sequence similarity is associated with higher coverage variability in the rRNA-depleted library (p = 9.006e-05) and lower variability in the no selection library (p = 0.0367; Fig. 4.5C). It is unsurprising that similarity to rRNA sequences contributes to variability in the rRNA-depleted data, given that rRNA-depletion is based upon pair-binding between probes and rRNA sequences. While it is unclear why this trend is reversed in the no selection library, it is striking given the significant increase in within-transcript variability between the no selection and rRNA-depleted libraries (Fig. 4.4B). Furthermore, we see a slight but highly significant correlation (Pearson R^2 = 0.308452; p < 2.2e-16) between a transcript sequence's similarity to rRNA, and the magnitude of the difference in coverage between the no selection and rRNA-depleted libraries (Fig. S4.9). While the majority of the factors contributing to the extreme bias in sequence coverage we see in the rRNA-depleted data remain unclear, our data suggests this is could be partially due to depletion of sequences homologous to rRNA.

Taken together, all of our data demonstrates the utility and potential of the IVT-seq method to identify sources of technical bias introduced by sequencing platforms and library preparation protocols.

4.4 Conclusions

In this study, we present IVT-seq as a method for assessing the technical variability of RNA sequencing technologies and platforms. We created a pool of *in vitro* transcribed RNAs from a collection of full length human cDNAs, followed by high-throughput sequencing (Fig. 4.1). Since we know the identities and sequences of these IVT transcripts, and since they were created under conditions not affected by splicing and post-transcriptional modification, they are ideal for identifying technical biases introduced during RNA-seq library generation and sequencing. We used this method to demonstrate that library generation introduces significant biases in RNA-seq data, adding extreme variability to coverage and read-depth along the length of sequenced transcripts (Fig. 4.2). Our most striking finding is that over 50% of the IVT transcripts show > 2-fold differences in this within-transcript coverage attributable to library preparation and sequencing, in the polyA and rRNA-depleted data (Table 4.2). We prepared all RNA-seq libraries from the same pool of IVT RNA, so these differences are due to library construction and sequencing methods. Furthermore, 6% of the IVT transcripts contain regions with high unpredictable coverage variability (huncs) across different dilutions of IVT and mouse liver RNA (Fig. 4.3). We found it particularly concerning that these huncs are consistent between replicates, as this means these regions cannot be indentified and avoided by making replicate libraries. While the exact cause of this effect is unclear, it could be due to some trans interaction between different RNA (human IVT RNA and the background mouse liver RNA). If this is the case, it could prove difficult to account for, given the challenges we have already encountered making predictions for miRNA targets and RNA secondary structure. Based on these results, we strongly recommend caution in interpreting exon-level quantification data, particularly for identifying and quantifying alternative splicing events, without further understanding of these biases.

Using simulated data and by sequencing at various stages of the process (plasmids, unselected IVT RNAs, rRNA-depleted, and polyA selected), we found each step introduces bias. Regions of certain IVT transcripts are underrepresented in both DNA and RNA, suggesting something inherent in their structure may resist cloning and sequencing properly. The IVT reaction has its own biases, however, by and large, it worked extremely efficiently with 90%of the input templates producing transcripts at detectable levels. PolyA sequencing revealed the well described 3' bias. Finally, we saw extreme bias introduced by the rRNA-depletion step. Though we have yet to find the majority of the sources for this extreme bias, knowing that it occurs and that is at least partially due to rRNA sequence similarity is an important first step. By making this data available to the community, we hope that new experimental and analysis methods can be developed to account for the biases inherent in various aspects of RNA-seq.

Moreover, IVT-seq could be more broadly employed. By itself, the MGC collection has cDNAs derived from more than 16,000 mouse and human genes, including hundreds of genes for which there are more than one form. Therefore, in principle, it is possible to generate sequence profiles for representatives for nearly 2/3 of the mammalian transcriptome, or spike in datasets to develop new and better methods for splice form detection and quantification. Similar profiling approaches could do the same for other organisms. In addition, IVT-seq is also immediately relevant to RNA-seq method development, e.g. developing new protocols or refining existing ones. Finally, the method is not specific to Illumina sequencing and could be used to account for bias in other sequencing chemistries and methods (e.g. SOLID, Ion Torrent, PacBio, etc.).

Importantly, we are not suggesting that current generation RNA-seq is not a fantastic new technology or that quantification data is incorrect, particularly given the validated, reproducible results researchers have been able to gain through its use. Rather, we wish to provide a cautionary note that our understanding of this technology is still relatively new and incomplete. It is our hope that through the use of this data and IVT-seq, we will develop the means to minimize or account for bias in RNA-seq and truly realize the vision of digital gene expression.

4.5 Methods

Amplification of plasmid library

Glycerol stocks containing individual cDNAs (cloned into pCMV-Sport 6 plasmid) from the Mammalian Gene Collection [152], were produced and plasmid DNA was extracted and plated at 50 ng per well in 384-well plates. The contents of three 384-well plates (total of human 1062 transcripts; Appendix D) were collected as follows: 10 µl sterile dH_2O was added to each well and incubated at 37^oC for 10 min to resuspend plasmid DNA in water. Plasmid DNAs were collected and combined in 1.5 mL tube with aid of multichannel pipette and concentrated by ethanol precipitation. To amplify the library 10 ng of plasmid library was transferred into *E.coli* DH5α (Invitrogen catalog no. 18258-012) with heat shock method. Cells were incubated with plasmid library for 5 min on ice and were subjected to 42^oC for 30 sec. Then cells were transferred back to ice and incubated for 2 min. Next, 0.95 mL SOC medium was added to the cells and incubated at 37 ^oC for 1 h by shaking at 225 rpm. Cells were plated on LB-agar plates containing 100 µg/ml ampicilin. Plates were incubated for 16h at 37^oC to grow the colonies and 3500 (approx 3-fold of library size) colonies were collected with liquid LB. Cells were transferred into 100 mL liquid LB and incubated at 37^oC for 2 h. Plasmids were purified using Qiagen maxiperep kit (catalog no. 12163), according to the manufacturer's protocol.

In vitro transcription from plasmid library

Plasmids were linearized by NotI-HF enzyme so that the SP6 polymerase promoter site will be upstream of the sequences to be transcribed. Reactions consists of 5 U NotI-HF (NEB catalog no. R3189L), 5 µg library plasmid DNA, 1 X NEBuffer 4 (supplied with enzyme) and 90 µl of dH_2O. Reaction was incubated at 37^{o}C for 2 h to achieve complete digestion. The complete digestion of plasmid DNA was assessed by DNA gel electrophoresis. To eliminate NotI-HF and possible RNase in reaction mixture, samples was subjected to Proteinase K treatment. SDS and Proteinase K were added to the reaction mixture to a final concentration of 0.5% and 100 µg/mL, respectively. Sample was incubated at 37^{o}C for 30 min. After Proteinase K treatment, sample was subjected to the phenol/chlororform extraction, followed by ethanol precipitation. Pellet was dissolved in 50 µl of RNase-free water. Next in vitro transcription was carried out using MAXIscript® SP6 Kit (Ambion catalog no: AM1308). Reaction composed of 1 µg of library plasmid, 1X transcription buffer, 0.5 mM of NTPs (GTP,ATP, CTP, and UTP), 40 U of SP6 RNA polymerase and 10 µl of RNase-free water. Reaction was incubated at 37^{o}C for 30 min. Next, samples were treated with TURBO DNase to remove the plasmid templates. Briefly, 10 U of TURBO DNase (included with MAXIscript SP6 kit) were added to reaction mixture and incubated at 37^{o}C for 15 min. To stop the reaction 1 µL of 0.5 M EDTA was added. To remove unincorporated NTPs and other impurities sample was precipitated with ammonium acetate/ethanol. The following reagents were added to the DNase -treated reaction mixture: 30 µL RNase-free water to bring the volume to 50 µL, 5 µL 5 M Ammonium Acetate, and 3 volumes 100% ethanol. Sample was chilled at

-20^{o}C for 30 min and then centrifuged at maximum speed in a 4^{o}C table-top microfuge. The supernatant was discarded and pellet was washed with ice-cold, 70% ethanol. Pellet was dissolved in 50 µL RNase-free water and quality of RNA was assessed by agrose gel electrophoresis. In addition, PCR was carried out with in vitro transcribed RNA to confirm total depletion of plasmid DNA as well.

Mouse liver collection and RNA extraction

WT 6-week old male C57/BL6 mice were acquired from Jackson Labs. Mice were sacrificed and liver samples were quickly dissected and snap-frozen in liquid nitrogen. RNA was isolated from frozen mouse liver samples by TRIzol reagent according to manufacturer's protocol (Invitrogen catalog no. 15596-026). All animal experiments were performed in accordance with the approval of the Institutional Animal Care and Use Committee.

Construction and sequencing of RNA-seq library from IVT RNA

IVT RNA (2500 ng, 150ng, 75ng, 15 ng, and 0 ng) was pooled with mouse liver RNA (0 ng, 2350 ng, 2425 ng, 2485 ng, and 2500 ng respectively) to a final quantity of 2.5 µg. Each pool was split into two replicate samples of 1 µg each. RNA pools were treated with Ribo-Zero Gold kit (Epicentre catalog no. RZHM11106) and converted into Illumina RNA-seq libraries with the TruSeq RNA sample prep kit (Ilumina catalog no. FC-122-1001). Briefly, rRNA was removed from 1 ug of pooled RNA using Ribo-Zero Gold kit and purified via ethanol/sodium acetate precipitation according to manufacturer's protocol. After drying, the RNA pellet was dissolved in 18 µL of Elute, Prime, Fragment mix (provided with TruSeq RNA sample prep kit). RNA was fragmented for 8 minutes and 17 uL of this fragmented RNA was used to make the RNA-seq library according to Illumina TruSeq RNA sample prep kit protocol. After fragmentation/ priming, first strand cDNA synthesis with SuperScript II (Invitrogen catalog no. 18064014), second-strand synthesis, end-repair, a-tailing, and adapter ligation, the library fragments were enriched with 15 cycles of PCR. Quality and size of library was assessed using Agilent 2100 BioAnalyzer. The five libraries from each replicate were pooled together and sequenced using a single lane from an Illumina HiSeq 2000 (paired 100 bp reads).

Construction and sequencing of plasmid library

MGC plasmids were linearized by NotI-HF enzyme as before. These linearized plasmids were then fragmented using a Covaris S220 Focused-ultrasonicator. Briefly, 1.2 µg of linearized plasmid in a final volume of 60 uL of H_2O was loaded into a microTUBE (Covaris catalog no. 520045). The ultrasonicator was de-gassed and prepared according to manufacturer's protocol. Linearized plasmids were sonicated using the following conditions: intensity 5, duty factor 10%, cycles per burst 200, time 120s, and water bath temperature 7^{o}C. Fragmented plasmids were gel-purified using a 1% agarose gel (BioRad catalog no. 161-3107) and TAE running buffer (BioRad catalog no. 161-0743). Gel slice between 100 bp and 700 bp was excised and DNA was purified using MinElute gel extraction kit (Qiagen catalog no. 28606) according to manufacturer's protocol. Fragmented DNA was converted into a sequencing library using TruSeq DNA sample prep kit (Illumina catalog no. FC-121-2001). End repair, adenylation, adapter ligation, gel size-selection, and PCR enrichment were performed according to manufacturer’s protocol. During the gel size-selection, a band between 300 bp and 500 bp was excised. Quality and size of library was assessed using Agilent 2100 BioAnalyzer. This library was sequenced using an Illumina MiSeq (paired 100 bp reads).

Construction and sequencing of no selection and polyA libraries

As with the other RNA-seq libraries, these libraries were prepared using the TruSeq RNA sample prep kit (Illumina catalog no. FC-122-1001). For the polyA sample, 1 µg of IVT RNA was treated with polyA selection reagents included with the TruSeq RNA sample prep kit according to manufacturer's protocol. The remainder of the library preparation was carried out using the same conditions as for the other IVT RNA samples. For the no selection sample, 100 ng of IVT RNA at a concentration of 100 ng/µL was diluted with 17 µL of Elute, Prime, Fragment mix (provided with TruSeq RNA sample prep kit). Again, the remainder of the library preparation was carried out as with the other samples. These samples were sequence in a single Illumina HiSeq 2500 lane (paired 100 bp reads).

Aligning, quantifying, and visualizing sequencing data

Raw reads from all sequencing samples were aligned to the human genome (GRCh37/hg19) using the RNA-seq Unified Mapper [61] (RUM; v2.0.4) with default parameters. Mapping stats for all libraries are included in Table S4.1. RUM also generated RNA-seq coverage plots in bedgraph format, and calculated transcript- and exon-level FPKM values for each IVT transcript (accession numbers listed in Appendix D). All analyses were performed using uniquely aligned reads (no multi-mappers) from the RUM_Unique and RUM_Unique.cov output files. Quantification was performed using annotations for the IVT transcripts that we downloaded from the MGC Genes track [152] on the UCSC Genome Browser [158]. Those IVT transcripts mapping to multiple loci, or overlapping other IVT transcripts were removed from further analysis (marked with * in Appendix D). All coverage plots in this paper were visualized in and captured from the UCSC Genome Browser. All statistical tests and correlation plots were performed in R.

Generating simulated data

Simulated data was generated using the BEERS software package (http://www.cbil.upenn.edu/BEERS/) from gene models for IVT transcripts, with an average coverage depth of 1000 reads (10,000,000 reads total). All error, intronic read, and polymorphism parameters were set to zero. Remaining parameters used default values. For the QM-simulated (Quantity Matched) data, FPKM values from replicate 1 of the IVT-only data were used as seeds for generating expression levels (40,000,000 reads total). This generated simulated data with FPKM values closely matching those from the real data (Fig. S4.2B). All other parameters were the same as for the other simulated data.

Processivity analysis

Coverage data for each IVT transcript was extracted from coverage files for the plasmid and no selection samples. For each transcript, base pair-level coverage data was extracted from the regions spanning 5-15% and 85-95% of the transcript, by length. For example, given a 1000 bp transcript, the first region spans base pairs 50-150, and the second region spans base pairs 850-950. These two coverage regions represent the 5’ end and 3’ end of the transcript, respectively. The first and last 5% of the transcript was excluded to avoid artifacts from the fragmentation process. Processivity of each transcript was assessed by the ratio of the mean depth of coverage from both of these regions (3’ region mean / 5’ region mean). These processivity ratios were calculated for all transcripts in the plasmid and no selection data, with expression > 5FPKM.

Calculating fold-change difference in within-transcript coverage

Coverage data for each of the IVT transcripts was extracted from the coverage files for the IVT-only, polyA, and no selection samples. The first and last 200 bp were trimmed from each transcript to prevent edge effects from interfering with the calculations. Due to this trimming, all IVT transcripts with less than 500 bp were discarded. All IVT transcripts expressed with FPKM < 5 in any of the samples were discarded from further analysis. Nucleotide-level coverage data was grouped into percentiles based on depth of coverage. Average coverage across the 10th percentile and 90th percentile were calculated. Fold-change difference in within-transcript coverage were calculated by dividing the 90th percentile average by the 10th percentile average. The list of transcripts with associated fold-change values is included in Supplementary digital file S3 (see Appendix A for details).

Identifying hunc regions

Coverage data for each of the IVT transcripts was extracted from the coverage files from each of the rRNA-depleted datasets (replicate dilution series: IVT-only, 1 IVT: 1 mouse, 1 IVT: 2 mouse, 1 IVT: 10 mouse, and mouse-only). These coverage plots were normalized between 0 and 1 to allow comparison between different dilutions. For each nucleotide position in a transcript, the deviation in coverage between each of the samples was calculated using the median absolute deviation (MAD), due to its resistance to outliers. MAD scores were calculated across the different dilutions using R's *mad* function with constant=1. Next, a sliding window was used to calculate the average MAD in the 100 bp windows centered on each nucleotide in the transcript. The first 300 and last 250 windows were trimmed from each transcript to avoid confounding variability due to edge effects or fragmentation artifacts. All analysis up until this point was carried out separately on the two replicate datasets. The 95th percentile of MAD scores was calculated for each of the replicates using R's *quantile* function (replicate 1: 0.08810424, replicate 2: 0.07183765). Only those regions with at least 20 contiguous windows having MAD scores above the appropriate 95th percentile values were retained for further analysis. Next, the BEDTools [140] intersect function was used to remove any regions with high MAD scores not present in both replicates. Finally, these remaining regions of high coverage variability were filtered for mouse reads misaligned to the human reference. Any region coming from a transcript with FPKM >= 5 in the mouse-only samples were discarded. To account for localized misalignment of mouse reads, any regions with an average coverage > 10 in the mouse-only samples or in the 100 bp on either side of the region were discarded. These remaining regions comprise the list final list of regions with high coverage variability. To search for hunc regions not maintained between replicates, windowed MAD scores from replicate 2 were subtracted from those of replicate 1. The 2.5th and 97.5th percentiles of these difference values were used as cutoffs (2.5th percentile: -0.07053690, 97.5th percentile: 0.09134876) to pull out the most extreme positive and negative difference values. Regions corresponding to these extreme difference values were filtered as above. Additionally, those difference regions within 200 bp of a previously identified hunc regions were filtered out. This last filtering step accounts for cases where a difference region with high MAD scores is just an extension of an existing hunc region. Hunc regions and difference regions were manually checked to determine whether or not they represent regions where expression patterns deviate from the remainder of the transcript.

Generating sequence characteristics

Sequences for each transcript were collected in R using the BSgenome, GenomicRanges, and GenomicFeatures packages. Hexamer entropy for each transcript was calculated as follows: occurrences of all possible hexamers in a given transcript were counted. These counts were converted into frequency space, and these frequency values were used to calculate the Shannon entropy. Shannon entropy is commonly used to represent complexity in nucleotide sequences or multiple alignments [154–156]. Similarity of transcripts to rRNA sequences were calculated as follows: each transcript was aligned to 45S (NR_046235.1) and 5S (X71804.1) rRNA using NCBI-BLAST [131] and the e-score for the best alignment was saved.

Sequence characteristic analysis

The list of IVT transcripts was sorted by transcript-level coefficients of variation for each library method (plasmid, no selection, polyA, replicate 1 of rRNA-depleted IVT-only). All transcripts with transcript-level FPKM <= 5 were excluded from further analysis. From this sorted list the transcripts with the 100 least and 100 most extreme coefficients of variation were collected for each of the above sequencing samples. The values for hexamer entropy, GC-content, and rRNA sequence similarity were compared between every pair of 100 least and 100 most extreme coefficients of variation using a Wilcoxon signed-rank test (implemented in R as the wilcox.test function). This entire analysis was repeated using transcript-level FPKM values instead of the coefficients of variation. All boxplots were prepared using R.

Description of window analysis of rRNA sequence similarity

All transcripts with zero read counts in the no selection and rRNA-depleted data were discarded. Nucleotide-level read counts for each of the remaining transcript were normalized between 0 and 1 as follows: The read depth for each nucleotide was divided by the transcript's maximum read depth. This step was performed to account for any lane effects. Read counts from the rRNA-depleted data were scaled to be strictly less than or equal to their corresponding read counts in the no selection data as follows: ignoring the first and last 100 bp of the transcript, the nucleotide position with the smallest ratio between the no selection and rRNA-depleted data was determined. All nucleotides within that transcript were multiplied by this ratio. This step was performed under the assumption that biases introduced by rRNA-depletion will always result in a loss of coverage relative to the no selection data. A 128 bp window slid across each transcript in 16 bp increments and calculated the following: 1) the highest Smith-Waterman alignment score between the sequence in the current window and a library of rRNA sequences (see below), using the SimMetrics Java package (http://sourceforge.net/projects/simmetrics/). 2) The average difference in coverage between the no selection and rRNA-depleted data. The Pearson correlation between this coverage difference and the Smith-Waterman score in each window was calculated using the cor.test function in R. The rRNA sequences used in this analysis came from the Refseq entries with the following IDs: NR_003286.2, NR_003287.2, NR_023365.1, NR_023366.1, NR_023367.1, NR_023368.1, NR_023369.1, NR_023370.1, NR_023371.1, NR_023372.1, NR_023373.1, NR_023374.1, NR_023375.1, NR_023376.1, NR_023377.1, NR_023378.1, NR_023379.1, NR_046235.1, NR_048572.1, and NR_049740.1.

Data access

We deposited all sequencing data in the NCBI Gene Expression Omnibus (GEO) under accession number GSE50445. We have also loaded the coverage tracks on the UCSC genome browser, and made them available at the following URLs: http://goo.gl/S7r5BG (comparison between different selection methods) and http://goo.gl/ISJUAH (comparison between replicates).

4.6 Tables

Table 4.1: Detection of source cDNA sequences in IVT-seq.

Total number of cDNA clones:	963	
	Replicate 1	Replicate 2
# of clones detected (FPKM ≥ 5):		
IVT Only	869	870
1:1 Mix	877	876
1:2 Mix	886	883
1:10 Mix	896	892
Mouse Only	278	271
PolyA Selection	829	-
No Selection	870	-
Plasmid Library	924	-
Average, normalized* depth of coverage for detected clones:		
IVT Only	76.09	80.22
1:1 Mix	75.15	75.06
1:2 Mix	65.79	69.40
1:10 Mix	37.50	47.46
Mouse Only	01.58	02.42
PolyA Selection	72.27	-
No Selection	72.74	-
Plasmid Library	42.08	-

*Average depth of coverage is normalized by the number of millions of fragments mapped to the human reference in each sample.

Table 4.2: Fold-change differences in within-transcript coverage by library type.

