Cross-regulation and interaction between eukaryotic gene regulatory processes

Noah Spies
Bachelor of Arts in Mathematics
Cornell University, 2006

Submitted to the Department of Biology in partial fulfillment of the requirements for the degree of **Doctor of Philosophy** at the **Massachusetts Institute of Technology**

June 2012

Noah Spies
Department of Biology
May 1, 2012
Author

David P Bartel
Professor of Biology, MIT
Thesis Advisor

Christopher B Burge
Professor of Biology, MIT
Thesis Advisor

Stephen P Bell
Professor of Biology, MIT
Chair, Biology Graduate Committee

Cross-regulation and interaction between eukaryotic gene regulatory processes

by Noah Spies

Submitted to the Department of Biology on May 1, 2012 in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biology

Thesis Supervisors: David P Bartel and Christopher B Burge, Professors of Biology

Abstract

Regulation of genes is fundamental to all living processes and can be exerted at many sequential steps. We studied several eukaryotic gene regulatory mechanisms with an emphasis on understanding the interplay between regulatory processes on a genome-wide scale.

Gene splicing involves the joining of exonic RNA stretches from within a precursor messenger RNA (mRNA). Splicing typically occurs co-transcriptionally as the pre-mRNA is being produced from the DNA. We explored the relationship between the chromatin state of the gene-encoding DNA and the splicing machinery. We found a marked enrichment for nucleosomes at exonic DNA in human T cells, as compared to surrounding introns, an effect mostly explained by the biased nucleotide content of exons. The use of nucleosome positioning information improved splicing simulation models, suggesting nucleosome positioning may help determine cellular splicing patterns. Additionally, we found several histone marks enriched or depleted at exons compared to the background nucleosome levels, indicative of a histone code for splicing. These results connect the chromatin regulation and mRNA splicing processes in a genome-wide fashion.

Another pre-mRNA processing step is cleavage and polyadenylation, which determines the $3'$ end of the mature mRNA. We found that 3P-Seq was able to quantify the levels of $3'$ end isoforms, in addition to the method's previous use for annotating mRNA $3'$ ends. Using 3P-Seq and a transcriptional shutoff experiment in mouse fibroblasts, we investigated the effect of nuclear alternative $3'$ end formation on mRNA stability, typically regulated in the cytoplasm. In genes with multiple, tandem $3'$ untranslated regions ($3'$ UTRs) produced by alternative cleavage and polyadenylation, we found the shorter UTRs were significantly more stable in general than the longer isoforms. This difference was in part explained by the loss of cis-regulatory motifs, such as microRNA targets and PUF-binding sites, between the proximal and distal isoforms.

Finally, we characterized the small interfering RNAs (siRNAs) produced from heterochromatic, silenced genomic regions in fission yeast. We observed a considerable bias for siRNAs with a $5'$ U, and used this bias to infer patterns of siRNA biogenesis. Furthermore, comparisons with between wild-type and the Cid14 non-canonical poly(A) polymerase mutant demonstrated that the exosome, the nuclear surveillance and processing complex, is required for RNA homeostasis. In the absence of a fully functional exosome complex, siRNAs are produced to normal exosome targets, including ribosomal and transfer RNAs, indicating these processes may compete for substrates and underscoring the interconnectedness of gene regulatory systems.

Acknowledgments

I am indebted to the excellent mentoring I have received both formally and informally in the course of my studies. First and foremost, I could not have completed this work without the direction and guidance of my two PhD advisors, Dave Bartel and Chris Burge. Chris and Dave have always been available to consult about the minutiae as well as the big picture of my work, and I feel I have learned a tremendous amount from my personal interactions with both of them. They have also fostered incredibly supportive lab environments, and I must thank everyone the both labs for scientific and non-scientific input, collaboration and friendship. In particular, I must thank Graham Ruby for early mentoring on the computational side and Calvin Jan for teaching me the ins and outs of wet lab experimentation.

Thanks to Jess Hurt for being an excellent bay mate and also Jess Hurt and Vincent Butty for discussion of the content of this thesis.

I would also like to thank my thesis committee, Phil Sharp and Aviv Regev, not only for the many productive discussions we have had about my projects, but also the encouragement they have provided me through my graduate studies. While I have had a broad network between my two labs and the biology community, it has always been good to know that I have the support of my thesis committee. Thanks also to Angela DePace for joining us for my thesis defense.

Finally, I would like to thank my friends and family. It has been great fun moving to Cambridge and making new friends. It would not have been worth it if I didn't have the love and friendship of Torrey, who has been supportive and interested in my work. And finally, I have to thank my parents, Wendy and Rupert, who have supported me at every step, and instilled a sense of curiosity in me from an early age. Who knew that would lead me down the road of the scientist?

– Noah Spies

Contents

Chapter 1

Introduction

1.1 The importance of gene regulation

The genome is life's blueprint. Contained within is the complete information required to produce all the proteins used to piece together the cell's structures and all the enzymes required to perform nearly all biological processes. The DNA-encoded genomes of thousands of organisms are now completely sequenced, but we are only beginning to understand how the different genetic sequences are combined to produce the complex cellular actions our bodies perform every living minute. Key to our understanding of molecular biology are two things: (1) the knowledge of what each gene in the genome actually does and (2) an understanding of under what conditions those genes are activated or deactivated. This thesis focuses on the latter of these two: gene regulation.

The central dogma of biology provides a simple framework for understanding gene regulation. A gene is first turned on by transcribing its DNA into messenger RNA. The mRNA is then processed and exported from the nucleus to the cytoplasm where it is translated into a protein. The protein folds into a three-dimensional structure capable of performing its cellular role until it is finally degraded. Each of these steps is potentially the target of gene regulatory mechanisms. This chapter describes the mRNA regulatory systems. In the interest of clarity this introduction will focus on gene regulation in mammals but most of these fundamental processes are understood through research performed in more basal organisms such as yeast. Most of the regulatory systems are described here from a mechanistic perspective. However, the biological role of regulation cannot be understated and will be touched upon in specific cases relevant to the research described in subsequent chapters.

An important theme of this introduction is the cross-talk between gene regulatory systems. We will see examples of how transcription can affect splicing through modulating the rate of the transcribing polymerase; how $3'$ end processing factors are recruited at transcription initiation; how splicing places an exon-junction complex on an mRNA, enabling translation-dependent quality control of splicing; and how chromatin-modifying enzymes are recruited to the fission yeast centromeres co-transcriptionally via siRNAs.

This thesis explores several topics related to the interaction between regulatory processes. The second chapter examines the interplay between co-transcriptional gene splicing and the chromatin state of that gene. The third chapter explores how regulation of a nuclear process, cleavage and polyadenylation, creates isoform variants that are differentially regulated by various degradation pathways in the cytoplasm. Finally, the fourth chapter characterizes the fission yeast small RNAs involved in co-transcriptional silencing of centromeres. The particular fission yeast mutants studied demonstrate that the nuclear exosome degradation pathway can compete with RNAi components for substrates.

1.2 Chromatin

Packaging of DNA into chromatin DNA is stored in the nucleus in a compact form we know as chromatin, so-called because of its ability to be stained and viewed under a microscope (Flemming 1882). The protein constituents of chromatin are the histones, which were identified early on but it wasn't until much later that it was understood how chromatin formed around them. Nearly 100 years after the discovery of chromatin, the histone octamer, or nucleosome, was identified as the core repeating protein structure around which DNA was wrapped (Kornberg 1974; Olins and Olins 1974). The nucleosome is made up of four subunits, histones H2A, H2B, H3 and H4. An $(H3–H4)_2$ tetramer forms the center of the nucleosome, with two H2A–H2B dimers bound to either side (Luger et al. 1997). A fifth histone, H1, serves as a linker between nucleosomes. Together, these histones are responsible for packing over a meter worth of DNA, end-to-end, into a nucleus $^1/_{10,000}$ that size (Woodcock and Ghosh 2010).

The nucleosome not only acts to compact DNA but also performs a vital role in regulating the activity of the genes it packages. The histone subunits contain flexible tails that can be modified so as to mark the DNA regions that are wrapped around them, thereby helping recruit regulatory factors (Brownell et al. 1996). In the extreme case, these factors can tightly condense the DNA into what is called heterochromatin, blocking access to RNA transcriptional machinery and leading to nearly complete gene silencing (Trojer and Reinberg 2007).

Post-translational modification of histones Histones in the vicinity of genes are generally post-translationally modified in a manner indicative of the gene's activity level. These modifications typically involve the covalent addition of acetyl, methyl, or ubiquitin groups to lysine and arginine residues on the N-terminal "tail" of each of the histones, although there is a large number of possible modifications (Suganuma and Workman 2011). Promoters of silent genes are enriched for H3K9me3 (tri-methylation on lysine 9 of histone H3) and H3K27me3, whereas active genes typically show enrichment for H3K4me3 and various acetylations (Zhou et al. 2011). The body of transcribed regions is generally high for H3K36me3 and H3K79me2, marks established by the elongating polymerase II complex.

The locations of these modified histones can be assayed using a method called chromatin immunoprecipitation, or ChIP (Solomon et al. 1988). Proteins are first cross-linked to DNA using a chemical, frequently formaldehyde. Following purification of chromatin from the cell, antibodies specific for certain histone tail modifications can be used to immunoprecipitate DNA regions bound by modified histones (or, depending on the choice of antibody, other chromatin-bound factors). The DNA can be digested or fragmented resulting in ~146 bp fragments, the length of DNA bound and protected by a single nucleosome. The cross-links can be reversed, and the bound DNA regions can be interrogated by high-throughput methods such as micro-array (Blat and Kleckner 1999; Ren et al. 2000), known as ChIP-Chip, or high-throughput sequencing, known as ChIP-Seq (Robertson et al. 2007; Johnson et al. 2007; Barski et al. 2007). These methods have immensely increased our understanding of the global distribution of transcription factors and modified histones. However, these results are frequently difficult to translate into understanding of cause and effect because of the global nature of perturbations and the impracticality of genetically modifying histone genes, which are highly duplicated in the genome (Henikoff and Shilatifard 2011).

Nucleosome positioning Active genes are typically characterized by an entirely nucleosome free region surrounding the transcription start site (TSS), allowing easy access to transcription factors and the RNA polymerase II (Yuan et al. 2005). Genes that are active in one cell type lose this nucleosome-free region (perhaps more accurately, but less commonly, known as the nucleosome-depleted region) when silenced in a different tissue, suggesting an active process can regulate the openness of this chromatin stretch (Ozsolak et al. 2007). It remains unclear whether the nucleosome-free region is established prior to transcription, or as a side-effect of recruiting general transcription factors and the polymerase complex to the DNA. Maintenance of the nucleosome-free region at active promoters pushes neighboring nucleosomes into well-defined positions just downstream of the TSS (the +1 nucleosome) and upstream of the promoter (alternatively called the −1 or −2 nucleosome, depending on species).

Nucleosomes preferentially bind to some DNA sequences. This inherent sequence bias can be understood at the structural level: the nucleosome induces significant bending of bound DNA, and this bending is achieved more readily for some sequences (Luger et al. 1997). For example, G·C base pairs are preferred when the major groove faces in toward the nucleosome, and indeed these sequences are preferred every ~10 nucleotides (which coincides with a full twist of DNA wrapped around the nucleosome) (Kaplan et al. 2009). While the nucleosome contacts DNA primarily through the sequence-independent phosphate backbone, some additional amino acid-base contacts also increase nucleosome affinity for A·T base pairs in the minor groove.