	# of IVT-transcripts with fold-change differences:		
	> 2	> 10	> 100
rRNA-Depleted	713 (74.0%)	110 (11.4%)	17 (1.7%)
PolyA Selection	678 (70.4%)	163 (16.9%)	7 (0.7%)
No Selection	400 (41.5%)	31 (3.2%)	3 (0.3%)
Plasmid	189 (19.6%)	14 (1.5%)	3 (0.3%)
Simulated	0	0	0
QM-Simulated	0	0	0

The plasmid data provides a measure of bias from library preparation/sequencing, while the ‘no selection’ data accounts for potential artifacts from the *in vitro* transcription step. To calculate the percentage of transcripts affected by bias due specifically to library preparation and sequencing, but not sequence or *in vitro* transcription artifacts, we perform the following calculation: rRNA-depletion % - no selection % + plasmid %. So we find 74% – 41.5% + 19.6% = 52.1% of transcripts in the rRNA-depleted data have > 2-fold difference in coverage, and 11.4% – 3.2% + 1.5% = 9.7% have > 10-fod difference in coverage.

4.7 Figures

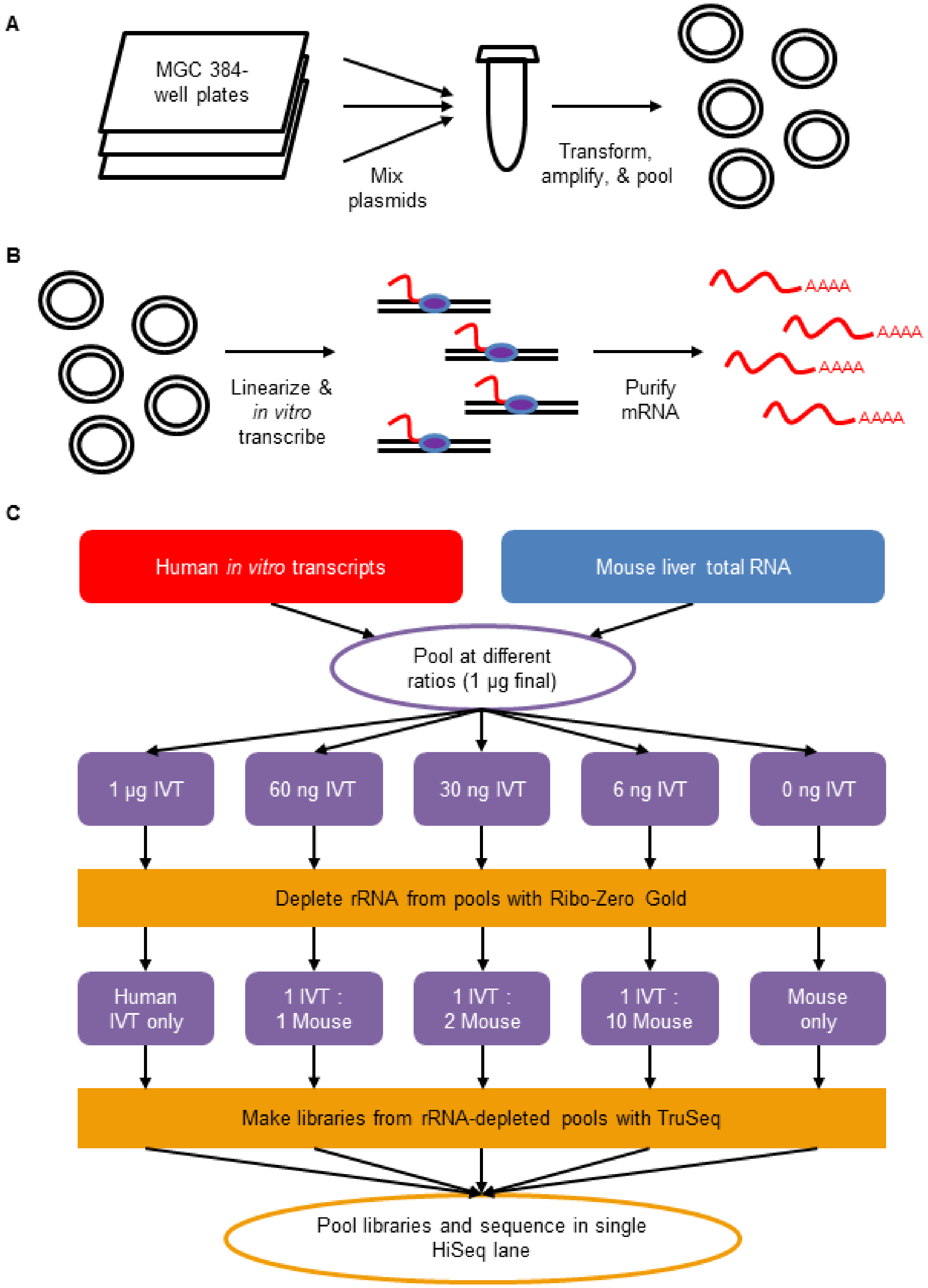
A
MGC 384-well plates
Mix plasmids
Transform, amplify, & pool
B
Linearize & *in vitro* transcribe
Purify mRNA
AAAA
AAAA
AAAA
AAAA
C
Human *in vitro* transcripts
Mouse liver total RNA
Pool at different ratios (1 µg final)
1 µg IVT
60 ng IVT
30 ng IVT
6 ng IVT
0 ng IVT
Deplete rRNA from pools with Ribo-Zero Gold
Human IVT only
1 IVT : 1 Mouse
1 IVT : 2 Mouse
1 IVT : 10 Mouse
Mouse only
Make libraries from rRNA-depleted pools with TruSeq
Pool libraries and sequence in single HiSeq lane

Fig. 4.1: Construction of IVT-seq libraries.

(A) Prepare pool of 1062 human cDNA plasmids. Contents of three 384-well plates containing MGC plasmids were pooled together. Pool was amplified via transformation in *E. coli* , and resulting clones were purified, and re-pooled. **(B)** Generate IVT transcripts. Pool of MGC plasmids were linearized and used a template for an *in vitro* transcription reaction. Enzymes and un-incorporated nucleotides were purified, leaving pool of poly(A) transcripts. **(C)** Create IVT-seq libraries. Listed quantities of IVT RNA were mixed with mouse liver total RNA to create six pools with final RNA quantities of 1 µg. Ribosomal RNA was depleted from these pools using the Ribo-Zero Gold kit. IVT RNA and mouse RNA are now present in pools at the listed ratios, following depletion of rRNA from mouse total RNA. These pools were used to generate RNA-seq libraries using Illumina’s TruSeq kit/protocol. This entire process was performed in duplicate. Replicate libraries were pooled separately and sequenced in separate HiSeq 2000 lanes (two lanes total).

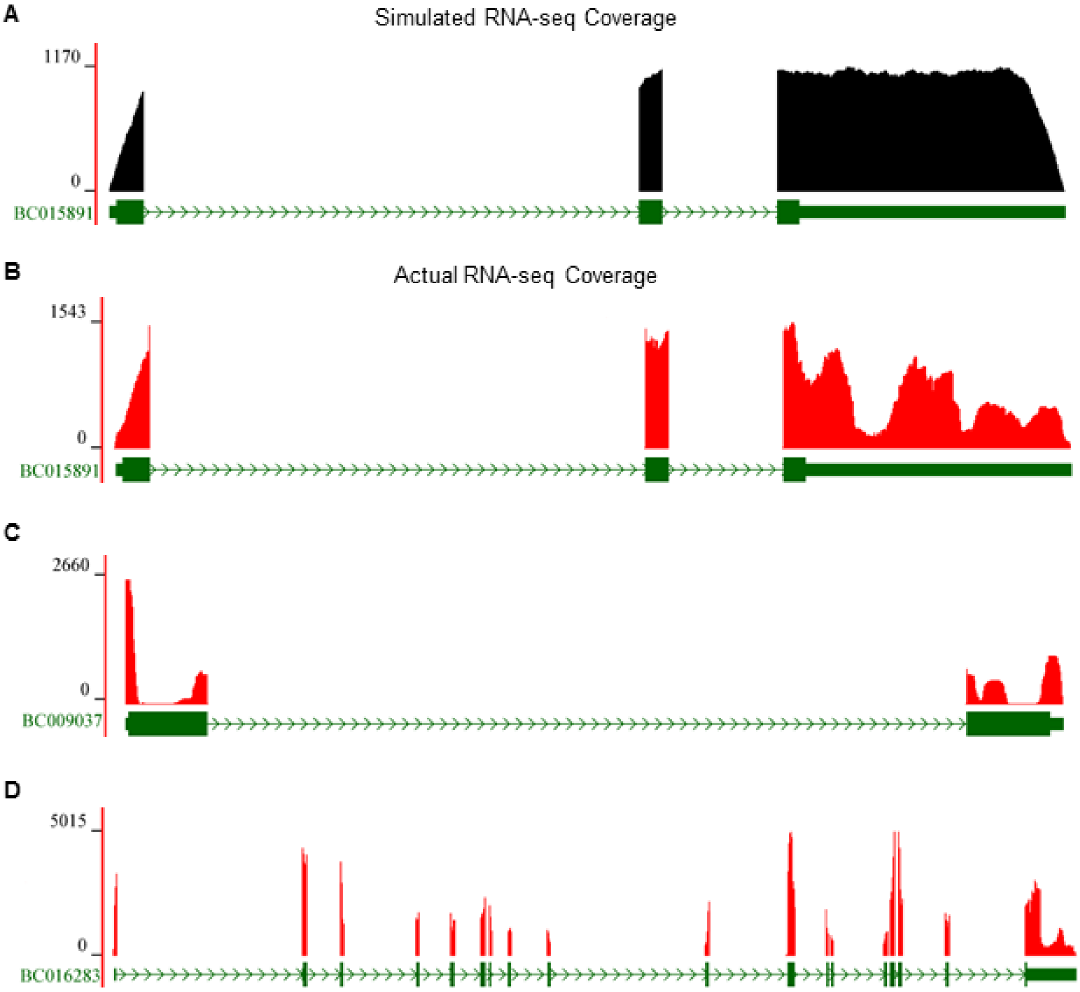
A
Simulated RNA-seq Coverage
1170
0
BC015891
B
Actual RNA-seq Coverage
1543
0
BC015891
C
2660
0
BC009037
D
5015
0
BC016283

Fig. 4.2: Within-transcript variations in RNA-seq coverage.

(A) Simulated RNA-seq coverage for a representative IVT transcript (BC015891). RNA-seq coverage plot (black) is displayed according to the gene model (green), as it is mapped to the reference genome. Blocks correspond to exons and lines indicate introns. The chevrons within the intronic lines indicated the direction of transcription. Numbers on y-axis refer to RNA-seq read-depth at a given nucleotide position. **(B)** The actual RNA-seq coverage plot for BC015891 in the IVT-only sample. Representative coverage plots for the IVT transcripts **(C)** BC009037 and **(D)** BC016283 are displayed according to the same conventions used above. All transcripts are displayed in the 5′ to 3′ direction.

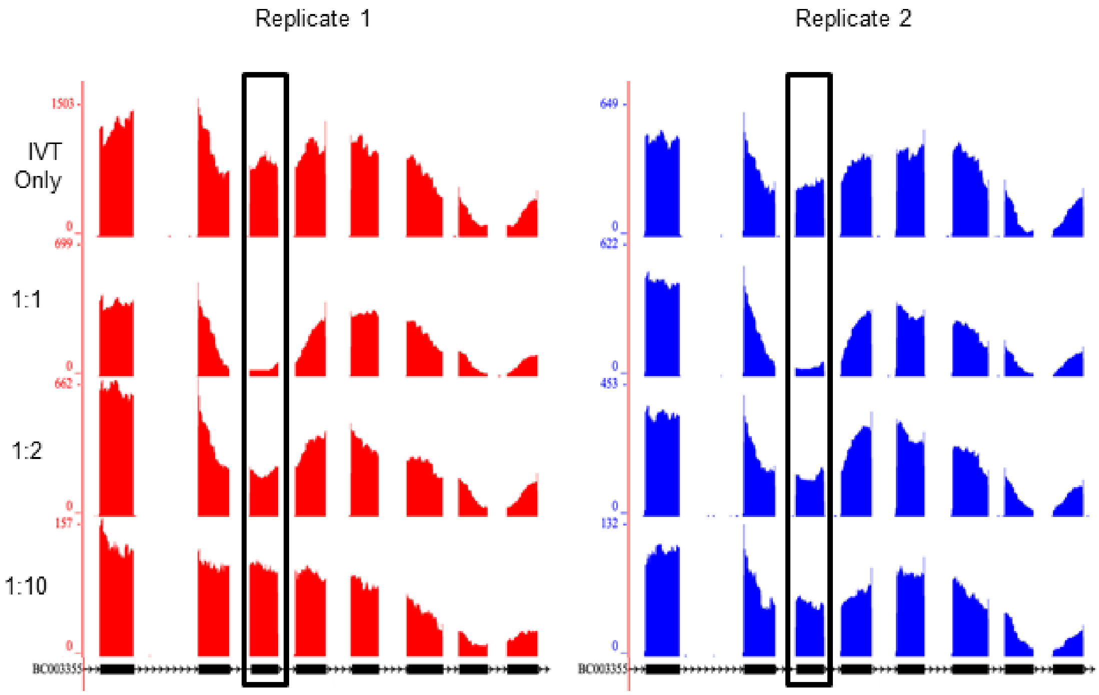

Fig. 4.3: Between-sample variations in RNA-seq coverage.

RNA-seq coverage plots across all samples for exons 4 – 11 of the IVT transcript BC003355. The black rectangles identify exon six, which shows extreme variability in coverage relative to the rest of the transcript when viewed across all of the samples. The ratio of IVT RNA to moue RNA is listed to the left of each sample's coverage plots. Coverage plots (red for first replicate; blue for second replicate) are displayed according to the gene model (black), as it is mapped to the reference genome. Blocks in the gene model correspond to exons and lines indicate introns. The chevrons within the intronic lines indicated the direction of transcription. Numbers on y-axes refer to RNA-seq read-depth at a given nucleotide position.

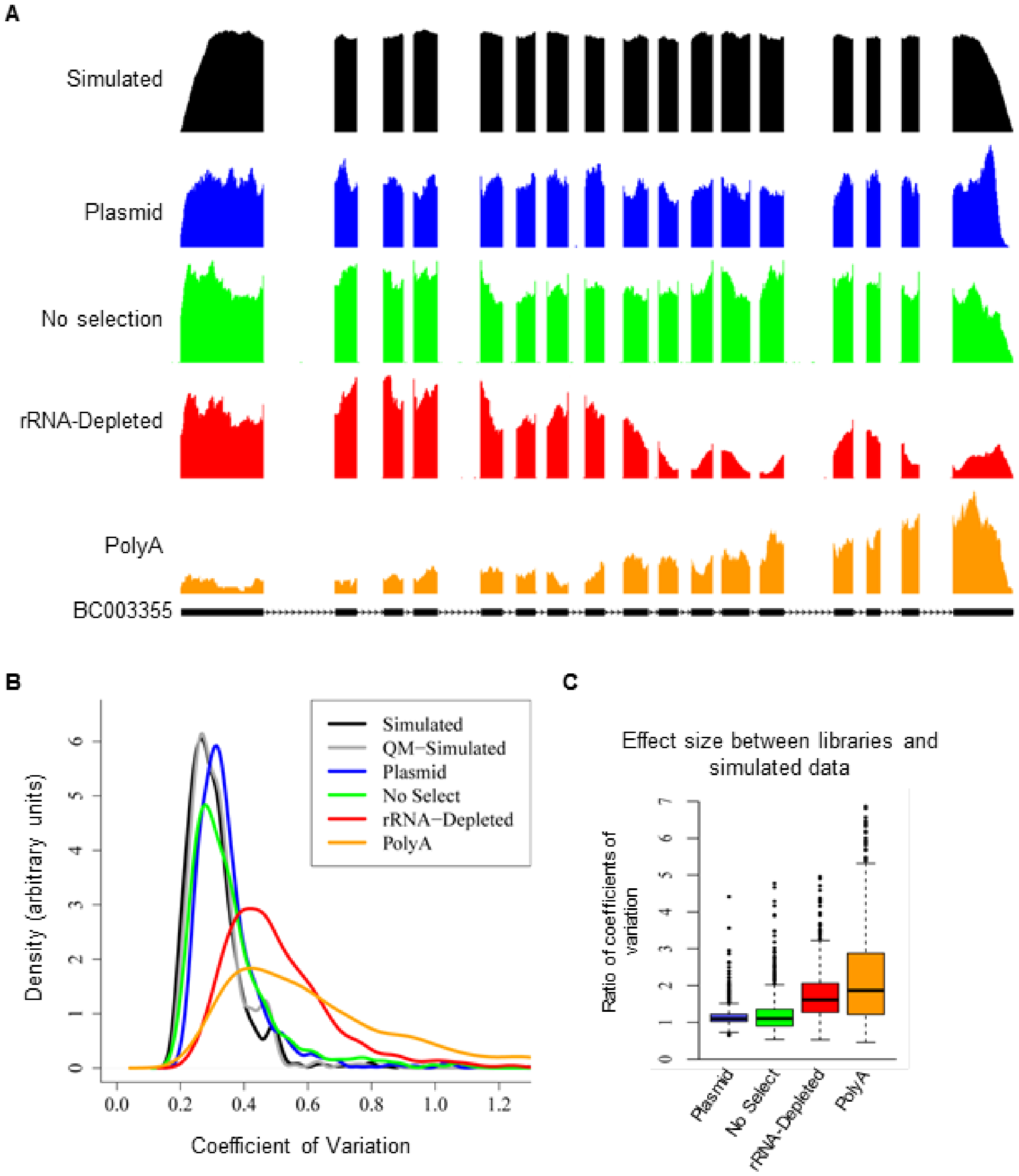
A
Simulated
Plasmid
No selection
rRNA-Depleted
PolyA
BC003355
B
Simulated
QM−Simulated
Plasmid
No Select
rRNA−Depleted
PolyA
Density (arbitrary units)
0
1
2
3
4
5
6
0.0
0.2
0.4
0.6
0.8
1.0
1.2
Coefficient of Variation
C
Effect size between libraries and simulated data
Ratio of coefficients of variation
0
1
2
3
4
5
6
7
Plasmid
No Select
rRNA-Depleted
PolyA

Fig. 4.4: Sources of bias in RNA-seq coverage.

(A) RNA-seq coverage plots for IVT transcript BC003355 from simulated (black), plasmid (blue), no selection (green), rRNA-depleted (red), and polyA (orange) data. The gene model is displayed in black, below all of the coverage plots. Blocks correspond to exons and lines indicate introns. The chevrons within the intronic lines indicated the direction of transcription. **(B)** Distributions for coefficients of variation across data displayed above, with the addition of the QM-simulated data (gray). Note that while the graph is cutoff at a coefficient of variation of 1.3, the tails for the Ribo-Zero and PolyA distributions extend out to 2.13 and 2.7, respectively. **(C)** Effect sizes for the differences in distribution of coefficients of variation between sequencing libraries and simulated data. Effect sizes are calculated as the per-transcript ratios of coefficients of variation between a given library and the simulated dataset.

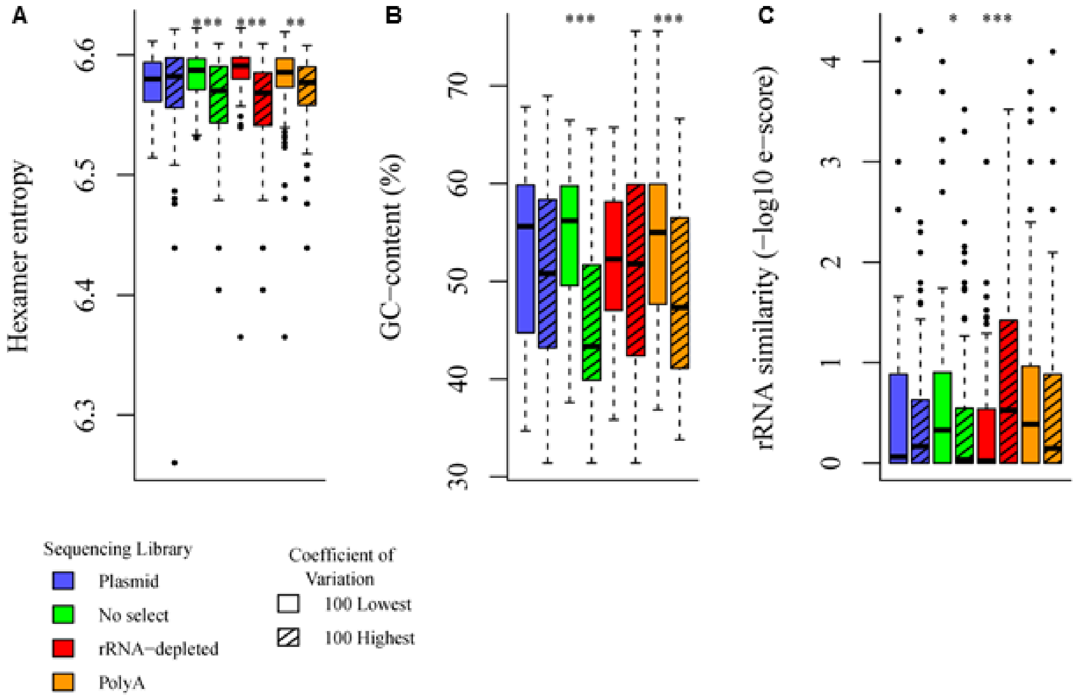

Fig. 4.5: Effects of sequence characteristics on coverage variability.

Distributions of **(A)** hexamer entropy, **(B)** GC-content, and **(C)** rRNA sequence similarity for the 100 transcripts with the highest and lowest coefficients of variation for transcript coverage from the plasmid, no selection, rRNA-depleted, and polyA libraries. Asterisks indicate the significance of a Wilcoxon signed-rank test comparing values for the listed sequence characteristics between each pair of groups from the same sample. * = *p*-value < 0.05; ** = *p*-value < 0.01; *** = *p*-value < 0.001.

4.8 Supplemental tables

Table S4.1: Alignment statistics for all sequencing data sets.