These sequence preferences lead to markedly lower affinities for nucleosomes in the nucleosome-free region at the TSS as well as near the transcription termination site (Kaplan et al. 2009). While the nucleosome-free region is regulated between cell types in mammals, in vitro experiments mixing yeast histones and genomic DNA recapitulate a nucleosome-free region, indicating promoter sequences are inherently unfavorable to nucleosome binding (Kaplan et al. 2009). Recent results suggest ATP-dependent chromatin remodelers are required for specific placement of the +1 and subsequent nucleosomes near the 5′ end of the transcript (Zhang et al. 2011). It remains an open question how much DNA sequence affects nucleosome positioning outside of these highly stereotyped regions and in species other than yeast, where most of these studies have been performed (Zhang et al. 2009; Kaplan et al. 2009).

1.3 Transcription

Polymerase II transcription initiation and elongation The DNA-dependent RNA polymerase II (pol II) transcribes protein-coding genes into messenger RNA. Early studies of partially purified polymerases demonstrated separate activities for three individual enzyme complexes, named pol I, pol II and pol III based on the purification scheme used (Roeder and Rutter 1969). While work initially focused on pol III, researchers in the late 1970s and early 1980s purified a set of basal transcription factors which assist in recruiting the 12 subunit pol II to DNA, converting it into an elongation-competent form (Thomas and Chiang 2006).

However, pol II transcription is a far more complex process in the context of a living eukaryote. In vivo transcription begins with the binding of transcription factors to regulatory elements in the core promoter regions near the transcription start site and to enhancer elements which can be many hundreds of kilobases distant from the TSS (Visel et al. 2009). These transcription factors recruit chromatin-modifying enzymes, such as histone acetylases, which open the chromatin around the TSS. The open chromatin allows basal transcription factors to bind core promoter elements such as the TATA box, recruiting pol II to the DNA (Lee and Young 2000).

Once recruited to the DNA, pol II transcribes a short distance, clearing the core promoter. In many genes, the polymerase stalls a short distance into the transcript. This paused pol II may in part be responsible for positioning the +1 nucleosome immediately downstream (Valouev et al. 2011). Binding of transcription factors to the promoter region, and the action of the positive transcription elongation factor B (P-TEFb) kinase, may release the polymerase from this pause into an elongating form (Rahl et al. 2010).

While pol II and the basal transcription factors TFIIB/D/E/F/H together can transcribe naked DNA in vitro, these factors are insufficient for efficient transcription elongation along a nucleosome-bound DNA-template. Pol II has some ability to transcribe through a nucleosome-bound region, although this activity requires the DNA sequence to have a relatively low inherent affinity for nucleosomes (Bondarenko et al. 2006).

There are several distinct but non-exclusive models describing how pol II may transcribe past nucleosomes in vivo. Biochemical complementation assays identified an additional factor, named FACT (facilitates chromatin transcription), which enables pol II to elongate efficiently along a chromatinized DNA template (Orphanides et al. 1998). FACT acts as a histone chaperone, likely by removing one of the outer histone H2A–H2B dimers while leaving the core $(H3–H4)_2$ tetramer intact (Selth et al. 2010). Removal of the H2A–H2B dimer appears to be sufficient to allow pol II to transcribe through the nucleosome, although it is also possible that H2A–H2B dimer displacement is merely a side-effect of increased DNA accessibility caused by FACT activity (Winkler and Luger 2011). In a second model, the entire nucleosome is evicted from the DNA by a histone chaperone. Under a third model, post-translational modification of the histone tails – in particular, acetylation – can reduce nucleosome affinity for DNA, potentially enhancing pol II's inherent ability to move past nucleosome-bound DNA (Selth et al. 2010).

Co-transcriptional regulation through the pol II CTD A number of kinases are involved in transcription by polymerase II, often acting through the carboxy-terminal domain (CTD) repeats of the largest pol II subunit. The consensus repeat amino acids Tyr-Ser-Pro-Thr-Ser-Pro-Ser appear 52 times in the human pol II, and provide a target for these kinases (Lee and Young 2000). Although fewer repeats exist in other organisms such as yeast, the consensus sequence remains the same. Early during transcription, the CTD becomes phosphorylated at Ser5 of these repeats. Following pause release of pol II, Ser2 becomes phosphorylated. By using phospho-specific antibodies, Ser2 phosphorylation can be used to distinguish the fully elongation-competent form of pol II from the initiating or early elongating forms.

Phosphorylation of the pol II CTD is important in regulating not only the action of the polymerase itself but also the post-transcriptional modifications of the nascent mRNA. Following the first CTD phosphorylation event on Ser5, capping enzymes are recruited to the polymerase. This ensures that the $5'$ end of the mRNA receives a 7meG cap immediately upon exit from the polymerase complex and thus protects the message from degradation by $5' \to 3'$ exonucleases (Moore and Proudfoot 2009).

The pol II CTD also recruits the U1 snRNP splicing complex to the nascent transcript as it is being produced, a step that appears to be necessary for efficient pre-mRNA splicing to occur (Das et al. 2006; Das et al. 2007). Splicing reactions can occur even as pol II continues to transcribe. Splicing reactions can finish in 5–10 min and transcription elongation has been measured at approximately 4kb/min. Therefore, in moderately-sized (> 20kb) genes, earlier introns are likely to complete splicing prior to transcription termination (Singh and Padgett 2009).

It has been suggested that the pol II elongation rate itself may be a determinant in the efficiency of co-transcriptional gene modifications. To test this hypothesis, de la Mata et al. (2003) blocked the endogenous pol II complex using the transcription-inhibiting drug α-amanitin, and replaced its action with an elongation-impaired (and α-amanitin-resistant) pol II mutant. The authors found that an alternatively skipped exon in their reporter gene was included at a higher frequency in the final transcript when transcription was switched to the slow pol II mutant. Furthermore, a genome-wide survey of splicing found increased levels of exon inclusion when using the slow polymerase mutant or after partially inhibiting pol II elongation using drugs (Ip et al. 2011). These studies support the hypothesis that slower transcription elongation allows trans factors more time to be recruited to the pre-mRNA, enhancing the recognition of splicing cis-regulatory motifs. However, the pleiotropic effects of genome-wide pol II inhibition are likely to be marked and it is possible that the observed increase in splicing efficiency is a secondary effect of aberrant levels of the trans-factors. Additionally, the biological relevance of pol II elongation rate on splicing has yet to be demonstrated, in large part due to the difficulty in directly measuring the rate at higher resolution. Similar work has suggested a kinetic model may also regulate poly(A) site selection (Pinto et al. 2011).

Transcription termination While the $3'$ end of the mRNA transcript is determined by cleavage and subsequent mRNA polyadenylation, pol II transcription often continues several kb downstream of the cleavage and polyadenylation site (Ford and Hsu 1978; Nevins and Darnell 1978). Proper transcription termination is likely to be important not only for avoiding transcription of downstream genes and recycling pol II, but also for reinitiation of pol II. It has been suggested that chromatin forms loops between the initiating and terminating regions of a gene (Richard and Manley 2009), and a recent study showed that failure to terminate transcription led to downregulation of transcription initiation of the same gene (Mapendano et al. 2010).

Pol II termination and release from DNA have been the subject of many research studies, but the actual molecular mechanisms governing these steps are still poorly understood. Two predominant models have been put forth, and recent work has suggested these mechanisms may work either in parallel or even in concert.

The first model for pol II transcription termination involves co-transcriptional degradation of the pol II-associated RNA following cleavage and polyadenylation of the mRNA. Under this model, the nuclear $5' \rightarrow 3'$ exonuclease Xrn2 (known as Rat1 in yeast) accesses the free $5'$ end of the downstream RNA cleavage product, and degradation proceeds quickly enough that the Xrn2 is able to catch up to the pol II, "torpedoing" it and causing it to terminate transcription (West et al. 2004; Kim et al. 2004).

In contrast to the torpedo model is the so-called allosteric model, in which the cleavage and polyadenylation process directly affects the elongating pol II complex, reducing its stability and leading to drop-off at some point downstream of the cleavage site. In support of this model, it appears that transcription termination can occur prior to cleavage and polyadenylation in vivo in some cases (Rosonina et al. 2006), implying that termination would precede Xrn2 binding and degradation.

A combined termination model has also been proposed. Replacement of the nuclear-localized Xrn2 by its cytoplasmic counterpart Xrn1 led to degradation of the downstream cleavage product following cleavage and polyadenylation, but importantly termination was impaired (Luo et al. 2006). This suggests that Xrn2 plays an important role in transcription termination that is separate from its exonuclease activity. Xrn2 appears to recruit a number of other factors which could potentially mediate this activity (Richard and Manley 2009). The details of such a hybrid model remain to be elucidated, and the mechanisms involved are likely to require a number of factors which may vary from gene to gene and possibly from cell type to cell type.

1.4 Messenger RNA Maturation

5′ end capping Prior to translation into protein, an mRNA must undergo several maturation steps. The first modification is a capping of the 5′ end of the mRNA, which occurs shortly after transcription initiation. The cap was originally discovered in 1974 as a methylated nucleotide in bulk mRNA, after it became possible to cleanly separate poly(A)-containing messenger RNA from ribosomal and transfer RNA (Desrosiers et al. 1974; Perry and Kelley 1974)*. The structure of this methylated nucleotide was soon to be determined as a 7meG, linked by a 5′–5′ triphosphate bridge. Addition of the 7meG cap serves not only to protect the message from degradation by 5′ → 3′ exonucleases but is also required for translation (Muthukrishnan et al. 1975).

Splicing Soon after the characterization of mRNA capping came the observation that nuclear mRNA is much longer than cytoplasmic mRNA, leading to the discovery that the mature cytoplasmic mRNA has removed portions of the DNA-encoded gene (Berget et al. 1977; Chow et al. 1977).

It was suggested and subsequently confirmed that, through base-complementarity, the snRNP U1 ribonucleo-protein complex recognizes a `GU`-containing consensus at the 5′ end of the intron being spliced out (Lerner et al. 1980; Rogers and Wall 1980). The `AG`-containing 3′ splice site sequence as well as an upstream pyrimidine-rich tract is bound by the U2 auxiliary factor U2AF and a further upstream branch point A is cooperatively bound by mBBP†(Wahl et al. 2009). The U2 snRNP displaces SF1/mBBP, leading to an intron bound at both ends by snRNP complexes. Subsequent to U1 and U2 binding, the pre-assembled tri-snRNP U4/U6·U5 is recruited, eventually displacing U1 from the 5′ splice site.

Under the guidance of the spliceosome, the 2′-hydroxyl of the branch point adenosine attacks the 5′ splice site phosphodiester bond, freeing the 3′ end of the upstream exon. The free 3′ hydroxyl then attacks the phosphodiester bond at the 3′ splice site, splicing together the upstream and downstream exons and releasing the noose-shaped intron lariat.

Pre-mRNA splicing is a highly dynamic process involving well over 100 factors (Wahl et al. 2009). Recognition of the correct splice site sequences is aided by several mechanisms. First, as was previously mentioned, splicing factors are recruited by the transcribing pol II complex via its CTD, allowing efficient loading of U1 onto the nascent transcript. In another example of cross-talk between gene regulatory stages, U1 recruitment may also be enhanced by binding to the 5′ cap of the nascent mRNA (Izaurralde et al. 1994; Konarska et al. 1984). Second, stepwise binding to the branch point A, polypyrimidine tract and the splice sites enables independent recognition and verification of the correct splicing sites by multiple factors. Finally, numerous auxiliary cis-regulatory splicing elements (splicing enhancers and silencers) are bound by serine-arginine repeat SR-proteins and interact with the core spliceosomal machinery to enhance recognition of correct splice sites and prevent splicing at incorrect locations (Matlin et al. 2005).

It is thought that splicing occurs with ex-

*Perry and Kelley (1974) was in the inaugural edition of the journal Cell, originally published by the MIT Press.