Library	Total Fragments	Total Fragments Mapped	Unique Mappers
Simulated data:			
Simulated	10,000,000	9,997,076 (99.9%)	9,631,147 (96.31%)
QM-Simulated	40,000,000	39,990,729 (99.9%)	39,308,168 (98.27%)
Replicate 1:			
IVT only	41,260,668	38,237,172 (92.6%)	36,740,018 (89.04%)
1:1 mix	35,315,448	23,824,059 (67.4%)	22,246,072 (62.99%)
1:2 mix	42,092,262	24,139,758 (57.3%)	21,365,369 (50.75%)
1:10 mix	37,640,274	12,989,430 (34.5%)	9,890,945 (26.27%)
Mouse only	39,243,399	10,697,075 (27.2%)	6,473,917 (16.49%)
Replicate 2:			
IVT only	32,240,162	29,278,157 (90.8%)	28,352,845 (87.94%)
1:1 mix	32,474,073	21,714,926 (66.8%)	19,836,639 (61.08%)
1:2 mix	36,655,155	19,582,020 (53.4%)	16,870,009 (46.02%)
1:10 mix	34,091,563	10,392,308 (30.4%)	6,459,324 (18.94%)
Mouse only	39,086,565	8,746,680 (22.3%)	3,391,749 (8.67%)
Other libraries			
Plasmid*	24,008,610	5,969,758 (24.8%)	5,650,393 (23.53%)
No Selection	105,970,103	97,410,895 (91.9%)	93,636,659 (88.36%)
PolyA	7,779,720	7,214,472 (92.7%)	6,919,962 (88.94%)

*Relatively low percentage of mapped reads is due to the presence of the plasmid backbone in library. The backbone does not map to the reference human genome.

4.9 Supplemental figures

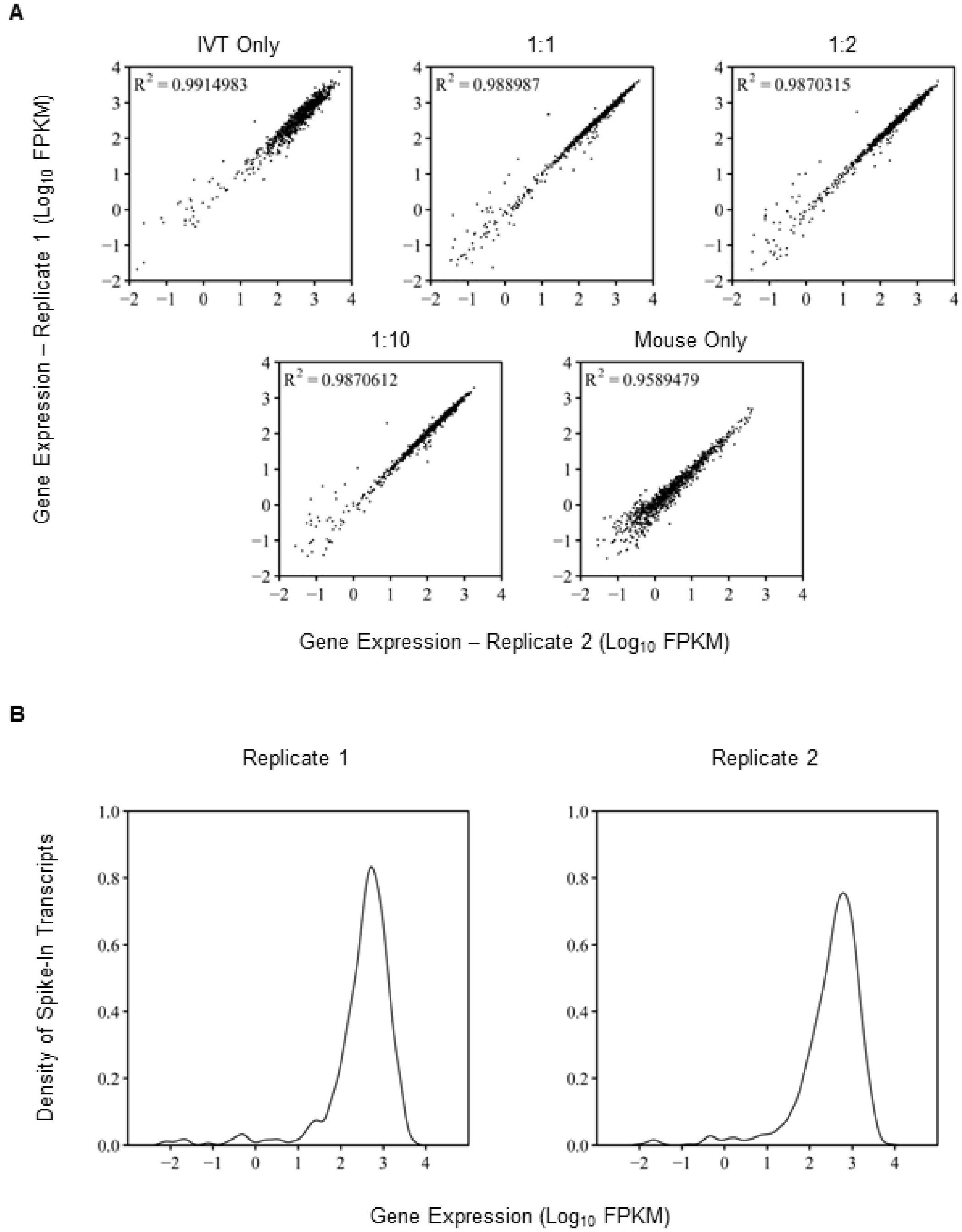

Fig. S4.1: Expression comparison between replicates.

(A) Correlation plots for log10 transcript-level FPKM values between replicate IVT-seq samples. Pearson R^2 values for the correlations are included as inserts in each plot. **(B)** Distribution of FPKM values in both replicates of the IVT-only sample. FPKM values are plotted on the x-axis in log10 space. The y-axis is plotted in arbitrary density units.

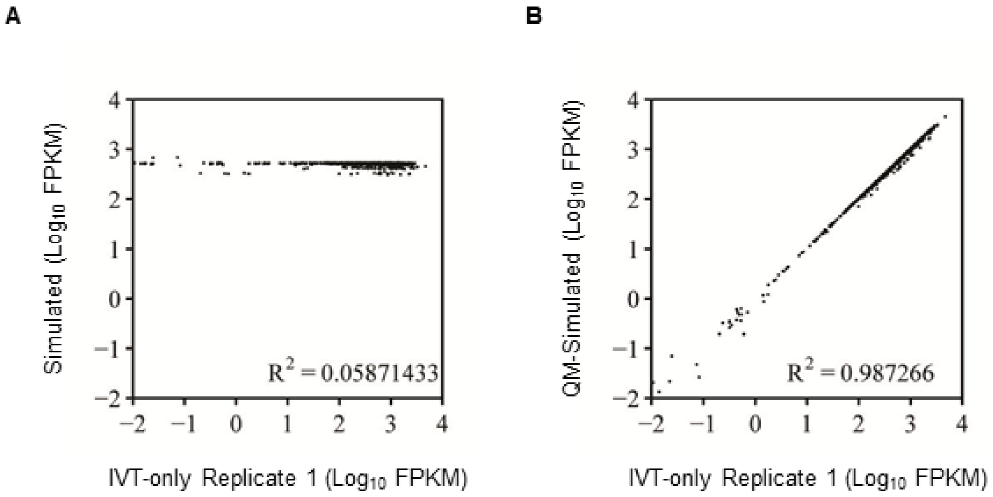

Fig. S4.2: Expression comparison between simulated and IVT data.

Correlation plots for log10 transcript-level FPKM values between **(A)** simulated data or **(B)** QM-simulated data, and replicate 1 of the IVT-only data. Pearson R^2 values for the correlations are included as inserts in each plot.

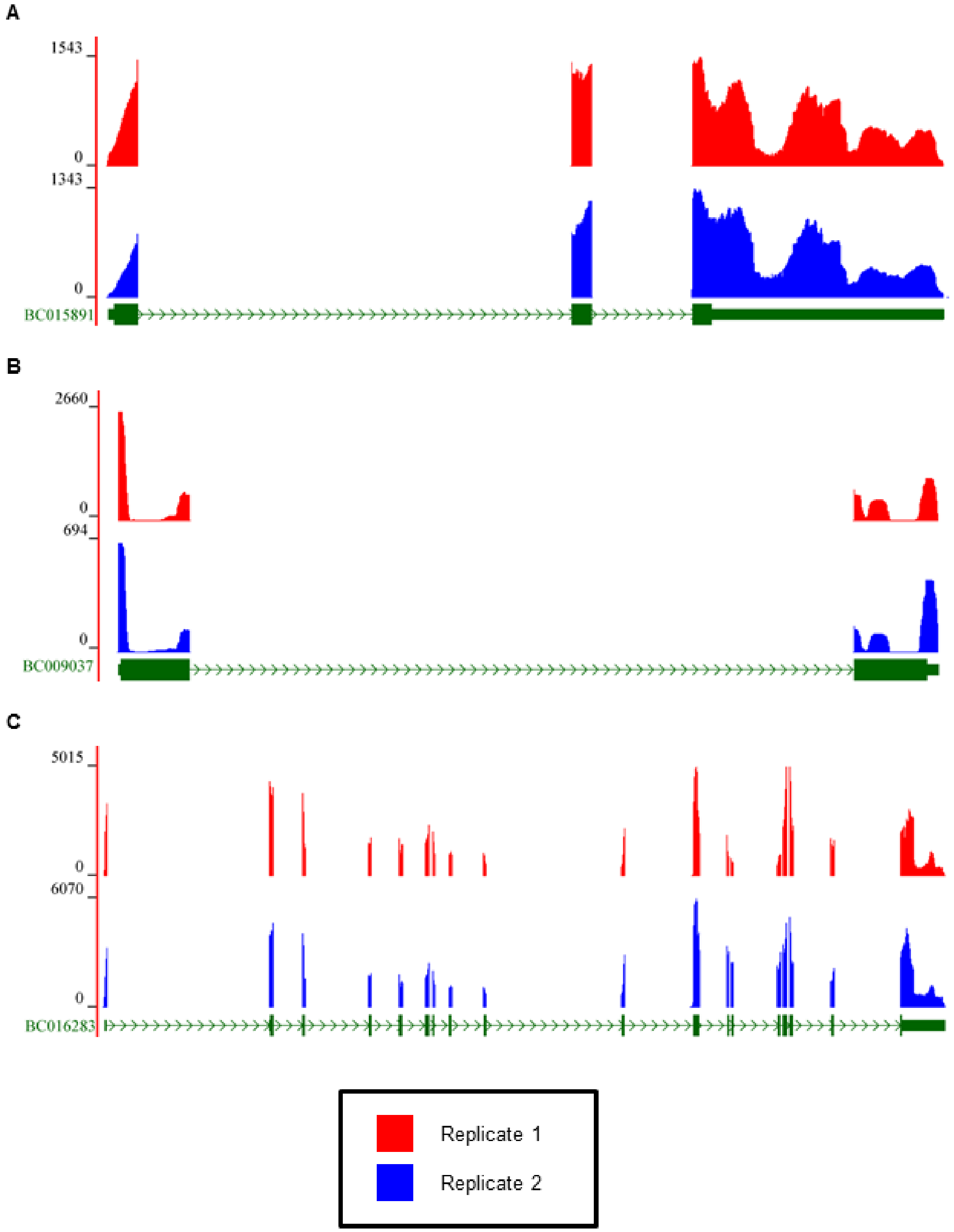
A
1543
0
1343
0
BC015891
B
2660
0
694
0
BC009037
C
5015
0
6070
0
BC016283
Replicate 1
Replicate 2

Fig. S4.3: Coverage patterns are reproducible across replicates.

Coverage patterns from both replicates for all transcripts in Fig 4.2. RNA-seq coverage plots from replicate IVT only samples (red – replicate 1l blue – replicate 2) for **(A)** BC015891, **(B)** BC009037, and **(C)** BC016283 are displayed according to the gene model (green), as it is mapped to the human reference genome. Blocks correspond to exons and lines indicate introns. The chevrons within the intronic lines indicated the direction of transcription. Numbers on y-axis refer to RNA-seq read-depth at a given nucleotide position. All transcripts are displayed in the 5′ to 3′ direction.

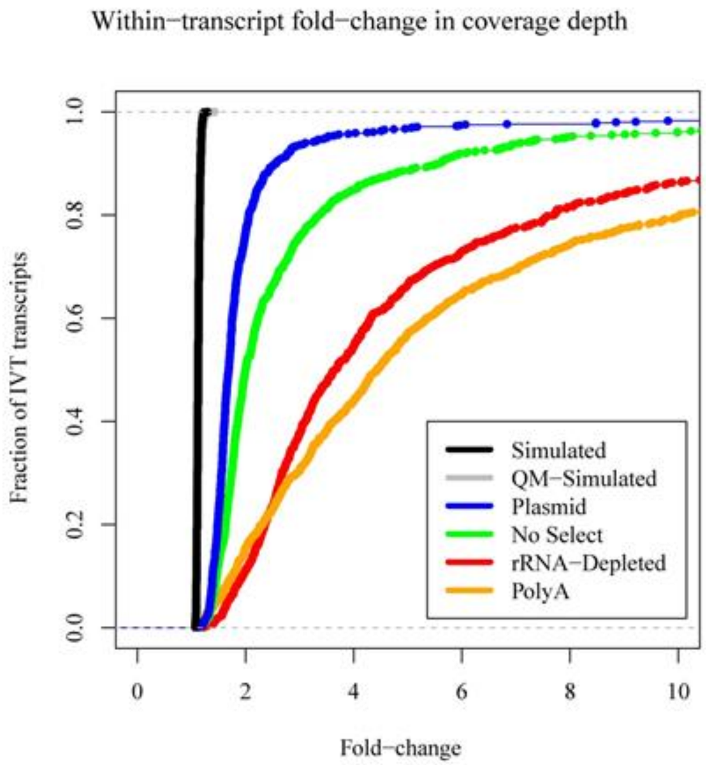

Fig. S4.4: Fold-change in within-transcript coverage across libraries.

The cumulative distribution functions for fold-change in within transcript coverage are displayed for the rRNA-depleted (red), polyA (orange), no selection (green), plasmid (blue), QM-simulated (gray), and simulated (black) datasets. Curves toward the left side of the plot indicate fewer genes contain high fold-change differences in coverage. Curves toward the right side of the plot indicate many genes contain high fold-change differences in coverage. The dotted lines indicate the y-axis values for none of the data (0.0) and all of the data (1.0). This plot is focused on the fold-change values between 1 and 10. See the *Materials and Methods* section for full details on the fold-change calculations.

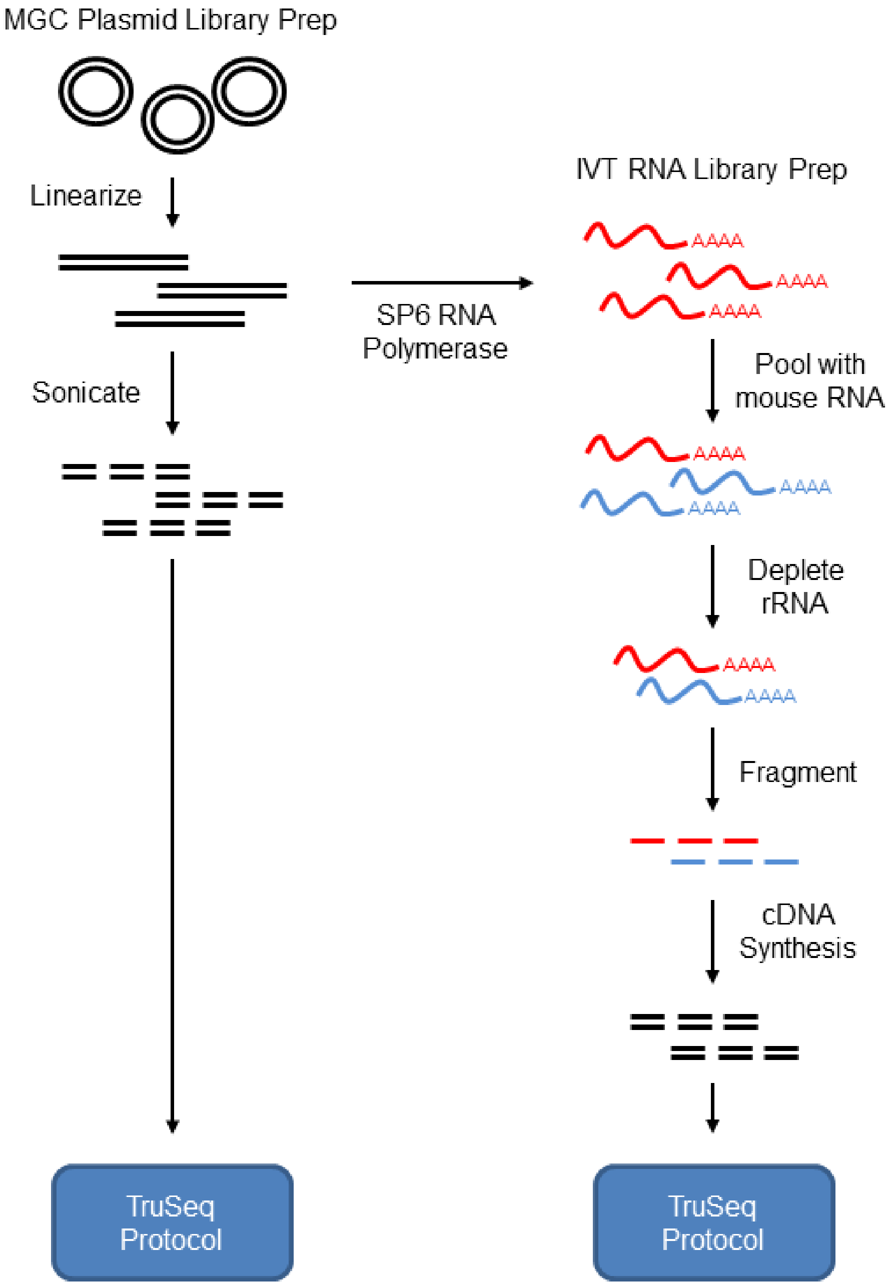
MGC Plasmid Library Prep
Linearize
Sonicate
SP6 RNA Polymerase
IVT RNA Library Prep
AAAA
AAAA
AAAA
Pool with mouse RNA
AAAA
AAAA
AAAA
Deplete rRNA
AAAA
AAAA
Fragment
cDNA Synthesis
TruSeq Protocol
TruSeq Protocol

Fig. S4.5: Plasmid sequencing protocol compared to IVT-seq.

The protocol for preparing MGC plasmids for DNA-sequencing library generation is displayed alongside the protocol for preparing IVT transcripts for RNA-seq library generation. Both protocols start by linearizing the plasmids. For DNA-sequencing, linearized plasmids are fragmented via Covaris sonication, and the resulting fragments are taken through the TruSeq protocol. For RNA-sequencing, the linearized plasmids are used as templates for an *in vitro* transcription reaction. IVT RNA is then pooled with mouse RNA, rRNA is removed from pool via Ribo-Zero Gold kit, rRNA-depleted pool is fragmented via metal-ion hydrolysis, and fragmented RNA is converted to cDNA via reverse transcription with random-hexamer priming. The resulting cDNA fragments are then taken through the TruSeq protocol.

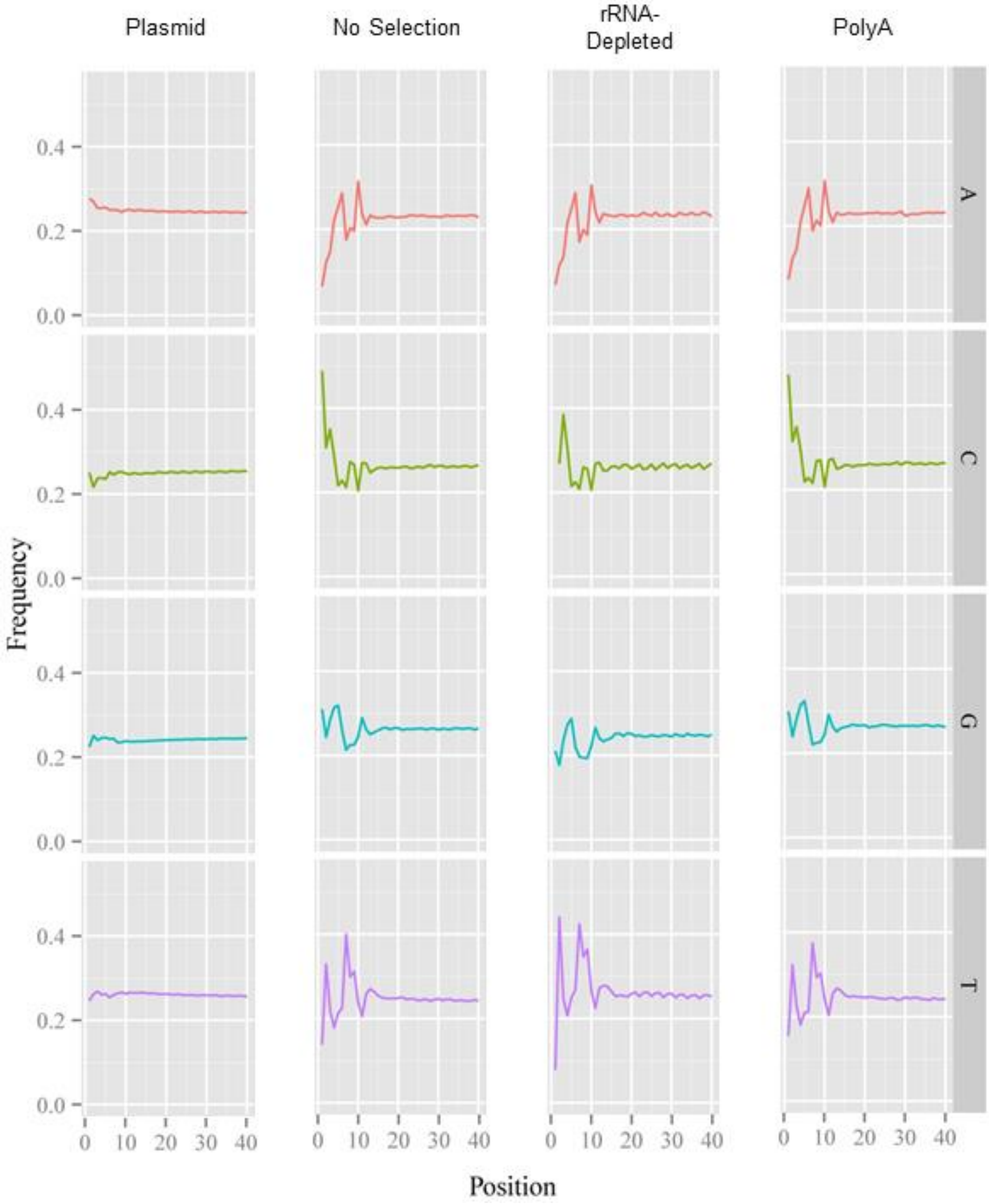

Fig. S4.6: Random hexamer bias across all selection methods.