†Also known, rather unimaginatively, as splicing factor 1, or SF1, not to be confused with the steroidogenic factor 1, also abbreviated SF1.

tremely high fidelity (Wang and Burge 2008). An ongoing challenge has been to understand the nature of this high precision given that modern splicing simulations poorly distinguish true splice sites from decoys, even when modeling all known cis-regulatory sequences.[‡]

Most mammalian genes contain multiple exons, with an average of more than 8 exons per gene in mouse and humans (Roy and Gilbert 2006). The process of splicing is important in joining the coding exons together, revealing the correct open reading frame for subsequent translation. Additionally, these exons can be joined in different patterns, in a process commonly referred to as alternative splicing, producing variation in the resulting proteins as well as varying non-coding regulatory portions of the messages. Recent advances in high throughput sequencing have enabled genome-wide identification of new splicing isoforms as well as quantification of tissue- and treatment-specific splicing patterns (Wang et al. 2008). It was recently estimated that about 90% of mammalian multi-exon genes undergo some sort of alternative splicing (Wang et al. 2008), underscoring the integral role splicing plays in contributing to genome complexity. A holy grail for the splicing field is to use sequence features to predict changes in splicing between tissues. A recent proof-of-principle study was able to accurately predict the direction of change for many alternative splicing events (Barash et al. 2010), although predicting the magnitude of such changes is still difficult.

Discovery of the poly(A) tail A flurry of activity in the late 1960s and '70s demonstrated the importance of the poly(A) tail. First came the observation by Edmonds and Caramela in 1969 of long homopolymeric stretches of adenine in nuclear RNA and the subsequent realization that these polyadenylated RNAs might be precursors to cytoplasmic messenger RNAs (Edmonds et al. 1971). Further characterization demonstrated that these poly(A) sequences came at the 3′ end of mRNA (Molloy et al. 1972).

In 1971, Darnell et al. made several key observations. First, by using a very short time-course, they were able to show that the amount of poly(A) incorporation following actinomycin D treatment significantly exceeded the amount of transcription, indicating that the poly(A) tail is added post-transcriptionally. Secondly, because nuclear RNA becomes polyadenylated prior to the appearance of polysomal, translating, polyadenylated mRNA, they suggested that the nuclear RNA was a precursor of the cytoplasmic mRNA.

Finally, the critical importance of the poly(A) tail was suggested by experiments in which the drug cordycepin was added to cells. Cordycepin is a modified form of adenine which, because it lacks a hydroxyl group its 3′ end, terminates RNA synthesis at A residues. Upon drug treatment, poly(A) tail formation was almost completely abrogated and newly synthesized RNA no longer appeared on polyribosomes. As a result of these experiments, it became clear that the poly(A) tail is an integral step in the maturation of messenger RNA and it was suggested that blocking poly(A) tail addition prevented transport of mRNA from the nucleus to the cytoplasm (Darnell et al. 1973), a suggestion that was ultimately shown to be correct.

Work later that decade explored how this poly(A) tail came to terminate mRNA. Pulse-labelling experiments showed longer transcription products that hybridized to their viral template DNA downstream of the poly(A) site, demonstrating that transcription proceeds beyond the polyadenylation site and that cleavage of the nascent mRNA transcript precedes polyadenylation (Ford and Hsu 1978; Nevins and Darnell 1978). (This was a key observation in understanding the process of transcription termination, discussed on p. 13.)

[‡] It should be noted that we are currently unable to identify the branch point site, let alone predict the positive regulatory effect of having a good or poor branch point sequence. These simulations can only model the efficacy of the splice site sequences and intronic and exonic splicing elements.

Sequence determinants of cleavage and polyadenylation Early sequencing at the 3′ end of mRNA's yielded the `AAUAAA` consensus poly(A) motif (Proudfoot and Brownlee 1976) and subsequent works deleting or mutating this sequence demonstrated its key role in cleavage and polyadenylation (Fitzgerald and Shenk 1981; Wickens and Stephenson 1984). However, the mere presence of the poly(A) signal motif was not sufficient in some cases for efficient cleavage and polyadenylation (Simonsen and Levinson 1983). Mutagenesis of the region downstream of the poly(A) signal of the well-studied simian virus SV40 polyadenylation site suggested the existence of an auxiliary downstream U-rich motif important in poly(A) site recognition (McDevitt et al. 1986).

Recent surveys of expressed sequence tags (ESTs) and cDNA sequences available in public databases gave a genome-wide view of the poly(A) signal sequences in mouse and human (Legendre and Gautheret 2003; Tian et al. 2005). Tian et al. (2005) identified 29,283 poly(A) sites in human and 31,179 poly(A) sites in mouse; of these, over 70% contained either the aforementioned `AAUAA` motif, or the closely related `AUUAAA` hexamer. The other minor-frequency poly(A) signal motifs were A-rich and all but one contained a third-position U. Cleavage and polyadenylation occurred on average at least 21bp downstream of this motif, although there was considerable heterogeneity in the exact cleavage site. The frequent presence of a U-rich downstream sequence element (DSE) 15–30bp downstream of the cleavage site was confirmed, as well as a less common U-rich upstream sequence element (USE) 5′ of the poly(A) site (Legendre and Gautheret 2003). A previously reported GU-rich DSE appears to be used infrequently (Cheng et al. 2006b).

Protein factors involved in poly(A) site recognition and cleavage and polyadenylation Determination of an mRNA's 3′ end is a multi-step process. First, the region must be transcribed by pol II, revealing the RNA cis motifs involved in poly(A) site recognition. Second, the 3′ end processing complex must be recruited to the nascent mRNA, where it cleaves the RNA. Finally, the upstream cleavage product – the mRNA – receives a poly(A) tail. The cleavage and polyadenylation process is mediated by a large protein complex, involving more than 14 core components in mammals (Mandel et al. 2008).

A number of protein subcomplexes are involved in poly(A) site recognition, including CPSF, CstF and cleavage factors CF $\mathrm{I_m}$ and CF $\mathrm{II_m}$. CPSF is first to arrive, at transcription initiation, recruited by the basal transcription factor TFIID to pol II (McCracken et al. 1997; Dantonel et al. 1997). Mammalian CPSF, short for cleavage and polyadenylation specificity factor, is composed of five subunits, which together bind to the `AAUAAA` (or similar) poly(A) signal once it is transcribed. The cleavage stimulation factor, or CstF, comprised of three subunits, binds the U-rich (or GU-rich) DSE downstream element, imparting additional specificity in poly(A) site selection (Takagaki and Manley 1997). CF $\mathrm{I_m}$ also increases specificity for the complex by binding the USE upstream element (Mandel et al. 2008). Despite the detailed dissection of these protein factors, it wasn't until very recently that it was directly shown that the actual endonuclease component is CPSF (Mandel et al. 2006; Takagaki and Manley 1997).

Also integral to the 3′ end complex are the poly(A) polymerase PAP and the nuclear poly(A) binding protein PABPN. Although these two factors are required for the cleavage activity, their main role is in creating the mRNA poly(A) following endonucleolytic cleavage at the poly(A) site. In vitro, PAP is able to add a poly(A) tail to an RNA molecule, but the CPSF complex as well as PABPN are required to specify the correct length of the poly(A) tail (Mandel et al. 2006). The poly(A) polymerase PAP is converted from a non-processive to processive form by association with CPSF and the poly(A) binding protein PABPN. Association of PABPN to the nascent poly(A) tail (positioned one PABPN every 11–14 nucleotides of poly(A) tail) enables the processive polyadenylation reaction only to a final length of 250–300 bp, thereby ensuring a uniform poly(A) tail length (Kühn et al. 2009).

Nearly all pol II transcripts are internally cleaved and polyadenylated by the canonical pathway discussed above. A major exception to this is the replication-dependent histone genes, including H2A, H2B, H3, H4 as well as the linker histone H1 (Marzluff et al. 2008). During S phase, the cellular histone content must double, and this is achieved by a rapid induction of histone gene transcription, followed by sudden degradation of these transcripts at the end of S phase. This sudden degradation is mediated through a stem-loop structure immediately downstream of the histone stop codon, which is recognized in the nucleus by the SLBP stem-loop binding complex (including some canonical cleavage and polyadenylation factors). The histone pre-mRNA is cleaved downstream of the stem-loop, and the bound SLBP performs many of the functions of a canonical poly(A) tail, including protecting the message from degradation and enhancing efficient translation (Marzluff et al. 2008). The sudden degradation of histone mRNAs following S phase is mediated by this structure as well (Pandey and Marzluff 1987).

Regulated cleavage and polyadenylation The number of potential poly(A) sites in the genome far exceeds the number of genes (Tian et al. 2005). At least some of these alternative poly(A) sites are used in a regulated fashion. There are several types of alternative cleavage and polyadenylation. First, alternative last exons include two poly(A) sites, one in an intron and one at the end of the longest isoform. A competition between splicing of the intron and cleavage and polyadenylation of the intronic poly(A) site determines whether the internal or final poly(A) site is used. Second, tandem UTRs include at least two poly(A) sites within the final UTR region. In this case, competition between the poly(A) sites determines which gets used. Lastly, regulation of the poly(A) tail length is also known to occur.

The accuracy of the cleavage and polyadenylation is important, as exemplified by biological regulation and mis-regulation of $3'$ end formation. Mutation of a sub-optimal early poly(A) signal to a higher efficiency form upregulates the prothrombin gene, leading to a hereditary high risk for thrombosis (Danckwardt et al. 2008). Use of proximal tandem $3'$ UTR isoforms is associated with increased proliferation (Sandberg et al. 2008) and oncogenic transformation (Mayr and Bartel 2009). Finally, the U1A gene component of the U1 snRNP provides an example of regulated length of poly(A) tail formation. Outside of the context of the full U1 snRNP complex, the U1A protein can bind to its own $3'$ UTR prior to cleavage, where it represses the action of the poly(A) polymerase, resulting in a short poly(A) tail and lower expression of the U1A gene (Gunderson et al. 1994; Boelens et al. 1993). This is an elegant self-regulatory mechanism which allows U1A to downregulate its own expression when it is in excess over the other U1 snRNP components.

1.5 Post-transcriptional regulation of gene expression

Export Because a failure to correctly splice together the coding exons could lead to aberrant translation of the intronic sequence, quality control mechanisms exist to ensure that only spliced messages are efficiently translated. A critical barrier to translating mis-processed mRNAs is the nuclear membrane. There is evidence suggesting involvement of each of the nuclear mRNA processing steps in regulating nucleo-cytoplasmic export: capping, splicing and cleavage and polyadenylation.

The 7meG cap at the $5'$ end of mRNAs binds to the aptly-named cap-binding complex in the nucleus, and this complex helps recruit export machinery to the mRNA in a splicing-dependent manner (Cheng et al. 2006a). While unspliced messages, most notably the histone genes, are able to be exported, in general, splicing does enhance mRNA export (Luo and Reed 1999; Valencia et al. 2008). Export factors are also recruited to the poly(A) tail, and it is likely that not only cleavage (to release the mRNA from chromatin) but also polyadenylation are required for most mRNAs to become fully export-competent. However since capping, splicing and cleavage and polyadenylation are interdependent processes, it is difficult to tease out direct from indirect effects on export (Bird et al. 2005).

As replication-dependent histones are not spliced nor cleaved as normal mRNAs, and don't receive a poly(A) tail, their export is most likely regulated by export sequences within the mRNA (Erkmann et al. 2005).

Translation and translational control An mRNA is prepared for translation in the cytoplasm by binding of a number of eukaryotic initiation factors (eIFs) to the 5′ cap and the poly(A) tail. Current models suggest these eIFs bridge the 5′ and 3′ ends of the mRNA, circularizing the message (Sonenberg and Hinnebusch 2009). The 40S small ribosomal subunit is recruited to the 5′ untranslated region, where it scans until it finds an `AUG` start codon. The 60S large ribosomal subunit joins with the 40S subunit to form the initiation complex, which can begin translating the mRNA into a polypeptide. Translation continues until a stop codon is reached, which is recognized by a eukaryotic release factor (Amrani et al. 2006).