Nucleotide frequency as a function of read position for sequencing reads at the 5' ends of cDNA fragments. Frequencies are plotted for plasmid, no selection, rRNA-depleted, and polyA datasets.

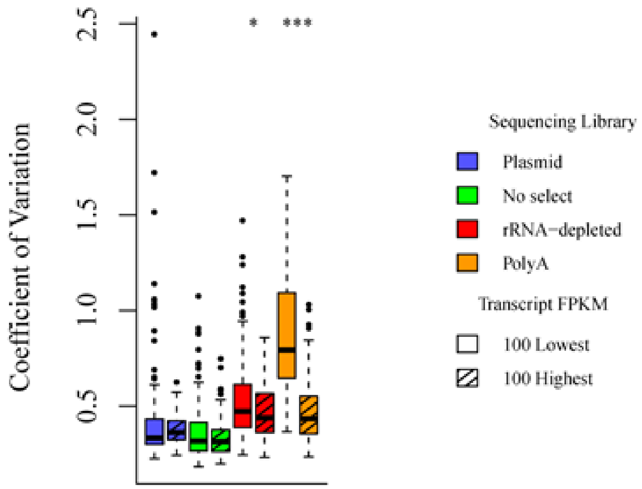

Fig. S4.7: Confounding effects between coverage depth and variability.

Distributions of transcript-level coefficients of variation for the 100 transcripts with the highest and lowest transcript-level FPKMs from the plasmid, no selection, rRNA-depleted, and polyA libraries. Asterisks indicate the significance of a Wilcoxon signed-rank test comparing values for the listed sequence characteristics between each pair of groups from the same sample. * = *p*-value < 0.05; *** = *p*-value < 0.001.

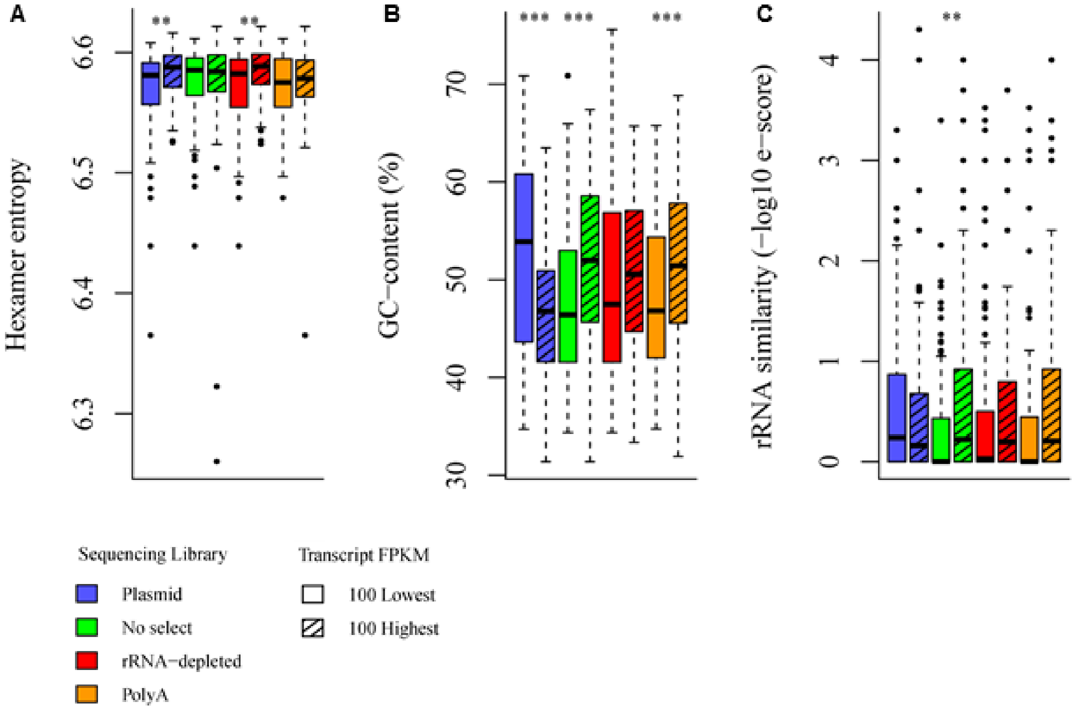

Fig. S4.8: Effects of sequence characteristics on coverage depth.

Distributions of **(A)** hexamer entropy, **(B)** GC-content, and **(C)** rRNA sequence similarity for the 100 transcripts with the highest and lowest transcript-level FPKMs from the plasmid, no selection, rRNA-depleted, and polyA libraries. Asterisks indicate the significance of a Wilcoxon signed-rank test comparing values for the listed sequence characteristics between each pair of groups from the same sample. ** = *p*-value < 0.01; *** = *p*-value < 0.001.

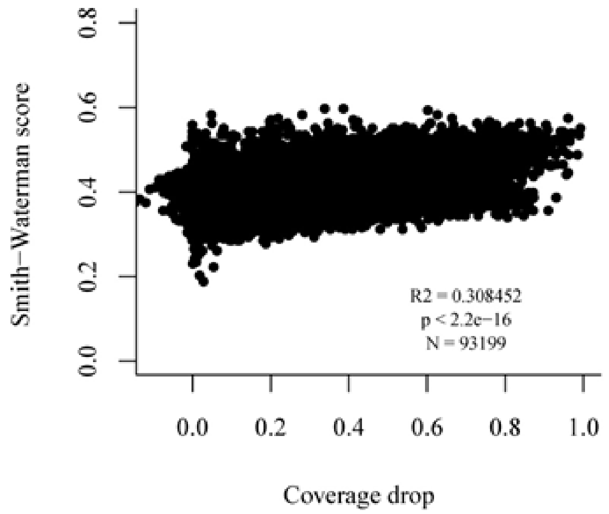

Fig. S4.9: rRNA sequence similarity and coverage bias in rRNA-depleted data. Correlation plot between Smith-Waterman alignment score to rRNA sequences and the magnitude of the decrease in coverage depth between no selection and rRNA-depleted samples. A coverage drop of 1.0 indicates a large decrease in coverage between the no selection and rRNA-depleted samples. A coverage drop of 0 indicates no difference between the two samples. For full details on this analysis, see Materials and methods section.

Chapter 5: Future directions

5.1 Circadian rhythms and RNA-seq

Confirming miRNA targets

We found examples of oscillating miRNA host transcripts in chapters 2 & 3. In chapter 2, we found a cluster of oscillating miRNA precursors for *Mir290*, *Mir291a*, *Mir292*, *Mir291b*, *Mir293*, *Mir294*, and *Mir295*. In chapter 3, we found the host gene for *Mir22* oscillates in heart and lung tissue. In both cases, these miRNAs have promising predicted targets (prediction by TargetScan, release 6.2 [134]). *Mir292* and *Mir291b* have predicted target sites in the 3' UTR of *Clock*. *Mir22* has one of the most broadly-conserved target sites located in the 3' UTR of *Ptgs1*. While both of these findings offer interesting functional implications, these targets must be experimentally confirmed and validated.

Over-expression studies provide the most direct approach for initial testing. Briefly, this involves transfecting miRNA mimics into cells expressing the predicted targets. If the predicted targets are valid, a Western blot should show decreased levels of the target protein in the presence of the miRNA mimic, and no change in the presence of a scrambled miRNA sequence (a control which accounts for any of the transfection effects). Alternatively, one could use an antagomir to knock down endogenous expression of the miRNA [159], and then check for effects on target gene expression/translation. While this second approach does avoid potential artifacts arising from super-biological levels of the miRNA of interest, it does require finding a cell line in which both the miRNA and its target are expressed at high enough levels to elicit a response. However, any differences in protein levels one might observe from either of these experimental paradigms could be due to interactions with intermediate genes, rather than a direct effect.

To confirm the direct interaction between miRNAs and target genes, begin by cloning the 3'UTR of the target transcript into a luciferase reporter vector, such that the target UTR sequence serves as the UTR for the luciferase transcript. In this way, any regulatory effect occurring through the UTR sequence should affect downstream luciferase levels as measure by luminescence. As a control, take the same luciferase-UTR construct and mutagenize the predicted miRNA seed sequence. As the seed sequence is necessary for correct miRNA binding and function, disrupting it should remove any regulatory effects caused by direct miRNA binding to that specific site in the target UTR (for full details on protocol, see [160]).

Lastly, to address the circadian aspect of this miRNA regulation, particularly in the case of the mir292 cluster which may target a core clock component, we should repeat these experiments listed above in a cell line with endogenous circadian rhythms (like U2OS or NIH3T3 cells [40, 161, 162]). For *Mir292* and *Mir291b*, we would expect miRNA mimics to affect oscillations in these cell lines similar to a *Clock* knockdown experiment [162]. To assess the effects of *Mir22* on oscillations in *Ptgs1* requires that we find a cell line in which *Ptgs1* oscillates. If no such cell line exists, we may need to look at endogenous rhythms in *Ptgs1* mRNA and protein from mouse lines deficient in *Mir22*. Given our data, the best tissues to examine in these mice would be heart and lung.

Expanding our analysis of circadian transcription across twelve mouse tissues

In chapter 3, we presented the results of an experiment using RNA-seq and microarrays to assay circadian transcription across twelve mouse tissues. We only scratched the surface of this huge dataset by focusing on non-coding transcripts. This data is also useful for a large-scale study of circadian mRNA splicing and (TSS) selection. While previous work has found several examples of alternative splicing with a time-of-day effect [41, 48, 163], these appear to form a relatively small portion of the oscillating transcriptome within a specific tissue. By using our data to expand this analysis across multiple tissues, we can look for correlations between differential spliceform usage and differential oscillations between tissues. For example, do rhythmic genes have different "oscillatory" and "non-oscillatory" spliceforms expressed in different tissues? These initial analyses should focus on alternative TSS usage as opposed to internal alternative exon usage, given evidence that clock regulation of alternative splicing seems to operate largely at the level of TSS selection [48]. This analysis may prove challenging for genes with low levels of expression, or particularly long transcripts, as we used polyA selection during the preparation of our RNA-seq libraries. Future studies seeking to analyze circadian TSS usage specifically should use alternative selection methods.

Given the limited knowledge of the functional role of ncRNAs, this dataset may also serve to guide future projects for the functional annotation of ncRNAs. For example, one could use the list of conserved ncRNAs as the basis for a cell-based functional screen. Researchers have already had success using such screens to identify the functional effects of miRNAs on the NF-κB signaling system [164]. We would begin by assembling a cDNA library of, or an siRNA library targeting the oscillating ncRNAs identified above. Next, we would perform a high-throughput transfection of these libraries into cells with a luciferase reporter for a specific pathway (like NF-κB signaling, CLOCK-BMAL1 activation of E-boxes, or oscillations in *Per2* promoter activity). We would then use any ncRNAs which affected these various pathway outputs for secondary screening and mechanism characterization. Additionally, once such a library of ncRNAs exists, we could use it to assay any pathway with an output measurable by luciferase reporter.

Lastly, while our current work is limited more to lincRNAs and host genes, future studies could repeat these experiments to look at mature miRNA sequences. The RNA purification and library preparation techniques we used will only yield expression data for transcripts greater than 100 bp in length. This misses all small RNAs, including mature miRNAs. Future work should use small RNA-seq protocols across multiple tissues to expand upon our work. Also, future studies could sample RNA-seq data at a higher sampling resolution to detect more oscillating ncRNAs with a likely greater degree of accuracy.

5.2 IVT-seq and bias analysis

Replicate IVT-seq experiment with different platforms

In chapter 4 we presented a method for assessing technical biases in RNA-seq experiments, and used it to study the effect of RNA-selection method on coverage. However, we only performed this analysis using Illumina sequencing technology. Future studies could repeat these experiments using 454, SOLiD, and PacBio sequencing. While there are existing studies comparing the differences between these technologies, none have examined whether or not they have differential effects on RNA-selection method. We could assess the effects of particular sequencing methodologies on the variability introduced by rRNA-depletion or polyA selection (ex. does a particular sequencing platform exacerbate these biases?). Additionally, we could assess how sequencing method affects the location and frequency of hunc regions. This expanded study would allow us to more precisely determine whether or not particular biases arise from RNA selection (which would conceivably be common to all sequencing methods), library preparation, and sequencing.

Identify cause of hunc regions

At the moment, we hypothesize that the hunc regions described in chapter 4 could arise from a trans interaction between the IVT sequences and the background RNA (mouse liver RNA in this case). We can test this by creating additional sequencing libraries using different background RNA. Briefly, we would prepare new RNA samples that consist of different proportions of the human IVT RNA and background RNA (1:1, 1:2, 1:10, and background only). For background RNA, we could use mouse brain tissue (cortex, hypothalamus), as it is quite functionally distinct from the liver and is therefore likely to express a significantly different population of transcripts. We could take existing transcriptome studies across multiple tissues (like the one presented in chapter 3) to confirm that the population of RNA presenting in new background samples are highly distinct from one another. Also, we could use background RNA from a completely new species, like *Drosophila* or *C. elegans*. However, we must take care when selecting RNA from a new species; we are using an rRNA-depletion kit targeting mammalian ribosomes and we do not want to pick RNA from an organisms with rRNA sequences not recognized by this kit. Finally, after creating a dilution series of RNA pools from different background RNA, we would prepare and sequence libraries using the same protocol described in chapter 4, and compare any resultant hunc regions to those we generated using the mouse liver RNA background. If our initial hypothesis is true, we would expect to see different hunc regions represented in libraries with different background RNAs. If we see the same hunc regions present regardless of background RNA, it suggests this effect may be the result of an interaction between some aspect of library preparation, and the quantity of IVT RNA present in a sample. We could also repeat the above dilution experiments using polyA selection to address: a) whether or not hunc regions are an effect of rRNA-depletion specifically, and b) whether or not polyA selection is differentially affect by any trans interactions between different RNA. These future studies would help us further elucidate the true sources of this bias.

Chapter 6: Summary and conclusions

The circadian clock appears to regulate an ever-expanding list of behaviors, biological processes, and disease phenotypes. With this list of large, system-level clock outputs comes the need to identify CCGs; the molecular outputs that mediate these grand rhythms. As I noted in chapter 1, the majority of studies to identify CCGs have been performed using microarrays. However, with the advent and growing popularity of RNA-seq, it appears most researchers have begun to favor sequencing over arrays. RNA-seq offers many advantages over microarrays, particularly for identifying novel transcripts, and for quantifying ncRNAs and other transcripts not commonly assayed by arrays. In this dissertation, I demonstrate the utility of using RNA-seq to find CCGs. However, RNA-seq is relatively young, having only emerged in the past 6-7 years. As a result, our understanding of its biases and limitations is still very much evolving. To address this shortcoming, I conclude the work in this dissertation by presenting a technique for assessing the sources of bias from sequencing experiments.

In chapter 2 I present a hybrid circadian expression profile which uses both microarrays and RNA-seq to identify oscillating transcripts in the mouse liver. This work demonstrates the feasibility of leveraging the low cost and established quantification pipelines for microarrays, with the nucleotide-level, genome-wide data provided by RNA-seq. In fact, a similar hybrid study found a new component of the zebrafish clock that appears to link molecular oscillations to rhythms in locomotor activity [165]. To begin, I analyzed the effect sampling resolution has on the ability to detect cycling transcripts, and found a steep drop in accuracy when sampling at greater than 4-hour resolution, or using only one data of data. As a result, I recommend that future circadian RNA profiles use at least a 2-hour sampling resolution over 48 hours to achieve the best balance between detection rates and cost of sample preparation and analysis.

From the microarray and RNA-seq expression data, I examined splicing in clock genes and their direct targets. I noted that ~84% of core clock genes have multiple annotated spliceforms, with ~58% expressing multiple spliceforms concurrently in the liver. These spliceforms appeared to oscillate in phase with each other, suggesting that circadian regulation of alternative splicing among core clock genes in the liver is not a significant factor in their expression. Additionally, I found evidence of a novel, alternative TSS for the *Dbp* gene, two novel lincRNAs, and a cluster of miRNAs, all with circadian expression. This oscillating cluster of miRNAs is predominantly expressed in ES cells [98, 105], and two of its members , *Mir292* and *Mir291b*, are predicted to target *Clock*. Given that ES cells appear to have no core clock rhythms [166], the expression of this miRNA cluster may help suppress these rhythms by targeting *Clock*. These results demonstrate the utility of a hybrid approach to circadian expression profiling, and of RNA-seq for identifying non-coding transcripts. Furthermore, this means we can leverage the vast body of existing array data to supplement and improve our future RNA-seq experiments.

In chapter 3 I applied this hybrid technique to twelve different mouse tissues to characterize the circadian non-coding transcriptome. As part of this work, I constructed a list of 1,016 ncRNAs conserved between humans and mice. Of this list, I found roughly 1/3rd oscillated in at least one of the sampled tissues, with no single ncRNA oscillating in more than five tissues. These oscillating ncRNAs included snoRNA and miRNA host genes. The snoRNAs associated with these host genes are predicted to modify 18s and 28s rRNA [167–169], and may contribute to the clock's emerging role in ribosome biogenesis [170]. One of the miRNAs with an oscillating host gene, *Mir22*, has been implicated in oncogenesis, arthritis, and cardiac stress [137, 171–173]. In the heart and liver, *Mir22hg* oscillates antiphase to the transcript levels of its predicted target, *Ptgs1*. This is significant as *Ptgs1* is the aspirin target thought to mediate the cardioprotective effect of low-dose aspirin therapy [138]. It is possible these *Mir22hg* oscillations may contribute to the rhythmic expression of *Ptgs1*. Additionally, I characterized 1,979 novel antisense transcripts, 43 of which oscillate at least eight hours out of phase with their sense transcripts. This includes novel a *Bmal1* antisense transcript. I also found oscillations of the previously characterized *Per2* antisense transcript [44, 51] in five tissues, the most out of any antisense transcript. Finally, I identified 5,154 putative ncRNAs. Of these, 712 oscillated in at least one tissue. These oscillating transcripts, both known and novel, provide excellent candidates for future functional analysis. Also, future analyses could likely improve the number and quality of the conserved ncRNAs I identified, by regenerating this list using successive versions of ncRNA databases, and by adding a step to filter out sequences with significant ORFs and coding potential.

In chapter 4, I presented a method, IVT-seq, for assessing the sources of technical bias in library preparation and sequencing. As this method is based around creating a pool of know IVT transcripts, it is applicable to any current or future technology that uses RNA as input. I applied this method to the Illumina sequencing methodology, in order to assess the contribution of different steps in library preparation to coverage bias. All stages of library preparation contribute bias, but it appears RNA selection (either polyA or rRNA-depletion) is responsible for > 2-fold differences in coverage across ~50% of IVT transcripts and > 10-fold differences across ~9% of transcripts. This finding suggests it is inadvisable to use exon-level quantification, or any other method which uses a subset of the whole transcript. I also identified regions in ~6% of transcripts which show variability in coverage level independent of the rest of the transcript. These hunc (high, unpredictable coverage) regions may be the result of trans interactions between different RNA sequences in the same sample. This has important implications for the interpretation of RNA-seq results, as gene-level quantification measures cease to be independent of one another. Finally, I attempted to determine sequence features responsible for these biases. While there was no single factor responsible for the majority of these biases, hexamer entropy and GC-content contribute to variability among RNA-specific aspects of library preparation, and rRNA sequence similarity contributes to variability in the rRNA-depleted library. This work demonstrates the utility of IVT-seq for bias analysis, and serves as a cautionary note for the interpretation of RNA-seq results.

In chapter 5, I conclude this dissertation by discussing future experiments that expand on the work presented in the previous chapters. Chief among these is the validation of *Clock* and *Ptgs1* as targets of the *Mir292* cluster and *Mir22*, respectively. If these are confirmed as direct targets, they have significant implications for regulation of the circadian clock and its effect on cardiac health. Also, to build upon the analysis of bias in RNA-seq data, I propose repeating the IVT-seq experiments using different sequencing technologies. This should serve to improve our understanding of bias sources in each of these different methods, and may also serve to highlight biases common to any sequencing-based technology. Lastly, I propose an experiment to examine whether or not hunc regions are the result of trans interactions between different RNA transcripts in the same sample.

Taken together, the work presented in this thesis demonstrates the power of RNA-seq data, as well as the pitfalls to consider when interpreting its results. The data presented herein should serve as an excellent resource to any researchers interested in circadian gene expression, cross-tissue transcriptome comparisons, benchmarking, and improving algorithms for the analysis of RNA-seq data. This work also provides additional evidence suggesting we put greater focus on ncRNAs in future studies of the circadian system, and ultimately any molecular network. Finally, this work adds to the call for continued development of RNA-seq analysis tools, standardized protocols, analysis pipelines, and data repositories. We continue to mine new insights from RNA-seq data, and with careful planning and analysis we can avoid getting buried by our own data.

Appendices

Appendix A – Names and descriptions of supplementary digital files

There are several data tables which are too large to include as part of this thesis. These have been uploaded as supplementary digital files. This section lists the names for these files, as well as a brief description of their contents.

Supplementary_digital_file_S1.xls - Conserved non-coding RNAs
This file lists annotation data for each conserved ncRNA identified in chapter 3 of this thesis. Annotation data includes genomic coordinates from the mouse geneome, the assigned RefSeq IDs and gene symbols from mouse and human, the direction of the alignment between the ncRNA and RefSeq sequences (sense or antisense), and the functional group assigned to each ncRNA. Additionally, this file lists the peak phase in expression for each tissue, as well as the number of tissues in which each ncRNA oscillates.

Supplementary_digital_file_S2.xls - Novel antisense transcripts
This file lists annotation data for each novel antisense transcript identified in chpater 3 of this thesis. Annotation data includes the genomic coordinates for each transcript, as well as the Ensembl gene ID and symbol for the overlapping sense transcript. Additionally, this file lists the difference between peak expression phase in the sense and antisense transcripts for each tissue. If these columns list "antisense_osc_only", or "sense_osc_only", it means that in the given tissue, only the antisense or sense transcript oscillated, respectively. Lastly, this file contains four summary columns that list the number of tissues in which the antisense transcript oscillated, the number in which the sense transcript oscillated, the number in which both oscillated, and the maximum difference in peak expression phase across all tissues.

Supplementary_digital_file_S3.xls - Loci for oscillating, novel transcripts
This file lists the genomic coordinates for each putative novel transcript with rhythmic expression, identified in chapter 3 of this thesis. It also lists the number and name of each tissue in which each novel transcript oscillated.

Supplementary_digital_file_S4.xls - List of transcripts with associated fold-change values in within-transcript coverage
This file lists the fold-change differences in coverage across all library preparation protocols for each IVT transcript in chapter 4.

Appendix B – Novel transcript sequences and qPCR amplicons

Novel Chromosome 6 Transcript - Predicted Sequence

```
>Ex1 chr6:121087472-121087563
AGAGCAGACTGAGAAAGCCATGTGGAGCAAGCCAGTAAACAGCACTCCTC
CACAGCCTCTGCATCAGCTCCTGCTCCAGGTTCCTGCCTCAT
>Ex2 chr6:121091595-121091797
GTTTGGCTGAAAGTGGATAAAGCTCTATTGTAAGAGACCCCGTGGGGTAG
AAAGCAGGAAGAGCTGTGTAGTTTGAAGCAACAAGCCAGAAAGACATATC
CTATGTTCCTGGCCTGTGTGTACCTCTGTACCTGGGGACCATTCATGTTC
CTTCATGGACCAGATCACATCTACTGGCGTGGAGACCAGAGGCCAGATTT
AAG
>Ex3 chr6:121093022-121093132
CTTTTCTTGCAGTCTCCGTGACATTGGGCTGTATGTGCCTCATTAGATGG
GATTCCATCCAGTGCGTCCCAGGAACCGCCCTGCCATCTCACTTAAGATT
CTTAGCAACCT
>Ex4 chr6:121096091-121096223
GTCATCCTGGGTCACAGACGCTCGTTCGCTCACTCCATCGTGGACCAGCA
GCTGTCAGCCAGTGCTGCCCACAGCTGTGTTCTCTGCCTCATGCTGCTTT
GCAGACCATTGAAAGCATCTCCCGCTCAGTGAG
>Ex5 chr6:121096313-121097060
GTCTGGATTTGAACCCAGGCCCCATGCATGCTAATGGGGAGTGAAGTGCC
ACAGCCCAGGGTCTTCTTTTGACAGCCTAGTCCGTCCACCTTGTCCACAA
GATGTCCAGGTTAGGCCTGAGCCTAACCTGCTTCCCTTCCCCTTGCCAAC
TGCTCTCCATCTATGAATGGTCCCCGGGAATTAGGAAGAATGGGGTGGGA
GTGGGATTGGCCACTCCTAGGAGTCTGGTGTTTGTTTCATCTCCTCTATA
AATGGTGTAAGAAAGATGTCCCAGCAAGACGTAGTAGCACACATCTGGGA
TACTAGCACTCAAGAGGTAGAGGCAAGGGTTATCCTGAGCTACATAGCCT
ATCCAAAACCAGTCTGGGTTACAGGAGACCCTGGTTTTATTCATTCATTC
ATTTATGATATTCTCTTTATCCAGTTCCAACCTAAACTAAATAAATTAAA
ATTTCAGTTTCTCTCTCTATGTAATTTTTCTCTTAGAGGAATCATTTGGA
TCCACAGCCAGGCGAATCCTGCAGTGGCATCTGCACTCTTGGGGCTGTGG
CCTTTGCATCCCTTAGCCCATGCCAGTTTCCAAAGCAGGCTGTGGGTGTC
TCTCAGAATGTACAATCAGTTTTTCCAGAGGCCTTGACATACTCCCATCC
CCACACCGAGTCTTTGCCTGTGATTTCTGGAAACGGTCCTGTTTCCCTTT
TCTCTCAGCCTTAAGTGGATTCTACCTTGGGTTCAGAGGAAGTTCAAG
```

Novel Chromosome 7 Transcript - Predicted Sequence

```
>Ex1 chr7: 35913467-35914000
TAACTTTAGATACATACCTGGCTTCGTGAAAAGGTAAATACATACCTTTT
CATTCAGAAAGCTGTTCCCTGAAAGAGTTGGGAGGCTTTGAATTCCCTTC
ATCACTGGGAGGAGAGCAGGAGCCATACACTGTGTGTGAGGGGTGTGGGG
ACCATTCCTACTGTCTCCCTCCCTCTGTTCTCCCCTAAATGATCCTGAGC
CAGGAAGAGTGATGCAGAATGTCTCTCACCCTGTGGAAGAGTTGGTCAGG
CTGGTCCTCAGACAACCAGGGAAGCTCTTGGGGTCCTGGAGAATAGGCAC
ATAGCAGATAAAAGGAGTTCTTAACCAAACTTCCCTAGAACGGAGGGAGC
TAACAAGAAAGAACTTTGGAAATCTACCCTCCTCTTTCCCTGTCACTGCC
AGGAATGTCACCATGAGAGCAGTTTCAGTTAATGAGCAAACTCCTCAGAC
AAGGCAGGAAGGCAGCTCTTGGGCCTCACTGTCAAGCACAGGAAGCGACT
GGATTCCACTTGCCCGGTGTAGGGATGACAGCAG
>Ex2 chr7:35922009-35922185
GTCAGCTCCTCTCCAATTACCCAGAATGGGTGAGAGTTGAAGATCTAGAA
ACAGATTGCTGAGGCAGGAGAACAAATGGTGACCCCTGCGCCTGCCCACG
TTCCTGGAGGAAAACAGGAGCCAGTAGAGAGTGAGCTGACTCGGTGTGAC
TCCCTTAACATTGACACAAAAGAGAAA
>Ex3 chr7:35924154-35924250
CTATCATAGGTCATGGGTACCAGCATCCCCAGTTCTCATCAGGAAGCAGT
ACCCCAGGGCACATGTCCCACAGACTGCTGAAAGAGATCTTGCTCAG
>Ex4 chr7:35924466-35924775
GGTTTCCCAGGATGAGGGGGTGCAGGAGCCCCCCAGAAGTGCCCATCCTC
TAACCAACTGGGGGCCAGCTAAGCAGACATCTTGGTCAAGCCTCAGCCCG
TAGACTTGTGCTCTTGTGGCATAAGATGAGCCTCTGGGAACCAACTCGAG
ACCTACTGTTTGGGAGTTCGCCGGAGCAACGAGCCCCTTCTGAGGCCTCT
AAGCTGCTGAGCTCTGCAGGATTGAGGTCATGATCCCTGCCATGTTCCAG
AGGCTTCACAAGAAGATGAAGGGACCCAGGAGGAGATTGTGGGTCCATGG
GAACTGTCAG
>Ex5 chr7:35927234-35927958
GCCCCAGAGAACCTCCTAGCCCATGCTGGAAGAGAAGGCCATTCCATCTG
GGAATCACATGGCACTGGGTGGAGAGAGAACCGACTGGGCCTGACGCCTT
GCAGAACCAGCATCCAGCCTGTGTCCAAAGTGCTCCTGGAACCACAGAAT
GTTTCATGCCTCCAACCCTGCCCCCTCTGTCTGTCTGTCTGTTTGTCCAT
CTGTCCATCTGTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCTCT
CTGTGTGTGTGTGTGTGTGTGTGTTACATACATGTATAACATTCAACACT
CACACTCATGCTCTGGCCCATGACTTTATCCATGTGCTCAGCATCCTCCC
ATCCAGAAATCTCCAAGGCTCCCAACTGCACAGGCGACCAGCCTGAGTTC
TCTGCCGGTCACTCCCAGGCTCCACCCAGCCCTTCCCCAGCCCCTTCACA
CAGCCCAGCTCAGCCAGGCAGGTCTCCCCTAGGGCACTATTTTCGAGCAC
ACTTTCTCTGCAGATGTTCAGTCATTTGCACCTGGGGCTGCTCTGCATTT
GTGCATCAGCCTCAGTAGGGGTGTCTGAATTGGGTGAGAGGGTGTGTCGG
AGAAGAATCCTGACTTGCATTGGTTTGAGCAGATAACGAATAAAAAGTGT
TCTCGCCGTGGAAAAATCCAGACGGGTGTTGGCCCAGGCACAGCTCTAGG
TCAAAATTTTAAGAACCCCCACTCA
```

Novel Chromosome 6 Transcript – qPCR Amplicon Sequences

: = Site of exon-exon junction in predicted sequence

>Ex1-2 Amplicon
GCATCAGCTCCTGCTCCAGGTTCCTGCCTCAT:GTTTGGCTGAAAGTGGA
TAAAGCTCTATTGTAAGAGACCCCGTGGGGTAGAAAGC

>Ex1-3 Amplicon
CAGCCTCTGCATCAGCTCCTGCTCCAGGTTCCTGCCTCAT:GTTTGGCTG
AAAGTGGATAAAGCTCTATTGTAAGAGACCCCGTGGGGTAGAAAGCAGGA
AGAGCTGTGTAGTTTGAAGCAACAAGCCAGAAAGACATATCCTATGTTCC
TGGCCTGTGTGTACCTCTGTACCTGGGGACCATTCATGTTCCTTCATGGA
CCAGATCACATCTACTGGCGTGGAGACCAGAGGCCAGATTTAAG:CTTTT
CTTGCAGTCTCCGTGACATTGGGCTGTATGTGCCTCATTAGATGGGATTC
CATCCAGTGCGTCCCAGGAACC

>Ex2-4 Amplicon
CCGTGGGGTAGAAAGCAGGAAGAGCTGTGTAGTTTGAAGCAACAAGCCAG
AAAGACATATCCTATGTTCCTGGCCTGTGTGTACCTCTGTACCTGGGGAC
CATTCATGTTCCTTCATGGACCAGATCACATCTACTGGCGTGGAGACCAG
AGGCCAGATTTAAG:CTTTTCTTGCAGTCTCCGTGACATTGGGCTGTATG
TGCCTCATTAGATGGGATTCCATCCAGTGCGTCCCAGGAACCGCCCTGCC
ATCTCACTTAAGATTCTTAGCAACCT:GTCATCCTGGGTCACAGACGCTC
GTTCGCTCACTCCA

Novel Chromosome 7 Transcript – qPCR Amplicon Sequences

: = Site of exon-exon junction in predicted sequence

>Ex1-3 Amplicon
AAGGCAGCTCTTGGGCCTCACTGTCAAGCACAGGAAGCGACTGGATTCCA
CTTGCCCGGTGTAGGGATGACAGCAG:GTCAGCTCCTCTCCAATTACCCA
GAATGGGTGAGAGTTGAAGATCTAGAAACAGATTGCTGAGGCAGGAGAAC
AAATGGTGACCCCTGCGCCTGCCCACGTTCCTGGAGGAAAACAGGAGCCA
GTAGAGAGTGAGCTGACTCGGTGTGACTCCCTTAACATTGACACAAAAGA
GAAA:CTATCATAGGTCATGGGTACCAGCATCCCCAGTTCTCATCAGGAA
GCAGTACCCCAGGGCACATGTCCCACAGACTGC

>Ex2-4 Amplicon
CAAATGGTGACCCCTGCGCCTGCCCACGTTCCTGGAGGAAAACAGGAGCC
AGTAGAGAGTGAGCTGACTCGGTGTGACTCCCTTAACATTGACACAAAAG
AGAAA:CTATCATAGGTCATGGGTACCAGCATCCCCAGTTCTCATCAGGA
AGCAGTACCCCAGGGCACATGTCCCACAGACTGCTGAAAGAGATCTTGCT
CAG:GGTTTCCCAGGATGAGGGGGTGCAGGAGCCCCCCAGAAGTGCCCAT
CCTCTAACCAACTGGGGGCCAGCTAAGCA

Appendix C – Detailed RNA-seq library construction protocol

Dynabead Purification (Invitrogen 610-06 – Dynabeads mRNA Purification Kit):
The following Dynabead protocol was adapted from Invotrogen's Dynabeads mRNA Purification Kit Protocol (rev no. 004), and the protocol used by the Gilad Lab [65].

1) Dilute 40 µg total RNA in 80 µL of RNase-free H_2O.
2) Heat RNA samples to 65ºC for 5 minutes, and then quickly put on ice. Keep samples on ice until step 6.
3) Re-suspend Dynabeads through vigorous vortexing. Transfer 160 µL of re-suspended beads to RNase-free 1.5mL microcentrifuge tube. Place tubes on magnetic stand, remove supernatant, and remove tubes from magnetic stand.
4) Re-suspend beads in 160 µL of Binding Buffer. Place tubes on magnetic stand, remove supernatant, and remove tubes from the magnetic stand.
5) Repeat step 4.
6) Add 80 µL of Binding Buffer to beads and combine with total RNA sample from 2. Mix by flicking tubes until resuspended.
7) Rotate for 5 minutes at room temperature.
8) Place tubes on magnetic stand, remove supernatant, and remove tubes from the magnetic stand.
9) Wash beads with 160 µL of Washing Buffer B. Place tubes on magnetic stand, remove supernatant, and remove tubes from the magnetic stand.
10) Repeat step 9.
11) Prepare a clean tube for each sample containing 140 µL of Binding Buffer. This tube will prepare the poly(A)+ selected mRNA for the second round of Dynabeads.
12) Remove supernatant from step 9, add 20 µL of Elution Buffer to beads, and mix by flicking.
13) Heat the beads to 80ºC for 2 minutes. Immediately put beads on magnetic stand and transfer supernatant to tubes containing Binding Buffer prepared during step 11. It is important to keep the eluting beads as hot as possible while removing the supernatant from the beads. As the eluate cools, the mRNA will start re-binding to the beads. It might be best to perform this step one sample at a time to keep the temperature as high as possible.
14) Heat eluted RNA samples from step 13 to 65ºC for 5 minutes, then quickly put on ice. Keep samples one ice until step 17.
15) Add 160 µL of Washing Buffer B to used beads from step 13. Place tubes on magnetic stand, remove supernatant, and remove tubes from the magnetic stand.
16) Repeat step 13.
17) Remove supernatant from beads from step 16 and add RNA sample from step 14. Rotate tubes at room temperature for 5 minutes.
18) Repeat steps 7-10 with the new samples.
19) Remove supernatant from step 18, add 9 µL of Elution Buffer to beads, and mix by flicking. Heat beads to 80ºC for 2 minutes. Immediately put beads on the magnetic stand and transfer supernatant to a clean 1.5 mL tube.
20) Holding point: store samples at -80ºC or proceed directly to fragmentation.

Fragment RNA and Precipitate (Ambion AM8740 – RNA Fragmentation Reagents):
*Keep all samples on ice unless otherwise specified (for all steps)

1) Add 1 µL of fragmentation reagent.
2) Mix by gentle flicking.
3) Spin samples briefly.
4) Incubate samples at 70ºC for 5 minutes. Replace on ice immediately.

5) Add 1 µL of stop buffer and mix by gentle flicking.
6) Spin briefly.
7) Ethanol precipitate fragmented mRNA.
    a. Bring sample volume up to 100 µL with nuclease-free H_2O.
    b. Add 10 µL 3M Sodium-Acetate (1/10th volume).
    c. Add 350 µL of 100% ethanol (3.5 volumes) at room temperature.
    d. Shake samples vigorously by hand (don't vortex).
    e. Freeze samples at -80ºC. Precipitate for at least 45 minutes (preferably overnight).
8) Holding point: keep samples at -80ºC or proceed to ds cDNA Synthesis.

ds cDNA Synthesis (Invitrogen 11917-010 – SuperScript double-stranded cDNA kit):

1) Complete precipitation:
    a. Spin samples at 4ºC in a refrigerated table-top microcentrifuge for 90 minutes at full speed.
    b. Decant Ethanol while taking care not to disturb the pellet.
    c. Add 750 µL of 80% ethanol.
    d. Spin samples at 4ºC for 5 minutes at full speed.
    e. Decant ethanol while taking care not to disturb the pellet.
    f. Invert tubes and allow pellets to dry for 10 minutes.
    g. Add 9 µL of nuclease-free H_2O to each pellet, mix by gentle flicking, and allow to dissolve on ice for 10 minutes.
    h. Spin samples briefly to collect at the bottom of tubes.
2) Prepare random hexamers (Promega C1181 – Random Primers) by diluting 10-fold. This dilution yields primers at 50 ng/µL.
3) Combine 9 µL of fragmented RNA sample with 4 µL of diluted hexamers in a Nuclease-free PCR tube. Mix by gentle flicking and spin briefly to collect in the bottom of the tube.
4) Heat mixture to 70ºC for 10 minutes, then cool to 4ºC. Once samples have cooled to 4ºC, place them on ice.
5) Combine the following reagents to create the first-strand synthesis master mix:

	Per Reaction
5x 1st-Strand Reaction Buffer	4 µL
0.1 M DTT	2 µL
10 mM dNTP mix	1 µL
SuperScript II (add last)	1 µL

    Add 8 µL of the above reaction mix to each sample/primer mix. Mix gently and spin.
6) Incubate sample-reactions at 45°C for 62 minutes, then cool to 4°C. Once sample-reactions have cooled to 4°C, place them on ice.
7) Combine the following reagents to create the second-strand synthesis master mix:

		Per Reaction
	Nuclease-free H_2O	91 µL
	5x 2nd-Strand Reaction Buffer	30 µL
	10 mM dNTP mix	3 µL
Add last	*E. coli* DNA Ligase	1 µL
	E. coli DNA Polymerase I	4 µL
	E. coli RNase H	1 µL

    Add 130 µL of the above reaction mix to each sample-reaction. Mix gently and spin.
8) Incubate sample-reactions at 16°C for 3 hours. After incubation, place the PCR tubes on ice (no need for cool-down this time).

9) Add 2 µL of T4 DNA Polymerase to each sample-reaction. Mix gently and spin.
10) Incubate sample-reactions at 16°C for 5 minutes, then cool to 4°C. Once sample-reactions have cooled to 4°C, place them on ice.
11) Add 10 µL of 0.5 M EDTA (**not** included with kit) to each sample-reaction. Mix gently and spin.
12) Perform Phenol:Chloroform extraction on all sample-reactions:
    a. Transfer reactions to clean, nuclease-free 1.5 mL microcentrifuge tube.
    b. Add 170 µL of Phenol:Chloroform to each sample. Mix Vigorously by vortexing for 30 seconds.
    c. Spin samples at max speed for 10 minutes.
    d. Remove 145 µL of the upper aqueous phase, and transfer to a clean, nuclease-free 2 mL microcentrifuge tube.
13) Perform ethanol precipitation
    a. Add 14.5 µL 3M Sodium-Acetate (1/10th volume).
    b. Add 560 µL of 100% ethanol (~3.5 volumes) at room temperature.
    c. Shake samples vigorously by hand (don't vortex).
    d. Freeze samples at -80ºC. Precipitate for at least 45 minutes (preferably overnight).
14) Holding point: keep samples at -80ºC or proceed to End Repair.

End Repair (Epicentre ER0720 – End-It DNA End-Repair Kit):

1) Complete precipitation:
    a. Spin samples at 4ºC in a refrigerated table-top microcentrifuge for 90 minutes at full speed.
    b. Decant Ethanol while taking care not to disturb the pellet.
    c. Add 750 µL of 80% ethanol.
    d. Spin samples at 4ºC for 5 minutes at full speed.
    e. Decant ethanol while taking care not to disturb the pellet.
    f. Invert tubes and allow pellets to dry for 10 minutes.
    g. Add 30 µL of nuclease-free H_2O to each pellet, mix by gentle flicking, and allow to dissolve on ice for 10 minutes.
    h. Spin samples briefly to collect at the bottom of tubes.
2) Combine the following reagents to create the end repair master mix:

	Per Reaction
10x End-Repair Buffer	5 µL
2.5 mM dNTP Mix	5 µL
10 mM ATP	5 µL
Nuclease-free H_2O	4 µL
End-Repair Enzyme Mix (add last)	1 µL

    Combine 31 µL of cDNA from each sample with 20 µL of the above reaction mix in RNase/DNase-free PCR reaction tubes (50 µL total reaction volume). Mix gently and spin.
3) Incubate sample-reaction at room temperature for 45 minutes.
4) Purify reaction products with the QIAquick PCR Purification Protocol and QIAquick columns, with the following modifications:
    a. 3 minute evaporation spin.
    b. Elute in 34 µL of EB and let columns stand for 5 minutes.
    c. 2 minute elution spin.

Add 'A' base to 3' Ends (NEB MO212s – Klenow Fragment 3' → 5' exo$^-$; 1 mM dATP – not included with Klenow kit):

1) Combine the following reagents to create the add ‘A’ master mix:

	Per Reaction
Klenow Buffer (NEB Buffer 2)	5 µL
1 mM ATP	10 µL
Klenow Fragment (3’ to 5’ exo$^-$) (add last)	1 µL

Combine 34 µL of cDNA from each sample with 16 µL of the above reaction mix in RNase/DNase-free PCR reaction tubes (50 µL total reaction volume). Mix gently and spin.