Translation is a tightly regulated process, including via global mechanisms affecting bulk translation. For example, translation is globally downregulated under stress conditions. Translation of specific messages is also regulated, for example, by upstream open reading frames, or uORFs, which place a stop codon upstream of the true start codon, thereby inhibiting translation initiation of the main ORF.

Nonsense-mediated decay is another important example of interplay between gene regulatory steps. Following splicing in the nucleus, a protein complex called the exon junction complex, or EJC, is placed about 20 bp upstream of the exon-exon junction. Once in the cytoplasm, EJCs are displaced by the ribosome during the pioneer round of translation, but if an error in splicing occurs, placing an in-frame stop codon far enough upstream of an EJC, the message will be recognized as aberrant and it will be degraded (Amrani et al. 2006). This mechanism is an example of how the translation machinery can correct for errors that occurred in the nucleus despite the physical separation of the processes.

Localization and compartmentalization of mRNAs The location of an mRNA within the cytoplasm can play an important regulatory role. Localization of specific messages is determined by binding of trans-factors to localization elements, typically found in the mRNA's 3′ UTR (Martin and Ephrussi 2009). For example, the β-actin mRNA is targeted to sites of active actin polymerization via binding of the zipcode binding protein ZBP1 to the β-actin 3′ UTR (Martin and Ephrussi 2009). ZBP1 enables the active translocation via binding to myosin motors, causing the mRNA to be pulled along the cytoskeleton to the leading edge of migrating cells (Oleynikov and Singer 2003).

In addition to targeted subcellular localization of individual messages, bulk mRNAs are delivered to two types of cytoplasmic foci, stress granules and P bodies. Stress granules compartmentalize mRNAs trapped in translation initiation, and may lead to differences in local concentrations of the translation machinery within the cytoplasm (Buchan and Parker 2009). mRNAs targeted for degradation may become aggregated into a different structure, called a P body. Within the P body, a degradation complex-associated mRNA may be stored or actively degraded. Targeting to P bodies can be modulated by processes such as nonsense mediated decay or stability-regulating factors such as AU-rich elements and microRNAs (see below) (Parker and Sheth 2007). It remains to be shown conclusively that targeting of messages to stress granules or P bodies is causative of translation inhibition or degradation, rather than a reaction to these processes.

Death of an mRNA mRNA degradation typically begins by degradation of the poly(A) tail (Garneau et al. 2007). Two poly(A) nucleases are involved sequentially in the degradation of the poly(A) tail (Yamashita et al. 2005). First, the PAN2/3 nucleases are responsible for steadily shortening the poly(A) tail from its starting length of ~250 bp to around 110 bp. At this point, the CCR4 complex performs a sudden further shortening of the poly(A) tail to a point where the mRNA itself becomes destabilized and undergoes $3' \rightarrow 5'$ degradation by the cytoplasmic exosome or decapping and $5' \rightarrow 3'$ degradation by Xrn1, or likely a combination of both (Garneau et al. 2007). As the poly(A) tail and $5'$ cap are required for efficient translation, deadenylation and decapping are associated with translational inhibition. In certain cases, degradation can begin by a targeted endonucleolytic cleavage event, such as mediated by an siRNA (see below), allowing the exonucleases to access the mRNA and degrade each cleavage fragment.

AU-Rich elements and other stability-regulating elements Because the translocating ribosome is able to displace most RNA-binding trans factors from the open reading frame, the $3'$ untranslated region ($3'$ UTR) plays an integral role in regulation of cytoplasmic mRNAs.

AU-rich elements (AREs) are an important class of $3'$ UTR regulatory sequences. Approximately 20 different RNA-binding factors bind AREs, including AUF1/hnRNP D, the Hu family proteins, and tristetraprolin (or TTP). Some of these factors, such as AUF1, tend to destabilize ARE-containing mRNAs, while others, such as HuR, are thought to antagonize these destabilizing effects through competitive binding to the AREs (Barreau et al. 2005). While AREs are characterized by the occurrence of an `AUUUA` motif, this consensus sequence is not generally sufficient to change a message's stability. Functional AREs are usually found within the context of a larger U or A/U rich region of the $3'$ UTR, but the degeneracy of these motifs makes it difficult to predict from sequence alone functional sites (Barreau et al. 2005).

To identify global binding preferences for RNA-binding factors, a method called CLIP-Seq may be used. CLIP-Seq, short for cross-linking, immunoprecipitation and high-throughput sequencing (the RNA-binding protein analogue to ChIP-Seq) uses UV light to directly cross-link protein to bound RNA, enabling stringent immunoprecipitation conditions to isolate bound RNA (Licatalosi et al. 2008). CLIP-Seq has recently been used to elucidate the binding patterns of HuR (Mukherjee et al. 2011; Lebedeva et al. 2011), and the application of CLIP-Seq or other genome-wide approaches to additional ARE-binding proteins may soon provide a more complete understanding of the ARE motifs and interactions between the various factors binding them.

microRNAs and siRNAs microRNAs are a class of ~22 nt long RNAs which mediate mRNA regulation, primarily through the $3'$ UTR. MicroRNAs, or miRNAs, are transcribed into a primary microRNA transcript by pol II. miRNAs fold back into a hairpin structure which is excised from the pri-miRNA by the nuclear RNase III endonuclease Drosha (Lee et al. 2003). The resulting pre-miRNA is exported to the cytoplasm where another RNase III enzyme, Dicer, cleaves the loop off the pre-miRNA hairpin, liberating the two strands (Bartel 2004). One of the strands, the mature miRNA, is loaded into the Argonaute protein. There are four Argonautes in mammals, Ago1–Ago4 (Filipowicz et al. 2008).

The miRNA recruits the ago-containing silencing complex to an mRNA with complementarity to the so-called miRNA seed sequence, the

2–8 most 5′ nucleotides of the miRNA (Bartel 2009). This silencing complex, known as the RNA-induced silencing complex, or RISC, speeds degradation of targeted messages (Lim et al. 2005) and can also additionally reduce translation (Guo et al. 2010; Hendrickson et al. 2009; Selbach et al. 2008; Baek et al. 2008). The destabilizing effect is generally mediated through enhanced deadenylation followed by decapping and degradation by an exonuclease (Fabian et al. 2010). However, like the intended targets of synthetic siRNAs, miRNA targets with near-perfect complementarity can be endonucleolytically cleaved by Ago2 (Filipowicz et al. 2008).

Importantly, efficacious microRNA target sites can be predicted accurately from the mRNA sequence alone. The accuracy of early prediction methods relied heavily on the level of conservation (Enright et al. 2003; Stark et al. 2003) or conservation above background (Lewis et al. 2003) of sequences complementary to the miRNA seed sequence. While hexamers complementary to nucleotides 2–7 of the microRNA show a small conservation signal and are slightly effective in downregulating the host mRNAs, a full 7mer match to nucleotides 2–8 is much more effective. In mammals, rather than direct sequence complementarity to the first base of the microRNA, an A opposite this position further improves microRNA targeting (Lewis et al. 2005). A number of additional features, such as being in an A/U rich region, improve the prediction of microRNA target sites such that they may be identified and ranked according to efficacy without the need for conservation (Nielsen et al. 2007; Grimson et al. 2007).

Small interfering RNAs, or siRNAs, are a class of short regulatory RNAs related to microRNAs. siRNAs are distinguished by their biogenesis from dsRNA precursors, which are cleaved directly by Dicer (rather than from hairpins which require Drosha cleavage first) and loading into RISC. Once in RISC, siRNAs and miRNAs are functionally equivalent, though because of their origins, siRNAs tend to show extensive complementarity to their targets, leading to endonucleolytic cleavage (Bartel 2004).

1.6 RNA-induced transcriptional silencing in fission yeast

An interesting merging of regulatory mechanisms occurs at the centromeres of the fission yeast *Schizosaccharomyces pombe*, where the RNA-interference machinery is used to recruit chromatin-modifying enzymes, leading to heterochromatinization of the centromeric DNA (Cam et al. 2009; Moazed 2009). Fission yeast is an excellent model organism for studying the effects of RNAi as each enzyme in that pathway is found in single copy in the genome, enabling simple genetic manipulation. Each of the aforementioned proteins is non-essential for viability, but RNAi mutants are unable to form heterochromatin at the centromeres and exhibit chromosome segregation defects (Volpe et al. 2003).

The process of silencing centromeres begins with transcription of the dg and dh repeats in the outer regions of each centromere (Cam et al. 2009; Moazed 2009). These repeat transcripts recruit the RNA-induced transcriptional silencing complex known as RITS. RITS includes the fission yeast homologs of the canonical RNA interference pathway, including Argonaute and Dicer, as well as the RNA-dependent RNA polymerase, an RNA-binding non-canonical poly(A) polymerase and several chromatin-modifying enzymes including most notably an H3K9 methyltransferase. The Dicer enzyme processes duplex repeat RNA into siRNAs that are loaded into Argonaute. The siRNA-containing RITS complex is then recruited to nascent transcripts as they are produced co-transcriptionally, bringing the H3K9 methyltransferase into proximity of the chromatin. Spreading of H3K9me along the DNA establishes heterochromatin throughout the centromere, up until boundary elements demarcated by tRNA genes or other unique sequences (Cam et al. 2005).

An apparent paradox is how transcription of a silenced, heterochromatic region can be responsible for establishing the heterochromatin itself. Recent work showed that transcription is tightly regulated by the cell cycle (Gullerova and Proudfoot 2008; Kloc et al. 2008; Chen et al. 2008). During S phase, pol II is permitted access to transcribe the dg and dh repeats, presumably loading the RITS complex with ample siRNAs for an entire cell cycle including an extended G2 phase. However, what event actually prompts the centromeres to be recognized for heterochromatinization is still unclear (Lejeune and Allshire 2011).

1.7 References

Amrani, N, MS Sachs, and A Jacobson (June 2006). “Early nonsense: mRNA decay solves a translational problem.” In: *Nat Rev Mol Cell Biol* 7.6, pp. 415–25. DOI: 10.1038/nrm1942 (cit. on p. 18).

Baek, D, J Villén, C Shin, FD Camargo, SP Gygi, and DP Bartel (Sept. 2008). “The impact of microRNAs on protein output.” In: *Nature* 455.7209, pp. 64–71. DOI: 10.1038/nature07242 (cit. on p. 20).

Barash, Y, JA Calarco, W Gao, Q Pan, X Wang, O Shai, BJ Blencowe, and BJ Frey (May 2010). “Deciphering the splicing code.” In: *Nature* 465.7294, pp. 53–9. DOI: 10.1038/nature09000 (cit. on p. 15).

Barreau, C, L Paillard, and HB Osborne (2005). “AU-rich elements and associated factors: are there unifying principles?” In: *Nucleic Acids Res* 33.22, pp. 7138–50. DOI: 10.1093/nar/gki1012 (cit. on p. 19).

Barski, A, S Cuddapah, K Cui, TY Roh, DE Schones, Z Wang, G Wei, I Chepelev, and K Zhao (May 2007). “High-resolution profiling of histone methylations in the human genome.” In: *Cell* 129.4, pp. 823–37. DOI: 10.1016/j.cell.2007.05.009 (cit. on p. 10).

Bartel, DP (Jan. 2004). “MicroRNAs: genomics, biogenesis, mechanism, and function.” In: *Cell* 116.2, pp. 281–97 (cit. on pp. 19, 20).

Bartel, DP (Jan. 2009). “MicroRNAs: target recognition and regulatory functions.” In: *Cell* 136.2, pp. 215–33. DOI: 10.1016/j.cell.2009.01.002 (cit. on p. 20).