2) Incubate sample-reaction at 37°C for 30 minutes.
3) Purify reaction products with the MinElute PCR Purification Protocol and QIAquick MinElute columns, with the following modifications:
   a. 3 minute evaporation spin.
   b. Elute in 10 µL of EB and let columns stand for 5 minutes.
   c. 2 minute elution spin.

Ligate Adapters (Promega M8221 – LigaFast Rapid DNA Ligation System; Illumina PE-102-1001 – Paired-End Adapters from PE DNA Sample Prep Kit):

1) Dilute the Paired-end adapter mix 1:10 in DEPC-treated H_2O before beginning.
2) Combine the following reagents to create the adapter ligation master mix:

	Per Reaction
DNA Ligase Buffer	15 µL
Nuclease-Free H_2O	2 µL
Adapter oligo mix (1:10 dilution)	1 µL
DNA Ligase (add last)	2 µL

Combine 10 µL of cDNA from each sample with 20 µL of the above reaction mix in RNase/DNase-free PCR reaction tubes (30 µL total reaction volume). Mix gently and spin.

3) Incubate sample-reaction at room temperature for 15 minutes.
4) Purify reaction products with the MinElute PCR Purification Protocol and QIAquick MinElute columns, with the following modifications:
   a. 3 minute evaporation spin.
   b. Elute in 15 µL of EB and let columns stand for 5 minutes.
   c. 2 minute elution spin.

Purify Ligation Products (Invitrogen 10416-014 – 50 bp DNA Ladder):

1) Prepare 2% agarose gel by dissolving 4 g of agarose in 200 mL of TAE. Add ethidium bromide to a final concentration of 0.5 µg/ml.
2) Pour separate gels for each sample or a single large gel. Both gel configurations should be 14 cm in length.
3) Combine the following reagents to create a 50 bp Ladder master mix:

	Per Reaction
50 bp Ladder	1 µL
Nuclease-Free H_2O	14 µL
6x Loading Buffer	5 µL

Prepare ladder by combining 1 µL of 50 bp Ladder, 5µL of 6x loading buffer, and 14 µL of Nuclease-free H_2O. Prepare 20 µL of 50 bp ladder mix for ever two cDNA samples.

4) Prepare samples by combining 15 µL cDNA sample and 5 µL 6x loading buffer.
5) Load ladder and samples into gel(s). For separate gels, leave an empty gel lane between the ladder and the sample. For a single large gel, load samples like so ("-" is an empty cel):
   – sample1 – ladder – sample2 – sample3 – ladder – sample4 – sample5 – ladder – sample6 – sample7 – ladder – sample8 –
   These steps facilitate cutting of gels later in this protocol.
6) Run gel at 150v for 51 minutes.
7) Image gel(s). Be sure to limit gel exposure time to UV light.
8) Cut slices from gel(s)s in the 350-500bp range (and/or any other lengths of interest) for each sample. Cut each sample with a new, clean blade. For a single large gel, divide the gel first such that each gel section contains two sample lanes separated by a ladder lane. Cute slices from these smaller gel pieces. Gel slices can be stored at -20°C.
9) Image gel(s) following cuts.
10) Holding point: Keep gel slices at -20°C or proceed to PCR amplification.

PCR Amplification (NEB F-531S – Phusion High-Fidelity PCR Master Mix):

1) Purify DNA from gel slices with the Qiagen Gel Extraction Kit, with the following modifications:
   a. 3 minute evaporation spin.
   b. Elute in 23 µL of EB and let columns stand for 5 minutes.
   c. 2 minute elution spin.
2) Combine the following reagents to create the PCR amplification master mix:

	Per Reaction
2x Phusion HF Mix	25 µL
25 µM Paired-End Primer Mix (sequences included at the end of this protocol)	2 µL

   Note that the Phusion Mix is a master mix that requires thawing.
   Combine 23 µL of cDNA from each sample with 27 µL of the above reaction mix in RNase/DNase-free PCR reaction tubes (50 µL total reaction volume). Mix gently and spin.
3) Incubate according to the following PCR protocol:
   a. 30 seconds at 98°C.
   b. 10 seconds at 98°C.
      30 seconds at 65°C.
      60 seconds at 72°C.
      Cycle 13 times
   c. 10 minutes at 72 °C.
   d. Hold at 4°C.
4) Purify reaction products with the QIAquick PCR Purification Protocol and QIAquick columns, with the following modifications:
   a. 3 minute evaporation spin.
   b. Elute in 50 µL of EB and let columns stand for 5 minutes.
   c. 2 minute elution spin.

Double-Stranded cDNA Synthesis PCR Program:
This thermocycler program is used to perform the Double-stranded cDNA synthesis step above.

1) 10 minutes at 70°C.

    - Incubate sample/primer mixture.

2) Hold at 4°C.
    - Once the samples have cooled to 4°C (~2 minutes), put them on ice, and add the 1st-Strand Master Mix.
    - Skip the remainder of this step once the master mix is added to the sample/primer mixture.
3) 62 minutes at 45°C.
    - Incubate 1st-strand synthesis sample-reactions
4) Hold at 4°C.
    - Once the samples have cooled to 4°C (~2 minutes), put them on ice, and add the 2nd-Strand Master Mix.
    - Skip the remainder of this step once the master mix is added to the sample/primer mixture.
5) 4 hours at 16°C.
    - Incubate 2nd-strand synthesis sample-reactions.
    - After 3 hours, put sample-reactions on ice and add 2 µL of T4 DNA Polymerase to each sample-reaction.
    - Once T4 DNA Polymerase has been added to each sample-reaction, return them to the PCR machine to incubate for 5 minutes.
6) Hold at 4°C.

Illumina Paired-End PCR Primers:

1) PE Primer 1.0:
   5’-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT-3’
2) PE Primer 2.0:
   5’-CAAGCAGAAGACGGCATACGAGATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3’

When ordering primers of this length, be sure to get them HPLC- or PAGE-purified to reduce the number of prematurely-terminated primers generated during the synthesis process.

Appendix D – Accession numbers for IVT transcripts

*indicates IVT transcript was removed from analysis for mapping to multiple loci, or overlapping another IVT transcript.

BC000128	BC001001	BC006772	BC009084	BC011353	BC015050
BC000129	BC001003	BC006781	BC009139	BC011355	BC015052
BC000130	BC001226	BC006784*	BC009523	BC011359*	BC015054
BC000131	BC001228	BC006786*	BC009524	BC011361	BC015056
BC000132	BC001229	BC006791	BC009529	BC011362	BC015164*
BC000133	BC001231	BC006793*	BC009530	BC011363	BC015165
BC000134	BC001232*	BC006794	BC009538*	BC011365*	BC015169*
BC000135	BC001233*	BC006795	BC009540	BC011368	BC015171
BC000138	BC001234	BC006804	BC009545	BC011369	BC015180
BC000140	BC001235	BC006807	BC009548*	BC011371	BC015202
BC000141	BC001236	BC006808	BC009552	BC011372	BC015219*
BC000142	BC001238*	BC006811	BC009553	BC011375	BC015231
BC000145	BC001239	BC006818	BC009561	BC011377	BC015236
BC000146	BC001240	BC006819	BC009564	BC011379	BC015474
BC000147	BC001241	BC006821	BC009614	BC011380	BC015480
BC000148*	BC001242	BC006823	BC009617	BC011381	BC015489
BC000149	BC001243	BC006825	BC009618	BC011382	BC015490
BC000150	BC001244	BC006831	BC009620	BC011384	BC015505
BC000151	BC001245	BC006832	BC009621	BC011387	BC015507
BC000152	BC001247	BC006837	BC009623	BC011392	BC015513
BC000153	BC001249*	BC006838	BC009624	BC011393	BC015541
BC000154	BC001250	BC006839*	BC009631	BC011394	BC015542
BC000155	BC001251	BC006849*	BC009642	BC011396	BC015555
BC000157	BC001252	BC006850	BC009644	BC011399	BC015701
BC000158	BC001253	BC007121	BC009670	BC011400	BC015703*
BC000159	BC001254	BC007122	BC009671	BC011402	BC015704
BC000160	BC001255	BC007123	BC009672	BC011404	BC015706
BC000161	BC001256	BC007124*	BC009674	BC011405	BC015708
BC000162	BC001257	BC008087	BC009675	BC011406	BC015710
BC000163	BC001258	BC008090	BC009677	BC011408	BC015711
BC000165	BC001259	BC008091	BC009678	BC011410	BC015712
BC000166	BC001261	BC008092*	BC009679	BC011414	BC015713

BC000168	BC001262	BC008094*	BC009680	BC011418	BC015714
BC000169	BC001263	BC008095	BC009681	BC011419	BC015715
BC000170	BC001264	BC008100	BC009684	BC011453	BC015722
BC000171	BC001265	BC008145	BC009685	BC011454	BC015725
BC000172	BC001267	BC008146	BC009686	BC011460	BC015794
BC000175	BC001268	BC008149	BC009687	BC011498	BC015796
BC000176	BC001269*	BC008151	BC009689	BC011502	BC015797
BC000177	BC001270	BC008178*	BC009691	BC011515	BC015799
BC000178	BC001271	BC008180	BC009693	BC011517	BC015801
BC000179	BC001272	BC008182	BC009694	BC011519	BC015802
BC000180*	BC001273	BC008183	BC009696	BC011520	BC015803
BC000181	BC001274	BC008185	BC009697	BC011522	BC015806
BC000793	BC001275	BC008188	BC009698	BC011523	BC015807
BC000794	BC001276	BC008191	BC009699	BC011524	BC015808
BC000795	BC001277	BC008194	BC009701	BC011526	BC015809
BC000797	BC001278	BC008195	BC009703	BC011529	BC015810
BC000799	BC001279	BC008196	BC009704	BC011534	BC015812
BC000802	BC001280	BC008197	BC009707	BC011535	BC015813
BC000803*	BC001281	BC008198*	BC009708	BC011537	BC015814
BC000804	BC001282	BC008200	BC009709	BC011538	BC015882
BC000805	BC001283	BC008201	BC009710	BC011539	BC015883*
BC000806	BC001284	BC008202	BC009711	BC011542	BC015886
BC000807	BC001285	BC008203	BC009712	BC011948	BC015887
BC000808*	BC001286	BC008205	BC009713	BC011992	BC015888*
BC000809	BC001287	BC008207	BC009714	BC012037	BC015890*
BC000810	BC001288	BC008211	BC009715	BC012040	BC015891
BC000813	BC001289*	BC008212	BC009716	BC012201	BC015893
BC000814	BC001291	BC008214	BC009717	BC012302*	BC015899
BC000819	BC001293	BC008215	BC009718	BC012304	BC015904
BC000820	BC001294	BC008219	BC009719	BC012372*	BC015925
BC000823	BC001295	BC008226	BC009720	BC012850	BC015926
BC000824	BC001297	BC008235	BC009722	BC012857*	BC015927
BC000827	BC001298*	BC008246	BC009726	BC012860	BC015928
BC000829	BC001300	BC008250	BC009727	BC012890*	BC015930
BC000830	BC001301*	BC008251	BC009731	BC012895	BC015931
BC000832	BC001302	BC008253	BC009733*	BC012925	BC015932
BC000834	BC001303	BC008254	BC009734	BC012926	BC015934*
BC000835	BC001304	BC008505*	BC009737	BC012932	BC015935

BC000836	BC001305	BC008506	BC009738	BC012941*	BC015936
BC000837	BC001308	BC008567	BC010235	BC012942	BC015937
BC000846	BC001309	BC008568	BC010439	BC012944	BC015938
BC000849	BC001310	BC008569	BC010441	BC012950	BC015939
BC000850	BC001311	BC008572	BC010444	BC013045	BC015940
BC000851	BC001312	BC008573	BC010446	BC013073	BC015941
BC000852*	BC001316	BC008584	BC010449	BC013142	BC015943
BC000853	BC001317	BC008585*	BC010450	BC013153	BC015944
BC000854	BC001318	BC008586*	BC010451	BC013155	BC015945
BC000855	BC001319	BC008594*	BC010456	BC013158	BC015946*
BC000857*	BC001321	BC008600	BC010458	BC013425	BC015947
BC000861	BC001326	BC008602	BC010460	BC013426	BC015948
BC000864	BC001327	BC008603	BC010463	BC013428	BC015949
BC000866	BC001328*	BC008604	BC010464	BC013433	BC016024
BC000868	BC001329	BC008605	BC010466	BC013435	BC016025
BC000870	BC001331	BC008607	BC010469	BC013436	BC016026
BC000871	BC001333*	BC008608	BC010471	BC013437	BC016028
BC000872	BC001334	BC008611	BC010522	BC013439*	BC016029
BC000873	BC001337	BC008613	BC010537	BC013566	BC016031
BC000875	BC001338	BC008618	BC010569	BC013567	BC016137
BC000877	BC001340	BC008621	BC010570	BC013568	BC016139
BC000878*	BC001341	BC008624	BC010571	BC013569	BC016140
BC000879	BC001342	BC008625	BC010574	BC013572	BC016145
BC000881	BC001344	BC008628*	BC010576	BC013575	BC016146
BC000882	BC001345	BC008629	BC010609	BC013576	BC016147
BC000884	BC001346	BC008631	BC010611	BC013577	BC016148
BC000887	BC001964	BC008632	BC010614	BC013580	BC016172
BC000889*	BC001965	BC008634	BC010616	BC013581	BC016174
BC000890	BC001966	BC008636*	BC010618	BC013583	BC016178
BC000891	BC001967*	BC008640	BC010620	BC013584	BC016179
BC000892	BC001968	BC008641	BC010623	BC013585	BC016277
BC000893	BC001970	BC008650	BC010626	BC013587	BC016279
BC000894	BC001971	BC008651	BC010628	BC013588	BC016281
BC000895	BC001972	BC008652	BC010629	BC013589	BC016282
BC000896	BC001979	BC008654	BC010632	BC013590	BC016283
BC000897	BC001980	BC008656	BC010634	BC013591	BC016284
BC000898	BC003352	BC008658*	BC010640	BC013592	BC016285
BC000899	BC003353	BC008659*	BC010641	BC013596	BC016286

BC000901	BC003354	BC008662	BC010647	BC013597	BC016288
BC000902	BC003355	BC008663	BC010648	BC013609	BC016292
BC000903	BC003356	BC008664	BC010649	BC013645	BC016294
BC000904	BC003357	BC008666	BC010652	BC013690	BC016295
BC000905	BC003358	BC008667	BC010653	BC013693	BC016445
BC000906	BC003359	BC008668	BC010658	BC013748	BC016472*
BC000907	BC003360*	BC008669	BC010659	BC013760	BC016474
BC000908*	BC003361	BC008671	BC010660*	BC013781	BC016509
BC000910	BC003362	BC008673	BC010661	BC013787	BC016582
BC000912	BC003364	BC008674*	BC010662	BC013788	BC016613
BC000913	BC003365	BC008675	BC010665	BC013789	BC016614*
BC000914	BC003366*	BC008678	BC010668	BC014787	BC016617
BC000915	BC003367	BC008679	BC010671	BC014789	BC016622
BC000916	BC003369	BC008680	BC010674	BC014846	BC016623
BC000918	BC003370	BC008682	BC010681	BC014861*	BC016633
BC000921*	BC003371	BC008684	BC010689	BC014879	BC016655
BC000923	BC003373	BC008685	BC010691	BC014880	BC016663
BC000924*	BC003375	BC008686	BC010692	BC014881	BC016664
BC000926	BC003376	BC008688	BC010696	BC014885	BC016852
BC000927	BC003377	BC008689	BC010697	BC014887	BC017025
BC000930*	BC003378	BC008690	BC010698	BC014888	BC017045
BC000931	BC003379	BC008691	BC010701	BC014889	BC017061
BC000932	BC003381	BC008972	BC010703	BC014890	BC017094
BC000933	BC003382	BC008973	BC010704	BC014891	BC017114
BC000934	BC003383	BC008975*	BC010708	BC014894	BC017115
BC000936	BC003384	BC008976	BC010732	BC014896	BC017117
BC000937*	BC003385*	BC008979	BC010734	BC014897	BC017119
BC000938	BC003387	BC008981	BC010735*	BC014898	BC017123
BC000939*	BC003388	BC008982	BC010737	BC014900	BC017163
BC000941	BC003389	BC008983	BC010738	BC014901*	BC017168
BC000942	BC003390	BC008984	BC010739	BC014904	BC017169
BC000944	BC003394	BC008986	BC010740	BC014907	BC017453
BC000948	BC003395	BC008987	BC010743	BC014908*	BC017469
BC000949	BC003397	BC008988	BC010744	BC014911	BC017471
BC000952	BC003398	BC008990	BC010846	BC014912	BC017472
BC000953	BC003400	BC008991	BC010849	BC014913	BC017492
BC000954	BC003401*	BC008992	BC010850	BC014916	BC017495*
BC000956*	BC003402	BC008993	BC010852	BC014918	BC017553

BC000959	BC003403	BC009009*	BC010853*	BC014919	BC017554
BC000961	BC003407*	BC009010	BC010854	BC014923	BC017555
BC000962	BC003408	BC009011	BC010855	BC014924	BC017556
BC000964	BC003409	BC009012	BC010856	BC014928	BC017558
BC000965	BC003410	BC009014	BC010857	BC014939*	BC017559
BC000966	BC003412	BC009015	BC010858	BC014940	BC017655
BC000968	BC003413	BC009016	BC010859	BC015012	BC017673
BC000971	BC003417	BC009017	BC010860	BC015013	BC018118
BC000972	BC003418	BC009025	BC010861	BC015014	BC018130
BC000973	BC003682	BC009026	BC010862	BC015016	BC018164
BC000974	BC003683	BC009031	BC010863	BC015017	BC018207
BC000977	BC003684	BC009032*	BC010866	BC015018	BC018295
BC000978	BC003685	BC009037	BC010868	BC015020*	BC018337
BC000979	BC003686*	BC009039	BC010874	BC015022	BC018349
BC000980	BC003688	BC009040	BC010876	BC015025	BC018426
BC000981	BC003689*	BC009041	BC011046	BC015026	BC018445
BC000983*	BC003690	BC009046	BC011047	BC015027	BC018466
BC000985	BC003691	BC009047	BC011048*	BC015028	BC018509
BC000986	BC003694	BC009048	BC011049	BC015030*	BC018514
BC000988	BC003701	BC009049	BC011051	BC015031	BC018528*
BC000989	BC004815	BC009051	BC011054	BC015032	BC020265
BC000990	BC004816	BC009052*	BC011057	BC015033	BC020492
BC000991	BC004817*	BC009053*	BC011175	BC015037	BC020493
BC000992	BC004818	BC009054	BC011249	BC015038	BC020494
BC000993	BC004819	BC009055	BC011262	BC015039	BC020518*
BC000994	BC004822*	BC009065	BC011263	BC015041	BC020965
BC000995	BC005404	BC009073	BC011267	BC015044	BC020973
BC000996	BC005408	BC009074	BC011268	BC015045	BC021892
BC000997	BC005700	BC009077	BC011348	BC015046	BC021958
BC000998*	BC006768*	BC009078	BC011349*	BC015047*	BC021959
BC001000	BC006769	BC009081	BC011350	BC015049	BC022096

Appendix E – List of hunc regions

Chr	Start	Stop	Accession	Strand	Block Count	Block Length	Block Starts
chr1	181059172	181059250	BC000128	+	1	78,	0,
chr19	58084981	58085042	BC000130	-	1	61,	0,
chr21	44949456	44949536	BC000153	-	1	80,	0,
chr19	1877291	1877385	BC000158	-	1	94,	0,
chr3	169802008	169802070	BC000181	+	1	62,	0,
chr9	139835174	139835213	BC000850	-	1	39,	0,
chr1	224380163	224380238	BC000961	+	1	75,	0,
chr1	224380275	224380371	BC000961	+	1	96,	0,
chrX	48121201	48121214	BC001003	+	1	13,	0,
chr19	55789027	55789128	BC001236	-	1	101,	0,
chr1	63894682	63894739	BC001253	+	1	57,	0,
chr6	3850619	3850729	BC001261	+	1	110,	0,
chr16	54147446	54147491	BC001284	+	1	45,	0,
chr11	60689485	60689526	BC001309	+	1	41,	0,
chr8	120846893	120847048	BC001316	-	1	155,	0,
chr11	66839035	66839073	BC001338	+	1	38,	0,
chr7	100075015	100075078	BC001966	-	1	63,	0,
chr20	60887523	60887695	BC003355	-	2	64,8,	0,164,
chr17	26672748	26672831	BC003694	+	1	83,	0,
chr15	101605981	101606062	BC005408	+	1	81,	0,
chr2	128460639	128460658	BC006808	+	1	19,	0,
chr2	128460786	128460885	BC006808	+	1	99,	0,
chr3	172538777	172538883	BC006838	+	1	106,	0,
chrX	153247810	153247972	BC008203	+	1	162,	0,
chr12	133587531	133587671	BC008211	+	1	140,	0,
chr7	8272267	8272298	BC008640	-	1	31,	0,
chr7	8257950	8260973	BC008640	-	2	182,69,	0,2954,
chr7	8183560	8198158	BC008640	-	3	41,99,2,	0,13185,14596,
chr4	845649	845679	BC008668	-	1	30,	0,
chr11	702952	703087	BC008671	+	1	135,	0,
chr11	114270761	114270857	BC009041	-	1	96,	0,
chr19	39898915	39899005	BC009693	+	1	90,	0,
chr19	39899600	39899631	BC009693	+	1	31,	0,
chr5	150175494	150175570	BC009719	+	1	76,	0,

chr2	220379553	220379750	BC010439	+	1	197,	0,
chr2	101019094	101023097	BC010441	-	2	22,60,	0,3943,
chr2	101014504	101014536	BC010441	-	1	32,	0,
chr2	101009725	101009843	BC010441	-	1	118,	0,
chr2	101009310	101009425	BC010441	-	1	115,	0,
chr2	101009082	101009194	BC010441	-	1	112,	0,
chr2	101008793	101008837	BC010441	-	1	44,	0,
chr1	156616744	156616942	BC010571	+	1	198,	0,
chr17	26732326	26732350	BC010691	-	1	24,	0,
chr2	232672717	232672809	BC010739	+	1	92,	0,
chr15	75641401	75641430	BC010876	+	1	29,	0,
chr22	29704246	29704382	BC011047	+	1	136,	0,
chr17	71205714	71205742	BC011054	-	1	28,	0,
chr19	39663585	39663681	BC011368	+	1	96,	0,
chr17	38191523	38192027	BC011375	-	2	78,53,	0,451,
chr8	145745721	145745902	BC011377	+	1	181,	0,
chr19	34180216	34180256	BC011380	+	1	40,	0,
chr1	165632263	165632381	BC011410	-	1	118,	0,
chr8	103662373	103662418	BC011538	-	1	45,	0,
chr8	103661393	103661567	BC011538	-	1	174,	0,
chr10	100018914	100019169	BC013153	-	2	70,22,	0,233,
chr10	100016604	100016609	BC013153	-	1	5,	0,
chr19	39913833	39913901	BC013426	+	1	68,	0,
chr19	39914208	39914268	BC013426	+	1	60,	0,
chrX	101970155	101970194	BC013576	+	1	39,	0,
chrX	101970457	101970503	BC013576	+	1	46,	0,
chrX	101970778	101971019	BC013576	+	1	241,	0,
chrX	101971326	101971540	BC013576	+	1	214,	0,
chrX	101972206	101972236	BC013576	+	1	30,	0,
chr3	99514066	99514102	BC013581	+	1	36,	0,
chr3	38180451	38180519	BC013589	+	1	68,	0,
chr17	42989048	42989069	BC013596	-	1	21,	0,
chr16	30536028	30536223	BC013760	-	1	195,	0,
chr9	115449230	115449332	BC014881	-	1	102,	0,
chrX	108787397	108787436	BC014888	+	1	39,	0,
chr11	8710413	8710450	BC014919	+	1	37,	0,
chr16	21191146	21191222	BC015812	+	1	76,	0,
chr15	93595560	93595618	BC015886	-	1	58,	0,

chr15	93588554	93588620	BC015886	-	1	66,	0,
chr8	141525920	141525990	BC015891	+	1	70,	0,
chr8	141526007	141526152	BC015891	+	1	145,	0,
chr1	112991680	112991725	BC016029	+	1	45,	0,
chr17	48046861	48046943	BC016145	+	1	82,	0,
chr3	179472630	179478901	BC016146	+	3	11,27,2,	0,2209,6269,
chr12	122729186	122729257	BC016617	-	1	71,	0,
chr12	122723160	122723196	BC016617	-	1	36,	0,
chr21	45738419	45739268	BC018295	+	2	58,35,	0,814,
chr21	45741639	45741729	BC018295	+	1	90,	0,
chr8	27145514	27145594	BC018337	-	1	80,	0,
chr16	67134042	67134152	BC018509	+	1	110,	0,
chr19	17691641	17691665	BC020492	+	1	24,	0,
chr19	17692293	17692361	BC020492	+	1	68,	0,

References

1. Hastings MH, Reddy AB, Maywood ES: **A clockwork web: circadian timing in brain and periphery, in health and disease**. *Nat Rev Neurosci* 2003, **4**:649–661.