Berget, SM, C Moore, and PA Sharp (Aug. 1977). “Spliced segments at the 5′ terminus of adenovirus 2 late mRNA.” In: *Proc Natl Acad Sci U S A* 74.8, pp. 3171–5 (cit. on p. 14).

Bird, G, N Fong, JC Gatlin, S Farabaugh, and DL Bentley (Dec. 2005). “Ribozyme cleavage reveals connections between mRNA release from the site of transcription and pre-mRNA processing.” In: *Mol Cell* 20.5, pp. 747–58. DOI: 10.1016/j.molcel.2005.11.009 (cit. on p. 18).

Blat, Y and N Kleckner (July 1999). “Cohesins bind to preferential sites along yeast chromosome III, with differential regulation along arms versus the centric region.” In: *Cell* 98.2, pp. 249–59 (cit. on p. 10).

Boelens, WC, EJ Jansen, WJ van Venrooij, R Stripecke, IW Mattaj, and SI Gunderson (Mar. 1993). “The human U1 snRNP-specific U1A protein inhibits polyadenylation of its own pre-mRNA.” In: *Cell* 72.6, pp. 881–92 (cit. on p. 17).

Bondarenko, VA, LM Steele, A Ujvári, DA Gaykalova, OI Kulaeva, YS Polikanov, DS Luse, and VM Studitsky (Nov. 2006). “Nucleosomes can form a polar barrier to transcript elongation by RNA polymerase II.” In: *Mol Cell* 24.3, pp. 469–79. DOI: 10.1016/j.molcel.2006.09.009 (cit. on p. 12).

Brownell, JE, J Zhou, T Ranalli, R Kobayashi, DG Edmondson, SY Roth, and CD Allis (Mar. 1996). “Tetrahymena histone acetyltransferase A: a homolog to yeast Gcn5p linking histone acetylation to gene activation.” In: *Cell* 84.6, pp. 843–51 (cit. on p. 10).

Buchan, JR and R Parker (Dec. 2009). “Eukaryotic stress granules: the ins and outs of translation.” In: *Mol Cell* 36.6, pp. 932–41. DOI: 10.1016/j.molcel.2009.11.020 (cit. on p. 18).

Cam, HP, T Sugiyama, ES Chen, X Chen, PC FitzGerald, and SIS Grewal (Aug. 2005). “Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome.” In: *Nat Genet* 37.8, pp. 809–19. DOI: 10.1038/ng1602 (cit. on p. 21).

Cam, HP, ES Chen, and SIS Grewal (Feb. 2009). “Transcriptional scaffolds for heterochromatin assembly.” In: *Cell* 136.4, pp. 610–4. DOI: 10.1016/j.cell.2009.02.004 (cit. on p. 20).

Chen, ES, K Zhang, E Nicolas, HP Cam, M Zofall, and SIS Grewal (Feb. 2008). “Cell cycle control of centromeric repeat transcription and heterochromatin assembly.” In: *Nature* 451.7179, pp. 734–7. DOI: 10.1038/nature06561 (cit. on p. 21).

Cheng, H, K Dufu, CS Lee, JL Hsu, A Dias, and R Reed (Dec. 2006a). “Human mRNA export machinery recruited to the 5’ end of mRNA.” In: *Cell* 127.7, pp. 1389–400. DOI: 10.1016/j.cell.2006.10.044 (cit. on p. 18).

Cheng, Y, RM Miura, and B Tian (Oct. 2006b). “Prediction of mRNA polyadenylation sites by support vector machine.” eng. In: *Bioinformatics* 22.19, pp. 2320–5. DOI: 10.1093/bioinformatics/btl394 (cit. on p. 16).

Chow, LT, RE Gelinas, TR Broker, and RJ Roberts (Sept. 1977). “An amazing sequence arrangement at the 5′ ends of adenovirus 2 messenger RNA.” In: *Cell* 12.1, pp. 1–8 (cit. on p. 14).

Danckwardt, S, MW Hentze, and AE Kulozik (Feb. 2008). “3’ end mRNA processing: molecular mechanisms and implications for health and disease.” In: *EMBO J* 27.3, pp. 482–98. DOI: 10.1038/sj.emboj.7601932 (cit. on p. 17).

Dantonel, JC, KG Murthy, JL Manley, and L Tora (Sept. 1997). “Transcription factor TFIID recruits factor CPSF for formation of 3′ end of mRNA.” In: *Nature* 389.6649, pp. 399–402. DOI: 10.1038/38763 (cit. on p. 16).

Darnell, JE, L Philipson, R Wall, and M Adesnik (Oct. 1971). “Polyadenylic acid sequences: role in conversion of nuclear RNA into messenger RNA.” eng. In: *Science* 174.8, pp. 507–10 (cit. on p. 15).

Darnell, JE, WR Jelinek, and GR Molloy (Sept. 1973). “Biogenesis of mRNA: genetic regulation in mammalian cells.” eng. In: *Science* 181.106. review covers early poly(A) developments, pp. 1215–21 (cit. on p. 15).

Das, R, K Dufu, B Romney, M Feldt, M Elenko, and R Reed (May 2006). “Functional coupling of RNAP II transcription to spliceosome assembly.” In: *Genes Dev* 20.9, pp. 1100–9. DOI: 10.1101/gad.1397406 (cit. on p. 12).

Das, R, J Yu, Z Zhang, MP Gygi, AR Krainer, SP Gygi, and R Reed (June 2007). “SR proteins function in coupling RNAP II transcription to pre-mRNA splicing.” In: *Mol Cell* 26.6, pp. 867–81. DOI: 10.1016/j.molcel.2007.05.036 (cit. on p. 12).

de la Mata, M, CR Alonso, S Kadener, JP Fededa, M Blaustein, F Pelisch, P Cramer, D Bentley, and AR Kornblihtt (Aug. 2003). “A slow RNA polymerase II affects alternative splicing in vivo.” In: *Mol Cell* 12.2, pp. 525–32 (cit. on p. 13).

Desrosiers, R, K Friderici, and F Rottman (Oct. 1974). “Identification of methylated nucleosides in messenger RNA from Novikoff hepatoma cells.” In: *Proc Natl Acad Sci U S A* 71.10, pp. 3971–5 (cit. on p. 14).

Edmonds, M and MG Caramela (Mar. 1969). “The isolation and characterization of adenosine monophosphate-rich polynucleotides synthesized by Ehrlich ascites cells.” eng. In: *J Biol Chem* 244.5, pp. 1314–24 (cit. on p. 15).

Edmonds, M, MH Vaughan, and H Nakazato (June 1971). “Polyadenylic acid sequences in the heterogeneous nuclear RNA and rapidly-labeled polyribosomal RNA of HeLa cells: possible evidence for a precursor relationship.” eng. In: *Proc Natl Acad Sci USA* 68.6, pp. 1336–40 (cit. on p. 15).

Enright, AJ, B John, U Gaul, T Tuschl, C Sander, and DS Marks (2003). “MicroRNA targets in Drosophila.” In: *Genome Biol* 5.1, R1. DOI: 10.1186/gb-2003-5-1-r1 (cit. on p. 20).

Erkmann, JA, R Sànchez, N Treichel, WF Marzluff, and U Kutay (Jan. 2005). “Nuclear export of metazoan replication-dependent histone mRNAs is dependent on RNA length and is mediated by TAP.” In: *RNA* 11.1, pp. 45–58. DOI: 10.1261/rna.7189205 (cit. on p. 18).

Fabian, MR, N Sonenberg, and W Filipowicz (2010). “Regulation of mRNA translation and stability by microRNAs.” In: *Annu Rev Biochem* 79, pp. 351–79. DOI: 10.1146/annurev-biochem-060308-103103 (cit. on p. 20).

Filipowicz, W, SN Bhattacharyya, and N Sonenberg (Feb. 2008). “Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?” In: *Nat Rev Genet* 9.2, pp. 102–14. DOI: 10.1038/nrg2290 (cit. on pp. 19, 20).

Fitzgerald, M and T Shenk (Apr. 1981). “The sequence 5′-AAUAAA-3′ forms parts of the recognition site for polyadenylation of late SV40 mRNAs.” In: *Cell* 24.1, pp. 251–60 (cit. on p. 16).

Flemming, W (1882). *Zellsubstanz, Kern und Zelltheilung.* Verlag von FCW Vogel (cit. on p. 10).

Ford, JP and MT Hsu (Dec. 1978). “Transcription pattern of in vivo-labeled late simian virus 40 RNA: equimolar transcription beyond the mRNA 3′ terminus.” eng. In: *J Virol* 28.3, pp. 795–801 (cit. on pp. 13, 15).

Garneau, NL, J Wilusz, and CJ Wilusz (Feb. 2007). “The highways and byways of mRNA decay.” In: *Nat Rev Mol Cell Biol* 8.2, pp. 113–26. DOI: 10.1038/nrm2104 (cit. on p. 19).

Grimson, A, KKH Farh, WK Johnston, P Garrett-Engele, LP Lim, and DP Bartel (July 2007). "MicroRNA targeting specificity in mammals: determinants beyond seed pairing." In: *Mol Cell* 27.1, pp. 91–105. DOI: 10.1016/j.molcel.2007.06.017 (cit. on p. 20).

Gullerova, M and NJ Proudfoot (Mar. 2008). "Cohesin complex promotes transcriptional termination between convergent genes in S. pombe." In: *Cell* 132.6, pp. 983–95. DOI: 10.1016/j.cell.2008.02.040 (cit. on p. 21).

Gunderson, SI, K Beyer, G Martin, W Keller, WC Boelens, and LW Mattaj (Feb. 1994). "The human U1A snRNP protein regulates polyadenylation via a direct interaction with poly(A) polymerase." In: *Cell* 76.3, pp. 531–41 (cit. on p. 17).

Guo, H, NT Ingolia, JS Weissman, and DP Bartel (Aug. 2010). "Mammalian microRNAs predominantly act to decrease target mRNA levels." In: *Nature* 466.7308, pp. 835–40. DOI: 10.1038/nature09267 (cit. on p. 20).

Hendrickson, DG, DJ Hogan, HL McCullough, JW Myers, D Herschlag, JE Ferrell, and PO Brown (Nov. 2009). "Concordant regulation of translation and mRNA abundance for hundreds of targets of a human microRNA." In: *PLoS Biol* 7.11, e1000238. DOI: 10.1371/journal.pbio.1000238 (cit. on p. 20).

Henikoff, S and A Shilatifard (July 2011). "Histone modification: cause or cog?" In: *Trends Genet.* DOI: 10.1016/j.tig.2011.06.006 (cit. on p. 11).

Ip, JY, D Schmidt, Q Pan, AK Ramani, AG Fraser, DT Odom, and BJ Blencowe (Mar. 2011). "Global impact of RNA polymerase II elongation inhibition on alternative splicing regulation." In: *Genome Res* 21.3, pp. 390–401. DOI: 10.1101/gr.111070.110 (cit. on p. 13).

Izaurralde, E, J Lewis, C McGuigan, M Jankowska, E Darzynkiewicz, and IW Mattaj (Aug. 1994). "A nuclear cap binding protein complex involved in pre-mRNA splicing." In: *Cell* 78.4, pp. 657–68 (cit. on p. 14).

Johnson, DS, A Mortazavi, RM Myers, and B Wold (June 2007). "Genome-wide mapping of in vivo protein-DNA interactions." In: *Science* 316.5830, pp. 1497–502. DOI: 10.1126/science.1141319 (cit. on p. 10).

Kaplan, N, IK Moore, Y Fondufe-Mittendorf, AJ Gossett, D Tillo, Y Field, EM LeProust, TR Hughes, JD Lieb, J Widom, and E Segal (Mar. 2009). "The DNA-encoded nucleosome organization of a eukaryotic genome." In: *Nature* 458.7236, pp. 362–6. DOI: 10.1038/nature07667 (cit. on p. 11).