2. Curtis AM, Fitzgerald GA: **Central and peripheral clocks in cardiovascular and metabolic function**. *Ann Med* 2006, **38**:552–559.

3. Green CB, Takahashi JS, Bass J: **The meter of metabolism**. *Cell* 2008, **134**:728–742.

4. Ramsey KM, Bass J: **Obeying the clock yields benefits for metabolism**. *Proc Natl Acad Sci U S A* 2009, **106**:4069–4070.

5. Paschos GK, Baggs JE, Hogenesch JB, FitzGerald GA: **The role of clock genes in pharmacology**. *Annu Rev Pharmacol Toxicol* 2010, **50**:187–214.

6. Marcheva B, Ramsey KM, Buhr ED, Kobayashi Y, Su H, Ko CH, Ivanova G, Omura C, Mo S, Vitaterna MH, Lopez JP, Philipson LH, Bradfield CA, Crosby SD, Jebailey L, Wang X, Takahashi JS, Bass J: **Disruption of the clock components CLOCK and BMAL1 leads to hypoinsulinaemia and diabetes**. *Nature* 2010, **466**:627–631.

7. Musiek ES, Lim MM, Yang G, Bauer AQ, Qi L, Lee Y, Roh JH, Ortiz-Gonzalez X, Dearborn JT, Culver JP, Herzog ED, Hogenesch JB, Wozniak DF, Dikranian K, Giasson BI, Weaver DR, Holtzman DM, Fitzgerald GA: **Circadian clock proteins regulate neuronal redox homeostasis and neurodegeneration**. *J Clin Invest* 2013, **123**:5389–5400.

8. Hastings MH, Goedert M: **Circadian clocks and neurodegenerative diseases: time to aggregate?** *Curr Opin Neurobiol* 2013, **23**:880–887.

9. DeBruyne JP, Weaver DR, Reppert SM: **CLOCK and NPAS2 have overlapping roles in the suprachiasmatic circadian clock**. *Nat Neurosci* 2007, **10**:543–545.

10. Debruyne JP: **Oscillating perceptions: the ups and downs of the CLOCK protein in the mouse circadian system**. *J Genet* 2008, **87**:437–446.

11. Etchegaray J-P, Lee C, Wade PA, Reppert SM: **Rhythmic histone acetylation underlies transcription in the mammalian circadian clock**. *Nature* 2003, **421**:177–182.

12. Yin L, Lazar MA: **The orphan nuclear receptor Rev-erbalpha recruits the N-CoR/histone deacetylase 3 corepressor to regulate the circadian Bmal1 gene**. *Mol Endocrinol Baltim Md* 2005, **19**:1452–1459.

13. Guillaumond F, Dardente H, Giguère V, Cermakian N: **Differential control of Bmal1 circadian transcription by REV-ERB and ROR nuclear receptors**. *J Biol Rhythms* 2005, **20**:391–403.

14. Sato TK, Panda S, Miraglia LJ, Reyes TM, Rudic RD, McNamara P, Naik KA, FitzGerald GA, Kay SA, Hogenesch JB: **A Functional Genomics Strategy Reveals Rora as a Component of the Mammalian Circadian Clock**. *Neuron* 2004, **43**:527–537.

15. Lowrey PL, Takahashi JS: **Genetics of circadian rhythms in Mammalian model organisms**. *Adv Genet* 2011, **74**:175–230.

16. Ko CH, Takahashi JS: **Molecular components of the mammalian circadian clock**. *Hum Mol Genet* 2006, **15 Spec No 2**:R271–277.

17. Welsh DK, Logothetis DE, Meister M, Reppert SM: **Individual neurons dissociated from rat suprachiasmatic nucleus express independently phased circadian firing rhythms**. *Neuron* 1995, **14**:697–706.

18. Balsalobre A, Damiola F, Schibler U: **A serum shock induces circadian gene expression in mammalian tissue culture cells**. *Cell* 1998, **93**:929–937.

19. Welsh DK, Yoo S-H, Liu AC, Takahashi JS, Kay SA: **Bioluminescence imaging of individual fibroblasts reveals persistent, independently phased circadian rhythms of clock gene expression**. *Curr Biol CB* 2004, **14**:2289–2295.

20. Weaver DR: **The suprachiasmatic nucleus: a 25-year retrospective**. *J Biol Rhythms* 1998, **13**:100–112.

21. Welsh DK, Takahashi JS, Kay SA: **Suprachiasmatic nucleus: cell autonomy and network properties**. *Annu Rev Physiol* 2010, **72**:551–577.

22. Berson DM, Dunn FA, Takao M: **Phototransduction by retinal ganglion cells that set the circadian clock**. *Science* 2002, **295**:1070–1073.

23. Chen S-K, Badea TC, Hattar S: **Photoentrainment and pupillary light reflex are mediated by distinct populations of ipRGCs**. *Nature* 2011, **476**:92–95.

24. Reppert SM, Weaver DR: **Coordination of circadian timing in mammals**. *Nature* 2002, **418**:935–941.

25. Yoo S-H, Yamazaki S, Lowrey PL, Shimomura K, Ko CH, Buhr ED, Siepka SM, Hong H-K, Oh WJ, Yoo OJ, Menaker M, Takahashi JS: **PERIOD2::LUCIFERASE real-time reporting of circadian dynamics reveals persistent circadian oscillations in mouse peripheral tissues**. *Proc Natl Acad Sci U S A* 2004, **101**:5339–5346.

26. Buhr ED, Yoo S-H, Takahashi JS: **Temperature as a universal resetting cue for mammalian circadian oscillators**. *Science* 2010, **330**:379–385.

27. Ueda HR, Hayashi S, Chen W, Sano M, Machida M, Shigeyoshi Y, Iino M, Hashimoto S: **System-level identification of transcriptional circuits underlying mammalian circadian clocks**. *Nat Genet* 2005, **37**:187–192.

28. Gachon F, Olela FF, Schaad O, Descombes P, Schibler U: **The circadian PAR-domain basic leucine zipper transcription factors DBP, TEF, and HLF modulate basal and inducible xenobiotic detoxification**. *Cell Metab* 2006, **4**:25–36.

29. Bozek K, Relógio A, Kielbasa SM, Heine M, Dame C, Kramer A, Herzel H: **Regulation of Clock-Controlled Genes in Mammals**. *PLoS ONE* 2009, **4**:e4882.

30. Schrem H, Klempnauer J, Borlak J: **Liver-enriched transcription factors in liver function and development. Part II: the C/EBPs and D site-binding protein in cell cycle control, carcinogenesis, circadian gene regulation, liver regeneration, apoptosis, and liver-specific gene regulation**. *Pharmacol Rev* 2004, **56**:291–330.

31. Lavery DJ, Schibler U: **Circadian transcription of the cholesterol 7 alpha hydroxylase gene may involve the liver-enriched bZIP protein DBP**. *Genes Dev* 1993, **7**:1871–1884.

32. McClung CA, Sidiropoulou K, Vitaterna M, Takahashi JS, White FJ, Cooper DC, Nestler EJ: **Regulation of dopaminergic transmission and cocaine reward by the Clock gene**. *Proc Natl Acad Sci U S A* 2005, **102**:9377–9381.

33. Hampp G, Ripperger JA, Houben T, Schmutz I, Blex C, Perreau-Lenz S, Brunk I, Spanagel R, Ahnert-Hilger G, Meijer JH, Albrecht U: **Regulation of monoamine oxidase A by circadian-clock components implies clock influence on mood**. *Curr Biol CB* 2008, **18**:678–683.

34. Panda S, Antoch MP, Miller BH, Su AI, Schook AB, Straume M, Schultz PG, Kay SA, Takahashi JS, Hogenesch JB: **Coordinated transcription of key pathways in the mouse by the circadian clock**. *Cell* 2002, **109**:307–320.

35. Le Martelot G, Claudel T, Gatfield D, Schaad O, Kornmann B, Sasso GL, Moschetta A, Schibler U: **REV-ERBalpha participates in circadian SREBP signaling and bile acid homeostasis**. *PLoS Biol* 2009, **7**:e1000181.

36. Storch K-F, Lipan O, Leykin I, Viswanathan N, Davis FC, Wong WH, Weitz CJ: **Extensive and divergent circadian gene expression in liver and heart**. *Nature* 2002, **417**:78–83.

37. Kornmann B, Schaad O, Bujard H, Takahashi JS, Schibler U: **System-driven and oscillator-dependent circadian transcription in mice with a conditionally active liver clock**. *PLoS Biol* 2007, **5**:e34.

38. Hughes ME, Hong H-K, Chong JL, Indacochea AA, Lee SS, Han M, Takahashi JS, Hogenesch JB: **Brain-specific rescue of clock reveals system-driven transcriptional rhythms in peripheral tissue**. *PLoS Genet* 2012, **8**:e1002835.

39. Hughes ME, Deharo L, Pulivarthy SR, Gu J, Hayes KR, Panda S, Hogenesch JB: **High-resolution time course analysis of gene expression from pituitary**. *Cold Spring Harb Symp Quant Biol* 2007, **72**:381–386.

40. Hughes ME, DiTacchio L, Hayes KR, Vollmers C, S Pulivarthy, Baggs JE, Panda S, Hogenesch JB: **Harmonics of circadian gene transcription in mammals**. *PLoS Genet* 2009, **5**:e1000442.

41. McGlincy NJ, Valomon A, Chesham JE, Maywood ES, Hastings MH, Ule J: **Regulation of alternative splicing by the circadian clock and food related cues**. *Genome Biol* 2012, **13**:R54.

42. Hatanaka F, Matsubara C, Myung J, Yoritaka T, Kamimura N, Tsutsumi S, Kanai A, Suzuki Y, Sassone-Corsi P, Aburatani H, Sugano S, Takumi T: **Genome-wide profiling of the core clock protein BMAL1 targets reveals strict relationship with metabolism**. *Mol Cell Biol* 2010.

43. Rey G, Cesbron F, Rougemont J, Reinke H, Brunner M, Naef F: **Genome-Wide and Phase-Specific DNA-Binding Rhythms of BMAL1 Control Circadian Output Functions in Mouse Liver**. *PLoS Biol* 2011, **9**:e1000595.

44. Koike N, Yoo S-H, Huang H-C, Kumar V, Lee C, Kim T-K, Takahashi JS: **Transcriptional Architecture and Chromatin Landscape of the Core Circadian Clock in Mammals**. *Science* 2012, **338**:349–354.

45. Feng D, Liu T, Sun Z, Bugge A, Mullican SE, Alenghat T, Liu XS, Lazar MA: **A circadian rhythm orchestrated by histone deacetylase 3 controls hepatic lipid metabolism**. *Science* 2011, **331**:1315–1319.

46. Cho H, Zhao X, Hatori M, Yu RT, Barish GD, Lam MT, Chong L-W, DiTacchio L, Atkins AR, Glass CK, Liddle C, Auwerx J, Downes M, Panda S, Evans RM: **Regulation of circadian behaviour and metabolism by REV-ERB-α and REV-ERB-β**. *Nature* 2012, **485**:123–127.

47. Bugge A, Feng D, Everett LJ, Briggs ER, Mullican SE, Wang F, Jager J, Lazar MA: **Rev-erbα and Rev-erbβ coordinately protect the circadian clock and normal metabolic function**. *Genes Dev* 2012, **26**:657–667.

48. Hughes ME, Grant GR, Paquin C, Qian J, Nitabach MN: **Deep sequencing the circadian and diurnal transcriptome of Drosophila brain**. *Genome Res* 2012, **22**:1266–1281.

49. Menet JS, Rodriguez J, Abruzzi KC, Rosbash M: **Nascent-Seq reveals novel features of mouse circadian transcriptional regulation**. *eLife* 2012, **1**:e00011.

50. Rodriguez J, Tang C-HA, Khodor YL, Vodala S, Menet JS, Rosbash M: **Nascent-Seq analysis of Drosophila cycling gene expression**. *Proc Natl Acad Sci U S A* 2013, **110**:E275–284.

51. Vollmers C, Schmitz RJ, Nathanson J, Yeo G, Ecker JR, Panda S: **Circadian oscillations of protein-coding and regulatory RNAs in a highly dynamic mammalian liver epigenome**. *Cell Metab* 2012, **16**:833–845.

52. Kadener S, Menet JS, Sugino K, Horwich MD, Weissbein U, Nawathean P, Vagin VV, Zamore PD, Nelson SB, Rosbash M: **A role for microRNAs in the Drosophila circadian clock**. *Genes Dev* 2009, **23**:2179–2191.

53. Coon SL, Munson PJ, Cherukuri PF, Sugden D, Rath MF, Møller M, Clokie SJH, Fu C, Olanich ME, Rangel Z, Werner T, Mullikin JC, Klein DC: **Circadian changes in long noncoding RNAs in the pineal gland**. *Proc Natl Acad Sci U S A* 2012, **109**:13319–13324.

54. Le Martelot G, Canella D, Symul L, Migliavacca E, Gilardi F, Liechti R, Martin O, Harshman K, Delorenzi M, Desvergne B, Herr W, Deplancke B, Schibler U, Rougemont J, Guex N, Hernandez N, Naef F, the CycliX consortium: **Genome-Wide RNA Polymerase II Profiles and RNA Accumulation Reveal Kinetics of Transcription and Associated Epigenetic Changes During Diurnal Cycles**. *PLoS Biol* 2012, **10**:e1001442.

55. Deckard A, Anafi RC, Hogenesch JB, Haase SB, Harer J: **Design and analysis of large-scale biological rhythm studies: a comparison of algorithms for detecting periodic signals in biological data**. *Bioinforma Oxf Engl* 2013, **29**:3174–3180.

56. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR, Boutell JM, Bryant J, Carter RJ, Keira Cheetham R, Cox AJ, Ellis DJ, Flatbush MR, Gormley NA, Humphray SJ, Irving LJ, Karbelashvili MS, Kirk SM, Li H, Liu X, Maisinger KS, Murray LJ, Obradovic B, Ost T, Parkinson

ML, Pratt MR, et al.: **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature* 2008, **456**:53–59.

57. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B: **Mapping and quantifying mammalian transcriptomes by RNA-Seq**. *Nat Methods* 2008, **5**:621–628.

58. Wang Z, Gerstein M, Snyder M: **RNA-Seq: a revolutionary tool for transcriptomics**. *Nat Rev Genet* 2009, **10**:57–63.

59. Wilhelm BT, Landry J-R: **RNA-Seq-quantitative measurement of expression through massively parallel RNA-sequencing**. *Methods San Diego Calif* 2009.

60. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome**. *Genome Biol* 2009, **10**:R25.

61. Grant GR, Farkas MH, Pizarro AD, Lahens NF, Schug J, Brunk BP, Stoeckert CJ, Hogenesch JB, Pierce EA: **Comparative analysis of RNA-Seq alignment algorithms and the RNA-Seq unified mapper (RUM)**. *Bioinforma Oxf Engl* 2011, **27**:2518–2528.

62. Wu TD, Nacu S: **Fast and SNP-tolerant detection of complex variants and splicing in short reads**. *Bioinforma Oxf Engl* 2010, **26**:873–881.

63. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR: **STAR: ultrafast universal RNA-seq aligner**. *Bioinforma Oxf Engl* 2013, **29**:15–21.

64. Engström PG, Steijger T, Sipos B, Grant GR, Kahles A, RGASP Consortium, Rätsch G, Goldman N, Hubbard TJ, Harrow J, Guigó R, Bertone P: **Systematic evaluation of spliced alignment programs for RNA-seq data**. *Nat Methods* 2013, **10**:1185–1191.

65. Marioni JC, Mason CE, Mane SM, Stephens M, Gilad Y: **RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays**. *Genome Res* 2008, **18**:1509–1517.

66. Fu X, Fu N, Guo S, Yan Z, Xu Y, Hu H, Menzel C, Chen W, Li Y, Zeng R, Khaitovich P: **Estimating accuracy of RNA-Seq and microarrays with proteomics**. *BMC Genomics* 2009, **10**:161.

67. Piskol R, Ramaswami G, Li JB: **Reliable identification of genomic variants from RNA-seq data**. *Am J Hum Genet* 2013, **93**:641–651.

68. Trapnell C, Pachter L, Salzberg SL: **TopHat: discovering splice junctions with RNA-Seq**. *Bioinforma Oxf Engl* 2009, **25**:1105–1111.

69. Au KF, Jiang H, Lin L, Xing Y, Wong WH: **Detection of splice junctions from paired-end RNA-seq data by SpliceMap**. *Nucleic Acids Res* 2010.

70. Farkas MH, Grant GR, White JA, Sousa ME, Consugar MB, Pierce EA: **Transcriptome analyses of the human retina identify unprecedented transcript diversity and 3.5 Mb of novel transcribed sequence via significant alternative splicing and novel genes**. *BMC Genomics* 2013, **14**:486.

71. Benjamini Y, Speed TP: **Summarizing and correcting the GC content bias in high-throughput sequencing**. *Nucleic Acids Res* 2012, **40**:e72.

72. Aird D, Ross MG, Chen W-S, Danielsson M, Fennell T, Russ C, Jaffe DB, Nusbaum C, Gnirke A: **Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries**. *Genome Biol* 2011, **12**:R18.

73. Hansen KD, Brenner SE, Dudoit S: **Biases in Illumina transcriptome sequencing caused by random hexamer priming**. *Nucleic Acids Res* 2010, **38**:e131–e131.

74. Nakamura K, Oshima T, Morimoto T, Ikeda S, Yoshikawa H, Shiwa Y, Ishikawa S, Linak MC, Hirai A, Takahashi H, Altaf-Ul-Amin M, Ogasawara N, Kanaya S: **Sequence-specific error profile of Illumina sequencers**. *Nucleic Acids Res* 2011, **39**:e90.

75. Jayaprakash AD, Jabado O, Brown BD, Sachidanandam R: **Identification and remediation of biases in the activity of RNA ligases in small-RNA deep sequencing**. *Nucleic Acids Res* 2011, **39**:e141.

76. Malone JH, Oliver B: **Microarrays, deep sequencing and the true measure of the transcriptome**. *BMC Biol* 2011, **9**:34.

77. Steijger T, Abril JF, Engström PG, Kokocinski F, RGASP Consortium, Abril JF, Akerman M, Alioto T, Ambrosini G, Antonarakis SE, Behr J, Bertone P: **Assessment of transcript reconstruction methods for RNA-seq**. *Nat Methods* 2013, **10**:1177–1184.

78. Bullard JH, Purdom E, Hansen KD, Dudoit S: **Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments**. *BMC Bioinformatics* 2010, **11**:94.