Kim, M, NJ Krogan, L Vasiljeva, OJ Rando, E Nedea, JF Greenblatt, and S Buratowski (Nov. 2004). "The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II." In: *Nature* 432.7016, pp. 517–22. DOI: 10.1038/nature03041 (cit. on p. 13).

Kloc, A, M Zaratiegui, E Nora, and R Martienssen (Apr. 2008). "RNA interference guides histone modification during the S phase of chromosomal replication." In: *Curr Biol* 18.7, pp. 490–5. DOI: 10.1016/j.cub.2008.03.016 (cit. on p. 21).

Konarska, MM, RA Padgett, and PA Sharp (Oct. 1984). "Recognition of cap structure in splicing in vitro of mRNA precursors." In: *Cell* 38.3, pp. 731–6 (cit. on p. 14).

Kornberg, RD (May 1974). "Chromatin structure: a repeating unit of histones and DNA." In: *Science* 184.139, pp. 868–71 (cit. on p. 10).

Kühn, U, M Gündel, A Knoth, Y Kerwitz, S Rüdel, and E Wahle (Aug. 2009). "Poly(A) tail length is controlled by the nuclear poly(A)-binding protein regulating the interaction between poly(A) polymerase and the cleavage and polyadenylation specificity factor." In: *J Biol Chem* 284.34, pp. 22803–14. DOI: 10.1074/jbc.M109.018226 (cit. on p. 17).

Lebedeva, S, M Jens, K Theil, B Schwanhäusser, M Selbach, M Landthaler, and N Rajewsky (Aug. 2011). "Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR." In: *Mol Cell* 43.3, pp. 340–52. DOI: 10.1016/j.molcel.2011.06.008 (cit. on p. 19).

Lee, TI and RA Young (2000). "Transcription of eukaryotic protein-coding genes." In: *Annu Rev Genet* 34, pp. 77–137. DOI: 10.1146/annurev.genet.34.1.77 (cit. on p. 12).

Lee, Y, C Ahn, J Han, H Choi, J Kim, J Yim, J Lee, P Provost, O Rådmark, S Kim, and VN Kim (Sept. 2003). "The nuclear RNase III Drosha initiates microRNA processing." In: *Nature* 425.6956, pp. 415–9. DOI: 10.1038/nature01957 (cit. on p. 19).

Legendre, M and D Gautheret (Feb. 2003). "Sequence determinants in human polyadenylation site selection." In: *BMC Genomics* 4.1, p. 7 (cit. on p. 16).

Lejeune, E and RC Allshire (June 2011). "Common ground: small RNA programming and chromatin modifications." In: *Curr Opin Cell Biol* 23.3, pp. 258–65. DOI: 10.1016/j.ceb.2011.03.005 (cit. on p. 21).

Lerner, MR, JA Boyle, SM Mount, SL Wolin, and JA Steitz (Jan. 1980). "Are snRNPs involved in splicing?" In: *Nature* 283.5743, pp. 220–4 (cit. on p. 14).

Lewis, BP, Ih Shih, MW Jones-Rhoades, DP Bartel, and CB Burge (Dec. 2003). "Prediction of mammalian microRNA targets." In: *Cell* 115.7, pp. 787–98 (cit. on p. 20).

Lewis, BP, CB Burge, and DP Bartel (Jan. 2005). "Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets." In: *Cell* 120.1, pp. 15–20. DOI: 10.1016/j.cell.2004.12.035 (cit. on p. 20).

Licatalosi, DD, A Mele, JJ Fak, J Ule, M Kayikci, SW Chi, TA Clark, AC Schweitzer, JE Blume, X Wang, JC Darnell, and RB Darnell (Nov. 2008). "HITS-CLIP yields genome-wide insights into brain alternative RNA processing." In: *Nature* 456.7221, pp. 464–9. DOI: 10.1038/nature07488 (cit. on p. 19).

Lim, LP, NC Lau, P Garrett-Engele, A Grimson, JM Schelter, J Castle, DP Bartel, PS Linsley, and JM Johnson (Feb. 2005). "Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs." In: *Nature* 433.7027, pp. 769–73. DOI: 10.1038/nature03315 (cit. on p. 20).

Luger, K, AW Mäder, RK Richmond, DF Sargent, and TJ Richmond (Sept. 1997). "Crystal structure of the nucleosome core particle at 2.8 A resolution." In: *Nature* 389.6648, pp. 251–60. DOI: 10.1038/38444 (cit. on pp. 10, 11).

Luo, MJ and R Reed (Dec. 1999). "Splicing is required for rapid and efficient mRNA export in metazoans." In: *Proc Natl Acad Sci U S A* 96.26, pp. 14937–42 (cit. on p. 18).

Luo, W, AW Johnson, and DL Bentley (Apr. 2006). "The role of Rat1 in coupling mRNA 3′-end processing to transcription termination: implications for a unified allosteric-torpedo model." In: *Genes Dev* 20.8, pp. 954–65. DOI: 10.1101/gad.1409106 (cit. on p. 13).

Mandel, CR, S Kaneko, H Zhang, D Gebauer, V Vethantham, JL Manley, and L Tong (Dec. 2006). "Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease." In: *Nature* 444.7121, pp. 953–6. DOI: 10.1038/nature05363 (cit. on p. 16).

Mandel, CR, Y Bai, and L Tong (Apr. 2008). "Protein factors in pre-mRNA 3′-end processing." In: *Cell Mol Life Sci* 65.7-8, pp. 1099–122. DOI: 10.1007/s00018-007-7474-3 (cit. on p. 16).

Mapendano, CK, S Lykke-Andersen, J Kjems, E Bertrand, and TH Jensen (Nov. 2010). "Crosstalk between mRNA 3′ end processing and transcription initiation." In: *Mol Cell* 40.3, pp. 410–22. DOI: 10.1016/j.molcel.2010.10.012 (cit. on p. 13).

Martin, KC and A Ephrussi (Feb. 2009). "mRNA localization: gene expression in the spatial dimension." In: *Cell* 136.4, pp. 719–30. DOI: 10.1016/j.cell.2009.01.044 (cit. on p. 18).

Marzluff, WF, EJ Wagner, and RJ Duronio (Nov. 2008). "Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail." In: *Nat Rev Genet* 9.11, pp. 843–54. DOI: 10.1038/nrg2438 (cit. on p. 17).

Matlin, AJ, F Clark, and CWJ Smith (May 2005). "Understanding alternative splicing: towards a cellular code." In: *Nat Rev Mol Cell Biol* 6.5, pp. 386–98. DOI: 10.1038/nrm1645 (cit. on p. 14).

Mayr, C and DP Bartel (Aug. 2009). "Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells." In: *Cell* 138.4, pp. 673–84. DOI: 10.1016/j.cell.2009.06.016 (cit. on p. 17).

McCracken, S, N Fong, K Yankulov, S Ballantyne, G Pan, J Greenblatt, SD Patterson, M Wickens, and DL Bentley (Jan. 1997). "The C-terminal domain of RNA polymerase II couples mRNA processing to transcription." In: *Nature* 385.6614, pp. 357–61. DOI: 10.1038/385357a0 (cit. on p. 16).

McDevitt, MA, RP Hart, WW Wong, and JR Nevins (Nov. 1986). "Sequences capable of restoring poly(A) site function define two distinct downstream elements." In: *EMBO J* 5.11, pp. 2907–13 (cit. on p. 16).

Moazed, D (Jan. 2009). "Small RNAs in transcriptional gene silencing and genome defence." In: *Nature* 457.7228, pp. 413–20. DOI: 10.1038/nature07756 (cit. on p. 20).

Molloy, GR, MB Sporn, DE Kelley, and RP Perry (Aug. 1972). "Localization of polyadenylic acid sequences in messenger ribonucleic acid of mammalian cells." eng. In: *Biochemistry* 11.17, pp. 3256–60 (cit. on p. 15).

Moore, MJ and NJ Proudfoot (Feb. 2009). "Pre-mRNA processing reaches back to transcription and ahead to translation." In: *Cell* 136.4, pp. 688–700. DOI: 10.1016/j.cell.2009.02.001 (cit. on p. 12).

Mukherjee, N, DL Corcoran, JD Nusbaum, DW Reid, S Georgiev, M Hafner, M Ascano Jr, T Tuschl, U Ohler, and JD Keene (Aug. 2011). "Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability." In: *Mol Cell* 43.3, pp. 327–39. DOI: 10.1016/j.molcel.2011.06.007 (cit. on p. 19).

Muthukrishnan, S, GW Both, Y Furuichi, and AJ Shatkin (May 1975). "5′-Terminal 7-methylguanosine in eukaryotic mRNA is required for translation." In: *Nature* 255.5503, pp. 33–7 (cit. on p. 14).

Nevins, JR and JE Darnell (Dec. 1978). “Steps in the processing of Ad2 mRNA: poly(A)+ nuclear sequences are conserved and poly(A) addition precedes splicing.” eng. In: *Cell* 15.4, pp. 1477–93 (cit. on pp. 13, 15).

Nielsen, CB, N Shomron, R Sandberg, E Hornstein, J Kitzman, and CB Burge (Nov. 2007). “Determinants of targeting by endogenous and exogenous microRNAs and siRNAs.” In: *RNA* 13.11, pp. 1894–910. DOI: 10.1261/rna.768207 (cit. on p. 20).

Oleynikov, Y and RH Singer (Feb. 2003). “Real-time visualization of ZBP1 association with beta-actin mRNA during transcription and localization.” In: *Curr Biol* 13.3, pp. 199–207 (cit. on p. 18).

Olins, AL and DE Olins (Jan. 1974). “Spheroid chromatin units (v bodies).” In: *Science* 183.4122, pp. 330–2 (cit. on p. 10).

Orphanides, G, G LeRoy, CH Chang, DS Luse, and D Reinberg (Jan. 1998). “FACT, a factor that facilitates transcript elongation through nucleosomes.” In: *Cell* 92.1, pp. 105–16 (cit. on p. 12).

Ozsolak, F, JS Song, XS Liu, and DE Fisher (Feb. 2007). “High-throughput mapping of the chromatin structure of human promoters.” In: *Nat Biotechnol* 25.2, pp. 244–8. DOI: 10.1038/nbt1279 (cit. on p. 11).

Pandey, NB and WF Marzluff (Dec. 1987). “The stem-loop structure at the 3’ end of histone mRNA is necessary and sufficient for regulation of histone mRNA stability.” In: *Mol Cell Biol* 7.12, pp. 4557–9 (cit. on p. 17).

Parker, R and U Sheth (Mar. 2007). “P bodies and the control of mRNA translation and degradation.” In: *Mol Cell* 25.5, pp. 635–46. DOI: 10.1016/j.molcel.2007.02.011 (cit. on p. 19).

Perry, R and D Kelley (1974). “Existence of Methylated Messenger-Rna in Mouse L Cells.” In: *Cell* 1.1, pp. 37–42 (cit. on p. 14).

Pinto, PAB, T Henriques, MO Freitas, T Martins, RG Domingues, PS Wyrzykowska, PA Coelho, AM Carmo, CE Sunkel, NJ Proudfoot, and A Moreira (2011). “RNA polymerase II kinetics in polo polyadenylation signal selection.” In: *EMBO J* 30.12, pp. 2431–44. DOI: 10.1038/emboj.2011.156 (cit. on p. 13).

Proudfoot, NJ and GG Brownlee (Sept. 1976). “3′ non-coding region sequences in eukaryotic messenger RNA.” In: *Nature* 263.5574, pp. 211–4 (cit. on p. 16).