79. MAQC Consortium, Shi L, Reid LH, Jones WD, Shippy R, Warrington JA, Baker SC, Collins PJ, de Longueville F, Kawasaki ES, Lee KY, Luo Y, Sun YA, Willey JC, Setterquist RA, Fischer GM, Tong W, Dragan YP, Dix DJ, Frueh FW, Goodsaid FM, Herman D, Jensen RV, Johnson CD, Lobenhofer EK, Puri RK, Schrf U, Thierry-Mieg J, Wang C, Wilson M, et al.: **The MicroArray Quality Control (MAQC) project shows**

inter- and intraplatform reproducibility of gene expression measurements. *Nat Biotechnol* 2006, **24**:1151–1161.

80. Reimers M: **Making Informed Choices about Microarray Data Analysis**. *PLoS Comput Biol* 2010, **6**:e1000786.

81. Soshnev AA, Ishimoto H, McAllister BF, Li X, Wehling MD, Kitamoto T, Geyer PK: **A conserved long noncoding RNA affects sleep behavior in Drosophila**. *Genetics* 2011, **189**:455–468.

82. Lowrey PL, Takahashi JS: **Mammalian circadian biology: elucidating genome-wide levels of temporal organization**. *Annu Rev Genomics Hum Genet* 2004, **5**:407–441.

83. Bunger MK, Wilsbacher LD, Moran SM, Clendenin C, Radcliffe LA, Hogenesch JB, Simon MC, Takahashi JS, Bradfield CA: **Mop3 is an essential component of the master circadian pacemaker in mammals**. *Cell* 2000, **103**:1009–1017.

84. Baggs JE, Hogenesch JB: **Genomics and systems approaches in the mammalian circadian clock**. *Curr Opin Genet Dev* 2010, **20**:581–587.

85. Allison DB, Cui X, Page GP, Sabripour M: **Microarray data analysis: from disarray to consolidation and consensus**. *Nat Rev Genet* 2006, **7**:55–65.

86. Filichkin SA, Priest HD, Givan SA, Shen R, Bryant DW, Fox SE, Wong W-K, Mockler TC: **Genome-wide mapping of alternative splicing in Arabidopsis thaliana**. *Genome Res* 2010, **20**:45–58.

87. Pepke S, Wold B, Mortazavi A: **Computation for ChIP-seq and RNA-seq studies**. *Nat Methods* 2009, **6**(11 Suppl):S22–32.

88. Auer PL, Doerge RW: **Statistical design and analysis of RNA sequencing data**. *Genetics* 2010, **185**.

89. Vitaterna MH, King DP, Chang AM, Kornhauser JM, Lowrey PL, McDonald JD, Dove WF, Pinto LH, Turek FW, Takahashi JS: **Mutagenesis and mapping of a mouse gene, Clock, essential for circadian behavior**. *Science* 1994, **264**:719–725.

90. King DP, Zhao Y, Sangoram AM, Wilsbacher LD, Tanaka M, Antoch MP, Steeves TD, Vitaterna MH, Kornhauser JM, Lowrey PL, Turek FW, Takahashi JS: **Positional cloning of the mouse circadian clock gene**. *Cell* 1997, **89**:641–653.

91. Hughes ME, Hogenesch JB, Kornacker K: **JTK_CYCLE: an efficient nonparametric algorithm for detecting rhythmic components in genome-scale data sets**. *J Biol Rhythms* 2010, **25**:372–380.

92. Ripperger JA, Schibler U: **Rhythmic CLOCK-BMAL1 binding to multiple E-box motifs drives circadian Dbp transcription and chromatin transitions**. *Nat Genet* 2006, **38**:369–374.

93. Godoy J, Nishimura M, Webster NJG: **Gonadotropin-releasing hormone induces miR-132 and miR-212 to regulate cellular morphology and migration in immortalized LbetaT2 pituitary gonadotrope cells**. *Mol Endocrinol Baltim Md* 2011, **25**:810–820.

94. Heintzman ND, Stuart RK, Hon G, Fu Y, Ching CW, Hawkins RD, Barrera LO, Van Calcar S, Qu C, Ching KA, Wang W, Weng Z, Green RD, Crawford GE, Ren B: **Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome**. *Nat Genet* 2007, **39**:311–318.

95. Guenther MG, Levine SS, Boyer LA, Jaenisch R, Young RA: **A chromatin landmark and transcription initiation at most promoters in human cells**. *Cell* 2007, **130**:77–88.

96. Heintzman ND, Hon GC, Hawkins RD, Kheradpour P, Stark A, Harp LF, Ye Z, Lee LK, Stuart RK, Ching CW, Ching KA, Antosiewicz-Bourget JE, Liu H, Zhang X, Green RD, Lobanenkov VV, Stewart R, Thomson JA, Crawford GE, Kellis M, Ren B: **Histone modifications at human enhancers reflect global cell-type-specific gene expression**. *Nature* 2009, **459**:108–112.

97. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data**. *Nucleic Acids Res* 2010, **39**(Database):D152–D157.

98. Chiang HR, Schoenfeld LW, Ruby JG, Auyeung VC, Spies N, Baek D, Johnston WK, Russ C, Luo S, Babiarz JE, Blelloch R, Schroth GP, Nusbaum C, Bartel DP: **Mammalian microRNAs: experimental evaluation of novel and previously annotated genes**. *Genes Dev* 2010, **24**:992–1009.

99. Pizarro A, Hayer K, Lahens NF, Hogenesch JB: **CircaDB: a database of mammalian circadian gene expression profiles**. *Nucleic Acids Res* 2012.

100. Filichkin SA, Mockler TC: **Unproductive alternative splicing and nonsense mRNAs: A widespread phenomenon among plant circadian clock genes**. *Biol Direct* 2012, **7**:20.

101. Amaral PP, Clark MB, Gascoigne DK, Dinger ME, Mattick JS: **lncRNAdb: a reference database for long noncoding RNAs**. *Nucleic Acids Res* 2011, **39**(Database issue):D146–151.

102. Moran VA, Perera RJ, Khalil AM: **Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs**. *Nucleic Acids Res* 2012, **40**:6391–6400.

103. Spizzo R, Almeida MI, Colombatti A, Calin GA: **Long non-coding RNAs and cancer: a new frontier of translational research?** *Oncogene* 2012, **31**:4577–4587.

104. Yang L, Duff MO, Graveley BR, Carmichael GG, Chen L-L: **Genomewide characterization of non-polyadenylated RNAs**. *Genome Biol* 2011, **12**:R16.

105. Zheng GXY, Ravi A, Calabrese JM, Medeiros LA, Kirak O, Dennis LM, Jaenisch R, Burge CB, Sharp PA: **A latent pro-survival function for the mir-290-295 cluster in mouse embryonic stem cells**. *PLoS Genet* 2011, **7**:e1002054.

106. Gatfield D, Le Martelot G, Vejnar CE, Gerlach D, Schaad O, Fleury-Olela F, Ruskeepää A-L, Oresic M, Esau CC, Zdobnov EM, Schibler U: **Integration of microRNA miR-122 in hepatic circadian gene expression**. *Genes Dev* 2009, **23**:1313–1326.

107. Cheng H-YM, Papp JW, Varlamova O, Dziema H, Russell B, Curfman JP, Nakazawa T, Shimizu K, Okamura H, Impey S, Obrietan K: **microRNA modulation of circadian-clock period and entrainment**. *Neuron* 2007, **54**:813–829.

108. Hsu F, Kent WJ, Clawson H, Kuhn RM, Diekhans M, Haussler D: **The UCSC Known Genes**. *Bioinformatics* 2006, **22**:1036–1046.

109. Pruitt KD, Tatusova T, Klimke W, Maglott DR: **NCBI Reference Sequences: current status, policy and new initiatives**. *Nucleic Acids Res* 2009, **37**(Database issue):D32–36.

110. Wilming LG, Gilbert JGR, Howe K, Trevanion S, Hubbard T, Harrow JL: **The vertebrate genome annotation (Vega) database**. *Nucleic Acids Res* 2008, **36**(Database issue):D753–760.

111. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Wheeler DL: **GenBank: update**. *Nucleic Acids Res* 2004, **32**(Database issue):D23–26.

112. Parra G, Agarwal P, Abril JF, Wiehe T, Fickett JW, Guigó R: **Comparative gene prediction in human and mouse**. *Genome Res* 2003, **13**:108–117.

113. Gross SS, Brent MR: **Using multiple alignments to improve gene prediction**. *J Comput Biol J Comput Mol Cell Biol* 2006, **13**:379–393.

114. Burge C, Karlin S: **Prediction of complete gene structures in human genomic DNA**. *J Mol Biol* 1997, **268**:78–94.

115. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kähäri AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, et al.: **Ensembl 2012**. *Nucleic Acids Res* 2012, **40**(Database issue):D84–90.

116. Thierry-Mieg D, Thierry-Mieg J: **AceView: a comprehensive cDNA-supported gene and transcripts annotation**. *Genome Biol* 2006, **7 Suppl 1**:S12.1–14.

117. Ye J, Coulouris G, Zaretskaya I, Cutcutache I, Rozen S, Madden T: **Primer-BLAST: A tool to design target-specific primers for polymerase chain reaction**. *BMC Bioinformatics* 2012, **13**:134.

118. Dalchau N, Webb AA: **Ticking over - Circadian systems across the kingdoms of life**. *Biochemist* 2011, **33**.

119. Levi F, Schibler U: **Circadian rhythms: mechanisms and therapeutic implications**. *Annu Rev Pharmacol Toxicol* 2007, **47**:593–628.

120. Klerman EB: **Clinical aspects of human circadian rhythms.** *J Biol Rhythms* 2005, **20**:375–86.

121. Akhtar RA, Reddy AB, Maywood ES, Clayton JD, King VM, Smith AG, Gant TW, Hastings MH, Kyriacou CP: **Circadian cycling of the mouse liver transcriptome, as revealed by cDNA microarray, is driven by the suprachiasmatic nucleus.** *Curr Biol CB* 2002, **12**:540–50.

122. Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, Regev A, Rinn JL: **Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses**. *Genes Dev* 2011.

123. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, Lagarde J, Veeravalli L, Ruan X, Ruan Y, Lassmann T, Carninci P, Brown JB, Lipovich L, Gonzalez JM, Thomas M, Davis CA, Shiekhattar R, Gingeras TR, Hubbard TJ, Notredame C, Harrow J, Guigó R: **The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression**. *Genome Res* 2012, **22**:1775–1789.

124. Washietl S, Kellis M, Garber M: **Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals**. *Genome Res* 2014.

125. Pelechano V, Steinmetz LM: **Gene regulation by antisense transcription**. *Nat Rev Genet* 2013, **14**:880–893.

126. Van Devondervoort IIGM, Gordebeke PM, Khoshab N, Tiesinga PHE, Buitelaar JK, Kozicz T, Aschrafi A, Glennon JC: **Long non-coding RNAs in neurodevelopmental disorders**. *Front Mol Neurosci* 2013, **6**:53.

127. Lim C, Allada R: **Emerging roles for post-transcriptional regulation in circadian clocks**. *Nat Neurosci* 2013, **16**:1544–1550.

128. Ling H, Fabbri M, Calin GA: **MicroRNAs and other non-coding RNAs as targets for anticancer drug development**. *Nat Rev Drug Discov* 2013, **12**:847–865.

129. Hauptman N, Glavac D: **MicroRNAs and long non-coding RNAs: prospects in diagnostics and therapy of cancer**. *Radiol Oncol* 2013, **47**:311–318.

130. Bu D, Yu K, Sun S, Xie C, Skogerbø G, Miao R, Xiao H, Liao Q, Luo H, Zhao G, Zhao H, Liu Z, Liu C, Chen R, Zhao Y: **NONCODE v3.0: integrative annotation of long noncoding RNAs**. *Nucleic Acids Res* 2012, **40**(Database issue):D210–215.

131. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**:403–410.

132. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, Zadissa A, Searle S, Barnes I, Bignell A, Boychenko V, Hunt T, Kay M, Mukherjee G, Rajan J, Despacio-Reyes G, Saunders G, Steward C, Harte R, Lin M, Howald C, Tanzer A, Derrien T, Chrast J, Walters N, Balasubramanian S, Pei B, Tress M, et al.: **GENCODE: the reference human genome annotation for The ENCODE Project**. *Genome Res* 2012, **22**:1760–1774.

133. Lewis BP, Burge CB, Bartel DP: **Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets**. *Cell* 2005, **120**:15–20.

134. Friedman RC, Farh KK-H, Burge CB, Bartel DP: **Most mammalian mRNAs are conserved targets of microRNAs**. *Genome Res* 2009, **19**:92–105.

135. Antithrombotic Trialists' Collaboration: **Collaborative meta-analysis of randomised trials of antiplatelet therapy for prevention of death, myocardial infarction, and stroke in high risk patients**. *BMJ* 2002, **324**:71–86.

136. Cohen MC, Rohtla KM, Lavery CE, Muller JE, Mittleman MA: **Meta-analysis of the morning excess of acute myocardial infarction and sudden cardiac death**. *Am J Cardiol* 1997, **79**:1512–1516.

137. Gurha P, Abreu-Goodger C, Wang T, Ramirez MO, Drumond AL, van Dongen S, Chen Y, Bartonicek N, Enright AJ, Lee B, Kelm RJ, Reddy AK, Taffet GE, Bradley A, Wehrens XH, Entman ML, Rodriguez A: **Targeted deletion of microRNA-22 promotes stress-induced cardiac dilation and contractile dysfunction**. *Circulation* 2012, **125**:2751–2761.

138. Patrignani P, Tacconelli S, Piazuelo E, Di Francesco L, Dovizio M, Sostres C, Marcantoni E, Guillem-Llobat P, Del Boccio P, Zucchelli M, Patrono C, Lanas A: **Reappraisal of the clinical pharmacology of low-dose aspirin by comparing novel direct and traditional indirect biomarkers of drug action**. *J Thromb Haemost JTH* 2014.

139. Anders S, Huber W: **Differential expression analysis for sequence count data**. *Genome Biol* 2010, **11**:R106.

140. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features**. *Bioinforma Oxf Engl* 2010, **26**:841–842.

141. Meyer LR, Zweig AS, Hinrichs AS, Karolchik D, Kuhn RM, Wong M, Sloan CA, Rosenbloom KR, Roe G, Rhead B, Raney BJ, Pohl A, Malladi VS, Li CH, Lee BT, Learned K, Kirkup V, Hsu F, Heitner S, Harte RA, Haeussler M, Guruvadoo L, Goldman M, Giardine BM, Fujita PA, Dreszer TR, Diekhans M, Cline MS, Clawson H, Barber GP, et al.: **The UCSC Genome Browser database: extensions and updates 2013**. *Nucleic Acids Res* 2013, **41**(Database issue):D64–69.

142. Park E, Williams B, Wold BJ, Mortazavi A: **RNA editing in the human ENCODE RNA-seq data**. *Genome Res* 2012, **22**:1626–1633.

143. Ilott NE, Ponting CP: **Predicting long non-coding RNAs using RNA sequencing**. *Methods San Diego Calif* 2013, **63**:50–59.

144. Nagalakshmi U, Waern K, Snyder M: **RNA-Seq: a method for comprehensive transcriptome analysis**. *Curr Protoc Mol Biol Ed Frederick M Ausubel Al* 2010, **Chapter 4**:Unit 4.11.1–13.

145. Levin JZ, Yassour M, Adiconis X, Nusbaum C, Thompson DA, Friedman N, Gnirke A, Regev A: **Comprehensive comparative analysis of strand-specific RNA sequencing methods**. *Nat Methods* 2010, **7**:709–715.

146. Cui P, Lin Q, Ding F, Xin C, Gong W, Zhang L, Geng J, Zhang B, Yu X, Yang J, Hu S, Yu J: **A comparison between ribo-minus RNA-sequencing and polyA-selected RNA-sequencing**. *Genomics* 2010, **96**:259–265.

147. Adiconis X, Borges-Rivera D, Satija R, DeLuca DS, Busby MA, Berlin AM, Sivachenko A, Thompson DA, Wysoker A, Fennell T, Gnirke A, Pochet N, Regev A, Levin JZ: **Comparative analysis of RNA sequencing methods for degraded or low-input samples**. *Nat Methods* 2013, **10**:623–629.

148. Spicuglia S, Maqbool MA, Puthier D, Andrau J-C: **An update on recent methods applied for deciphering the diversity of the noncoding RNA genome structure and function**. *Methods San Diego Calif* 2013, **63**:3–17.

149. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, Pachter L: **Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation**. *Nat Biotechnol* 2010, **28**:511–515.

150. Lee S, Seo CH, Lim B, Yang JO, Oh J, Kim M, Lee S, Lee B, Kang C, Lee S: **Accurate quantification of transcriptome from RNA-Seq data by effective length normalization**. *Nucleic Acids Res* 2011, **39**:e9–e9.

151. Jiang L, Schlesinger F, Davis CA, Zhang Y, Li R, Salit M, Gingeras TR, Oliver B: **Synthetic spike-in standards for RNA-seq experiments**. *Genome Res* 2011, **21**:1543–1551.

152. Temple G, Gerhard DS, Rasooly R, Feingold EA, Good PJ, Robinson C, Mandich A, Derge JG, Lewis J, Shoaf D, Collins FS, Jang W, Wagner L, Shenmen CM, Misquitta L, Schaefer CF, Buetow KH, Bonner TI, Yankie L, Ward M, Phan L, Astashyn A, Brown G, Farrell C, Hart J, Landrum M, Maidak BL, Murphy M, Murphy T, Rajput B, et al.: **The completion of the Mammalian Gene Collection (MGC)**. *Genome Res* 2009, **19**:2324–2333.

153. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS: **MEME SUITE: tools for motif discovery and searching**. *Nucleic Acids Res* 2009, **37**(Web Server issue):W202–208.

154. Schmitt AO, Herzel H: **Estimating the entropy of DNA sequences**. *J Theor Biol* 1997, **188**:369–377.

155. Mantegna RN, Buldyrev SV, Goldberger AL, Havlin S, Peng CK, Simons M, Stanley HE: **Linguistic features of noncoding DNA sequences**. *Phys Rev Lett* 1994, **73**:3169–3172.

156. Koonin EV: **A non-adaptationist perspective on evolution of genomic complexity or the continued dethroning of man**. *Cell Cycle Georget Tex* 2004, **3**:280–285.

157. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing**. *Nucleic Acids Res* 2008, **36**:e105.

158. Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ: **The UCSC Genome Browser database: update 2010**. *Nucleic Acids Res* 2010, **38**(Database issue):D613–619.

159. Velu CS, Grimes HL: **Utilizing antagomiR (antisense microRNA) to knock down microRNA in murine bone marrow cells**. *Methods Mol Biol Clifton NJ* 2012, **928**:185–195.

160. Nicolas FE: **Experimental validation of microRNA targets using a luciferase reporter system**. *Methods Mol Biol Clifton NJ* 2011, **732**:139–152.

161. Nagoshi E, Saini C, Bauer C, Laroche T, Naef F, Schibler U: **Circadian gene expression in individual fibroblasts: cell-autonomous and self-sustained oscillators pass time to daughter cells**. *Cell* 2004, **119**:693–705.

162. Baggs JE, Price TS, DiTacchio L, Panda S, Fitzgerald GA, Hogenesch JB: **Network features of the mammalian circadian clock**. *PLoS Biol* 2009, **7**:e52.

163. Bélanger V, Picard N, Cermakian N: **The circadian regulation of Presenilin-2 gene expression**. *Chronobiol Int* 2006, **23**:747–766.

164. Olarerin-George AO, Anton L, Hwang Y-C, Elovitz MA, Hogenesch JB: **A functional genomics screen for microRNA regulators of NF-kappaB signaling**. *BMC Biol* 2013, **11**:19.

165. Tovin A, Alon S, Ben-Moshe Z, Mracek P, Vatine G, Foulkes NS, Jacob-Hirsch J, Rechavi G, Toyama R, Coon SL, Klein DC, Eisenberg E, Gothilf Y: **Systematic identification of rhythmic genes reveals camk1gb as a new element in the circadian clockwork**. *PLoS Genet* 2012, **8**:e1003116.

166. Yagita K, Horie K, Koinuma S, Nakamura W, Yamanaka I, Urasaki A, Shigeyoshi Y, Kawakami K, Shimada S, Takeda J, Uchiyama Y: **Development of the circadian oscillator during differentiation of mouse embryonic stem cells in vitro**. *Proc Natl Acad Sci U S A* 2010, **107**:3846–3851.

167. Hüttenhofer A, Kiefmann M, Meier-Ewert S, O'Brien J, Lehrach H, Bachellerie JP, Brosius J: **RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse**. *EMBO J* 2001, **20**:2943–2953.

168. Kiss AM, Jády BE, Bertrand E, Kiss T: **Human box H/ACA pseudouridylation guide RNA machinery**. *Mol Cell Biol* 2004, **24**:5797–5807.

169. Schattner P, Barberan-Soler S, Lowe TM: **A computational screen for mammalian pseudouridylation guide H/ACA RNAs**. *RNA N Y N* 2006, **12**:15–25.

170. Jouffe C, Cretenet G, Symul L, Martin E, Atger F, Naef F, Gachon F: **The Circadian Clock Coordinates Ribosome Biogenesis**. *PLoS Biol* 2013, **11**:e1001455.

171. Song SJ, Ito K, Ala U, Kats L, Webster K, Sun SM, Jongen-Lavrencic M, Manova-Todorova K, Teruya-Feldstein J, Avigan DE, Delwel R, Pandolfi PP: **The oncogenic microRNA miR-22 targets the TET2 tumor suppressor to promote hematopoietic stem cell self-renewal and transformation**. *Cell Stem Cell* 2013, **13**:87–101.

172. Lin J, Huo R, Xiao L, Zhu X, Xie J, Sun S, He Y, Zhang J, Sun Y, Zhou Z, Wu P, Shen B, Li D, Li N: **A novel p53/microRNA-22/Cyr61 axis in synovial cells regulates inflammation in rheumatoid arthritis**. *Arthritis Rheumatol Hoboken NJ* 2014, **66**:49–59.

173. Huang Z-P, Chen J, Seok HY, Zhang Z, Kataoka M, Hu X, Wang D-Z: **MicroRNA-22 regulates cardiac hypertrophy and remodeling in response to stress**. *Circ Res* 2013, **112**:1234–1243.