Rahl, PB, CY Lin, AC Seila, RA Flynn, S McCuine, CB Burge, PA Sharp, and RA Young (Apr. 2010). “c-Myc regulates transcriptional pause release.” In: *Cell* 141.3, pp. 432–45. DOI: 10.1016/j.cell.2010.03.030 (cit. on p. 12).

Ren, B, F Robert, JJ Wyrick, O Aparicio, EG Jennings, I Simon, J Zeitlinger, J Schreiber, N Hannett, E Kanin, TL Volkert, CJ Wilson, SP Bell, and RA Young (Dec. 2000). “Genome-wide location and function of DNA binding proteins.” In: *Science* 290.5500, pp. 2306–9. DOI: 10.1126/science.290.5500.2306 (cit. on p. 10).

Richard, P and JL Manley (June 2009). “Transcription termination by nuclear RNA polymerases.” In: *Genes Dev* 23.11, pp. 1247–69. DOI: 10.1101/gad.1792809 (cit. on pp. 13, 14).

Robertson, G, M Hirst, M Bainbridge, M Bilenky, Y Zhao, T Zeng, G Euskirchen, B Bernier, R Varhol, A Delaney, N Thiessen, OL Griffith, A He, M Marra, M Snyder, and S Jones (Aug. 2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." In: *Nat Methods* 4.8, pp. 651–7. DOI: 10.1038/nmeth1068 (cit. on p. 10).

Roeder, RG and WJ Rutter (Oct. 1969). "Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms." In: *Nature* 224.5216, pp. 234–7 (cit. on p. 11).

Rogers, J and R Wall (Apr. 1980). "A mechanism for RNA splicing." In: *Proc Natl Acad Sci U S A* 77.4, pp. 1877–9 (cit. on p. 14).

Rosonina, E, S Kaneko, and JL Manley (May 2006). "Terminating the transcript: breaking up is hard to do." In: *Genes Dev* 20.9, pp. 1050–6. DOI: 10.1101/gad.1431606 (cit. on p. 13).

Roy, SW and W Gilbert (Mar. 2006). "The evolution of spliceosomal introns: patterns, puzzles and progress." In: *Nat Rev Genet* 7.3, pp. 211–21. DOI: 10.1038/nrg1807 (cit. on p. 15).

Sandberg, R, JR Neilson, A Sarma, PA Sharp, and CB Burge (June 2008). "Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites." In: *Science* 320.5883, pp. 1643–7. DOI: 10.1126/science.1155390 (cit. on p. 17).

Selbach, M, B Schwanhäusser, N Thierfelder, Z Fang, R Khanin, and N Rajewsky (Sept. 2008). "Widespread changes in protein synthesis induced by microRNAs." In: *Nature* 455.7209, pp. 58–63. DOI: 10.1038/nature07228 (cit. on p. 20).

Selth, LA, S Sigurdsson, and JQ Svejstrup (2010). "Transcript Elongation by RNA Polymerase II." In: *Annu Rev Biochem* 79, pp. 271–93. DOI: 10.1146/annurev.biochem.78.062807.091425 (cit. on p. 12).

Simonsen, CC and AD Levinson (Dec. 1983). "Analysis of processing and polyadenylation signals of the hepatitis B virus surface antigen gene by using simian virus 40-hepatitis B virus chimeric plasmids." In: *Mol Cell Biol* 3.12, pp. 2250–8 (cit. on p. 16).

Singh, J and RA Padgett (Nov. 2009). "Rates of in situ transcription and splicing in large human genes." In: *Nat Struct Mol Biol* 16.11, pp. 1128–33. DOI: 10.1038/nsmb.1666 (cit. on p. 13).

Solomon, MJ, PL Larsen, and A Varshavsky (June 1988). "Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene." In: *Cell* 53.6, pp. 937–47 (cit. on p. 10).

Sonenberg, N and AG Hinnebusch (Feb. 2009). "Regulation of translation initiation in eukaryotes: mechanisms and biological targets." In: *Cell* 136.4, pp. 731–45. DOI: 10.1016/j.cell.2009.01.042 (cit. on p. 18).

Stark, A, J Brennecke, RB Russell, and SM Cohen (Dec. 2003). "Identification of Drosophila MicroRNA targets." In: *PLoS Biol* 1.3, E60. DOI: 10.1371/journal.pbio.0000060 (cit. on p. 20).

Suganuma, T and JL Workman (June 2011). "Signals and combinatorial functions of histone modifications." In: *Annu Rev Biochem* 80, pp. 473–99. DOI: 10.1146/annurev-biochem-061809-175347 (cit. on p. 10).

Takagaki, Y and JL Manley (July 1997). “RNA recognition by the human polyadenylation factor CstF.” In: *Mol Cell Biol* 17.7, pp. 3907–14 (cit. on p. 16).

Thomas, MC and CM Chiang (2006). “The general transcription machinery and general cofactors.” In: *Crit Rev Biochem Mol Biol* 41.3, pp. 105–78. DOI: 10.1080/10409230600648736 (cit. on p. 11).

Tian, B, J Hu, H Zhang, and CS Lutz (Jan. 2005). “A large-scale analysis of mRNA polyadenylation of human and mouse genes.” eng. In: *Nucleic Acids Res* 33.1. - look at refs: 13 (poly-A tail length), 14, 15, pp. 201–12. DOI: 10.1093/nar/gki158 (cit. on pp. 16, 17).

Trojer, P and D Reinberg (Oct. 2007). “Facultative heterochromatin: is there a distinctive molecular signature?” In: *Mol Cell* 28.1, pp. 1–13. DOI: 10.1016/j.molcel.2007.09.011 (cit. on p. 10).

Valencia, P, AP Dias, and R Reed (Mar. 2008). “Splicing promotes rapid and efficient mRNA export in mammalian cells.” In: *Proc Natl Acad Sci U S A* 105.9, pp. 3386–91. DOI: 10.1073/pnas.0800250105 (cit. on p. 18).

Valouev, A, SM Johnson, SD Boyd, CL Smith, AZ Fire, and A Sidow (June 2011). “Determinants of nucleosome organization in primary human cells.” In: *Nature* 474.7352, pp. 516–20. DOI: 10.1038/nature10002 (cit. on p. 12).

Visel, A, EM Rubin, and LA Pennacchio (Sept. 2009). “Genomic views of distant-acting enhancers.” In: *Nature* 461.7261, pp. 199–205. DOI: 10.1038/nature08451 (cit. on p. 12).

Volpe, T, V Schramke, GL Hamilton, SA White, G Teng, RA Martienssen, and RC Allshire (2003). “RNA interference is required for normal centromere function in fission yeast.” In: *Chromosome Res* 11.2, pp. 137–46 (cit. on p. 20).

Wahl, MC, CL Will, and R Lührmann (Feb. 2009). “The spliceosome: design principles of a dynamic RNP machine.” In: *Cell* 136.4, pp. 701–18. DOI: 10.1016/j.cell.2009.02.009 (cit. on p. 14).

Wang, ET, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, SF Kingsmore, GP Schroth, and CB Burge (Nov. 2008). “Alternative isoform regulation in human tissue transcriptomes.” In: *Nature* 456.7221, pp. 470–6. DOI: 10.1038/nature07509 (cit. on p. 15).

Wang, Z and CB Burge (May 2008). “Splicing regulation: from a parts list of regulatory elements to an integrated splicing code.” In: *RNA* 14.5, pp. 802–13. DOI: 10.1261/rna.876308 (cit. on p. 15).

West, S, N Gromak, and NJ Proudfoot (Nov. 2004). “Human $5' \rightarrow 3'$ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites.” In: *Nature* 432.7016, pp. 522–5. DOI: 10.1038/nature03035 (cit. on p. 13).

Wickens, M and P Stephenson (Nov. 1984). “Role of the conserved AAUAAA sequence: four AAUAAA point mutants prevent messenger RNA $3'$ end formation.” In: *Science* 226.4678, pp. 1045–51 (cit. on p. 16).

Winkler, DD and K Luger (May 2011). “The histone chaperone FACT: structural insights and mechanisms for nucleosome reorganization.” In: *J Biol Chem* 286.21, pp. 18369–74. DOI: 10.1074/jbc.R110.180778 (cit. on p. 12).

Woodcock, CL and RP Ghosh (May 2010). "Chromatin higher-order structure and dynamics." In: *Cold Spring Harb Perspect Biol* 2.5, a000596. DOI: `10.1101/cshperspect.a000596` (cit. on p. 10).

Yamashita, A, TC Chang, Y Yamashita, W Zhu, Z Zhong, CYA Chen, and AB Shyu (Dec. 2005). "Concerted action of poly(A) nucleases and decapping enzyme in mammalian mRNA turnover." In: *Nat Struct Mol Biol* 12.12, pp. 1054–63. DOI: `10.1038/nsmb1016` (cit. on p. 19).

Yuan, GC, YJ Liu, MF Dion, MD Slack, LF Wu, SJ Altschuler, and OJ Rando (July 2005). "Genome-scale identification of nucleosome positions in S. cerevisiae." In: *Science* 309.5734, pp. 626–30. DOI: `10.1126/science.1112178` (cit. on p. 11).

Zhang, Y, Z Moqtaderi, BP Rattner, G Euskirchen, M Snyder, JT Kadonaga, XS Liu, and K Struhl (Aug. 2009). "Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo." In: *Nat Struct Mol Biol* 16.8, pp. 847–52. DOI: `10.1038/nsmb.1636` (cit. on p. 11).

Zhang, Z, CJ Wippo, M Wal, E Ward, P Korber, and BF Pugh (May 2011). "A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome." In: *Science* 332.6032, pp. 977–80. DOI: `10.1126/science.1200508` (cit. on p. 11).

Zhou, VW, A Goren, and BE Bernstein (Jan. 2011). "Charting histone modifications and the functional organization of mammalian genomes." In: *Nat Rev Genet* 12.1, pp. 7–18. DOI: `10.1038/nrg2905` (cit. on p. 10).

Chapter 2

Biased Chromatin Signatures around Polyadenylation Sites and Exons

Noah Spies, Cydney B Nielsen, Richard A Padgett and Christopher B Burge

Abstract

Core RNA-processing reactions in eukaryotic cells occur cotranscriptionally in a chromatin context, but the relationship between chromatin structure and pre-mRNA processing is poorly understood. We observed strong nucleosome depletion around human polyadenylation sites (PAS) and nucleosome enrichment just downstream of PAS. In genes with multiple alternative PAS, higher downstream nucleosome affinity was associated with higher PAS usage, independently of known PAS motifs that function at the RNA level. Conversely, exons were associated with distinct peaks in nucleosome density. Exons flanked by long introns or weak splice sites exhibited stronger nucleosome enrichment, and incorporation of nucleosome density data improved splicing simulation accuracy. Certain histone modifications, including H3K36me3 and H3K27me2, were specifically enriched on exons, suggesting active marking of exon locations at the chromatin level. Together, these findings provide evidence for extensive functional connections between chromatin structure and RNA processing.

Contents

This chapter was published as:

Spies, N, CB Nielsen, RA Padgett, and CB Burge (Oct. 2009). "Biased chromatin signatures around polyadenylation sites and exons." In: *Mol Cell* 36.2, pp. 245–54. DOI: 10.1016/j.molcel.2009.10.008.

Author contributions N.S. performed analyses of chromatin enrichment in relation to splicing, and C.B.N. performed similar analyses for poly(A) sites. The paper was written by N.S., C.B.N and C.B.B with input from R.A.P.

Acknowledgments We thank G. Frampton, C. Lin, R. Sandberg, P. Sharp, R. Young, and members of the C. Guthrie and C. Burge labs for helpful comments and discussions.

2.1 Introduction

In multicellular organisms, most primary RNA transcripts undergo extensive processing. Both pre-mRNA splicing and cleavage/polyadenylation are usually initiated or completed cotranscriptionally, and several mechanistic links between transcription and RNA processing are known. Upon phosphorylation of the C-terminal domain (CTD) of RNA polymerase II (Pol II) shortly after transcription initiation, 5′ end-capping enzymes are recruited to the nascent transcript (Moore and Proudfoot 2009). Factors central to pre-mRNA splicing are loaded onto the CTD (Kornblihtt et al. 2004), and some splicing factors, including the U1 and U2 snRNPs and SR proteins, are deposited on nascent pre-mRNAs as the 5′ and 3′ splice sites are transcribed (Görnemann et al. 2005; Lin et al. 2008). Similarly, cleavage and polyadenylation factors associate with the phosphorylated CTD and recognize the polyadenylation signal, often before Pol II termination (Moore and Proudfoot 2009). The kinetics of transcription can influence both pre-mRNA splicing (Howe et al. 2003; de la Mata et al. 2003) and cleavage and polyadenylation; for example, Pol II pausing in the 3′ region of a gene has been shown to favor use of the more 5′ among two alternative PAS (Peterson et al. 2002).

Pre-mRNA splicing requires extremely precise identification of the correct 5′ and 3′ splice sites, frequently from among many kilobases of intronic sequence containing a large excess of potential splice sites that are not used. Additional cis-regulatory RNA sequence elements, including exonic splicing enhancers (ESEs) and silencers (ESSs), assist in the accurate identification of splice sites, generally through recruitment of factors of the serine/arginine-rich (SR) protein and heterogeneous nuclear ribonucleoprotein (hnRNP) classes. Despite fairly extensive characterization of such elements, splicing simulators that incorporate these elements are still only able to correctly identify approximately half to two-thirds of exons from the sequence alone (Wang et al. 2004), underscoring that the complete set of rules for recognition of exons by the splicing machinery remains to be determined.

Transcription is influenced by chromatin structure and by histone modifications such as methylation and acetylation; both nucleosome-positioning and modification status are in turn influenced by the process of transcription (Li et al. 2007). The Pol II CTD functions not only to recruit RNA-processing factors but also chromatin-modifying factors. For example, the enzyme responsible for trimethylation of histone H3 lysine 36 (H3K36me3) is recruited to the Ser2-phosphorylated CTD, thereby establishing a pattern of this modification that is biased toward the downstream regions of expressed genes (Li et al. 2007).

A handful of recent studies have identified links between histone modifications and RNA processing (Kolasinska-Zwierz et al. 2009; Schor et al. 2009; Loomis et al. 2009; Sims et al. 2007), but whether aspects of nucleosome positioning and modification influence RNA processing generally (or vice versa) is not known. Here we show that specific chromatin signatures are associated with exons and with sites of cleavage and polyadenylation, and correlate with the strength or usage of these RNA elements, establishing a framework for interaction between chromatin structure and RNA processing.

2.2 Results

Nucleosomes Are Strongly Enriched on Exons We observed that nucleosomes are significantly enriched on DNA encoding internal exons compared to flanking introns (Figure 1) through analysis of high-throughput nucleosome chromatin immunoprecipitation and sequencing

(ChIP-Seq) data from human T cells (Schones et al. 2008). The magnitude of the enrichment on exons (1.41-fold enrichment for nucleosomes above background; 95% confidence interval, [1.408, 1.425], by resampling) rivals or exceeds that observed at the +1 nucleosome peak near the transcription start site (TSS) when plotted using the same data set and methods (Figure 1G). We also observed similar enrichment of nucleosomes on exons in published data from the Japanese killifish (data not shown; Sasaki et al. [2009]).

Biased Exon Composition Explains Nucleosome Enrichment We hypothesized that this enrichment might be explained at least partially by sequence features specific to exons, such as splice site- or ESE-related motifs. We analyzed sets of "decoy" 3′ splice site (3′ss) and decoy 5′ splice site (5′ss) sequences in introns, i.e., sequences that match the 3′ss or 5′ss consensus as well as authentic splice sites but are not observed to be used in splicing. Nucleosome density was only slightly enriched in the vicinity of decoy 3′ss and 5′ss (Figures 1D and 1F), suggesting that the enrichment on exons cannot be explained simply by effects of oligonucleotides that form parts of the splice site consensus motifs.

An alternative possibility was that the exonic bias of nucleosomes might be attributable to the distinctive oligonucleotide composition of exons (Denisov et al. 1997; Baldi et al. 1996). To explore this possibility, exon-sized stretches of nucleotides in intergenic regions or introns were identified that scored as high on exonic character as authentic exons but lacked evidence of splicing nearby and were flanked by regions of typically intronic character; we refer to these stretches as exonic composition regions (ECRs). Here, exonic or intronic character was assessed using homogeneous fifth-order Markov models (Burge and Karlin 1997) that captured the distinctive hexanucleotide (6-mer) compositions of human exons and introns but did not consider reading frame or splice site motifs. Notably, these ECRs exhibited strong enrichment for nucleosome density comparable in magnitude to that observed in authentic exons (Figure 1E). Nucleosome enrichment was similar for ECRs located in annotated intergenic regions or intronic regions, and remained when controlling for the mappability of genomic positions and for the biased 5′ nucleotide content of ChIP-Seq reads (see Figure S1, found in Appendix A). We conclude from these observations that nucleosomes are preferentially localized to exons and that the biased oligonucleotide content of exons can explain at least a major part of this effect. That oligonucleotide content could create such a strong bias in nucleosome position is supported by recent studies indicating that intrinsic DNA sequence preferences play a central role in determining nucleosome organization in vivo (Kaplan et al. 2009).

Specific Histone Marks Are Enriched on Exons In addition to nucleosome positions, the patterns of methylation marks on specific residues of the component histones can also play important roles in regulation of gene expression. These roles include demarcation of functional genomic regions and recruitment of protein factors to DNA, including both transcription and RNA-processing factors (Sims et al. 2007). Based on published genome-wide histone methylation data in human T cells (Barski et al. 2007), the enrichment of specific methylation marks on exons was assessed by calculating the ratio of ChIP-Seq read density in exons to that in the flanking introns. Because many histone methylation marks show increasing or decreasing densities from beginning to end of genes, flanking intronic read density was estimated based on the average of regions located equidistantly 5′ and 3′ of each exon. Using this measure, all methylated forms of histones except H3K9me3 were significantly enriched on exons relative to flanking intronic regions (Figure 2A; $p < 0.01$ after Bonferroni correction for multiple testing, bootstrap sampling test). ChIP-Seq data for the chromatin insulator factor CTCF from the same study did not exhibit a bias toward exons relative to introns (Figures 2A and 2F). Because

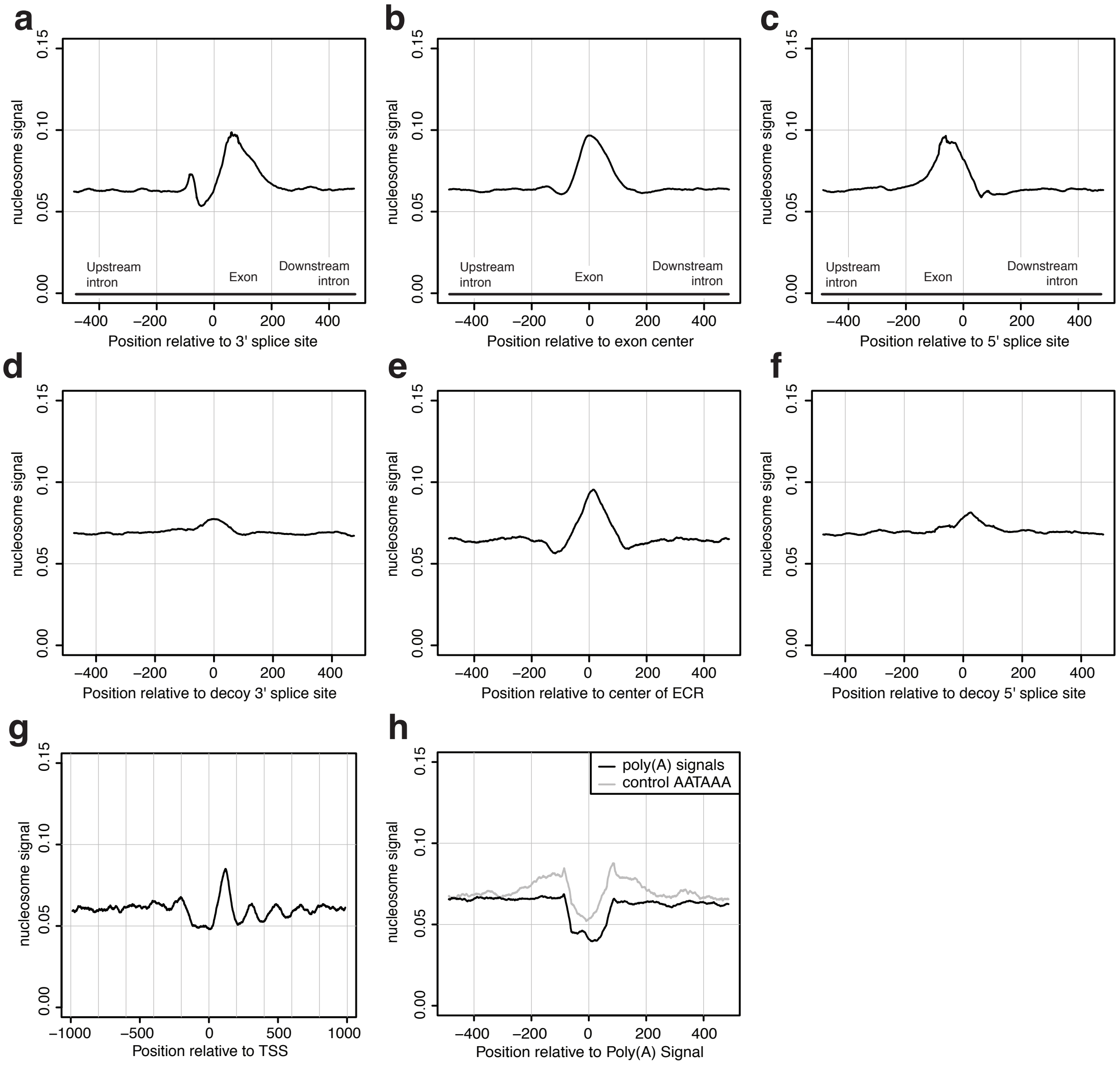


Figure 1: Nucleosome Enrichment and Depletion in the Vicinity of Core Sites of RNA Processing and Controls
Nucleosome read signal, centered on A 3′ss, B exon centers, and C 5′ss, with approximate exon sizes indicated by black box below. Nucleosome signal relative to D sequence-matched decoy 3′ss, E regions of exonic nucleotide composition, and F decoy 5′ ss. For reference, we have plotted nucleosome signal on the same y axis for G TSSs of expressed genes and H PAS.

CTCF is not associated with nucleosomes, these data serve as a type of negative control, indicating that the exonic biases observed are not simply some sort of artifact of the ChIP-Seq protocol.

Most of the observed overall average 1.3-fold enrichment of histone marks on exons can be attributed to the increased nucleosome density on exons observed above. However, two marks in particular were enriched in exons by 1.5-fold or more, significantly exceeding the average enrichment of nucleosomes (and of histone marks overall) on exons. These marks included not only the classical transcription elongation mark H3K36me3, whose enrichment on exon-associated nucleosomes has been previously noted (Kolasinska-Zwierz et al. 2009), but also H3K27me2, which is less associated with transcription elongation. H3K27me2 has generally been associated with repressed rather than active chromatin (Barski et al. 2007). Since many of these marks show distinctive patterns within gene bodies, we investigated whether their enrichment on exons was dependent on po-

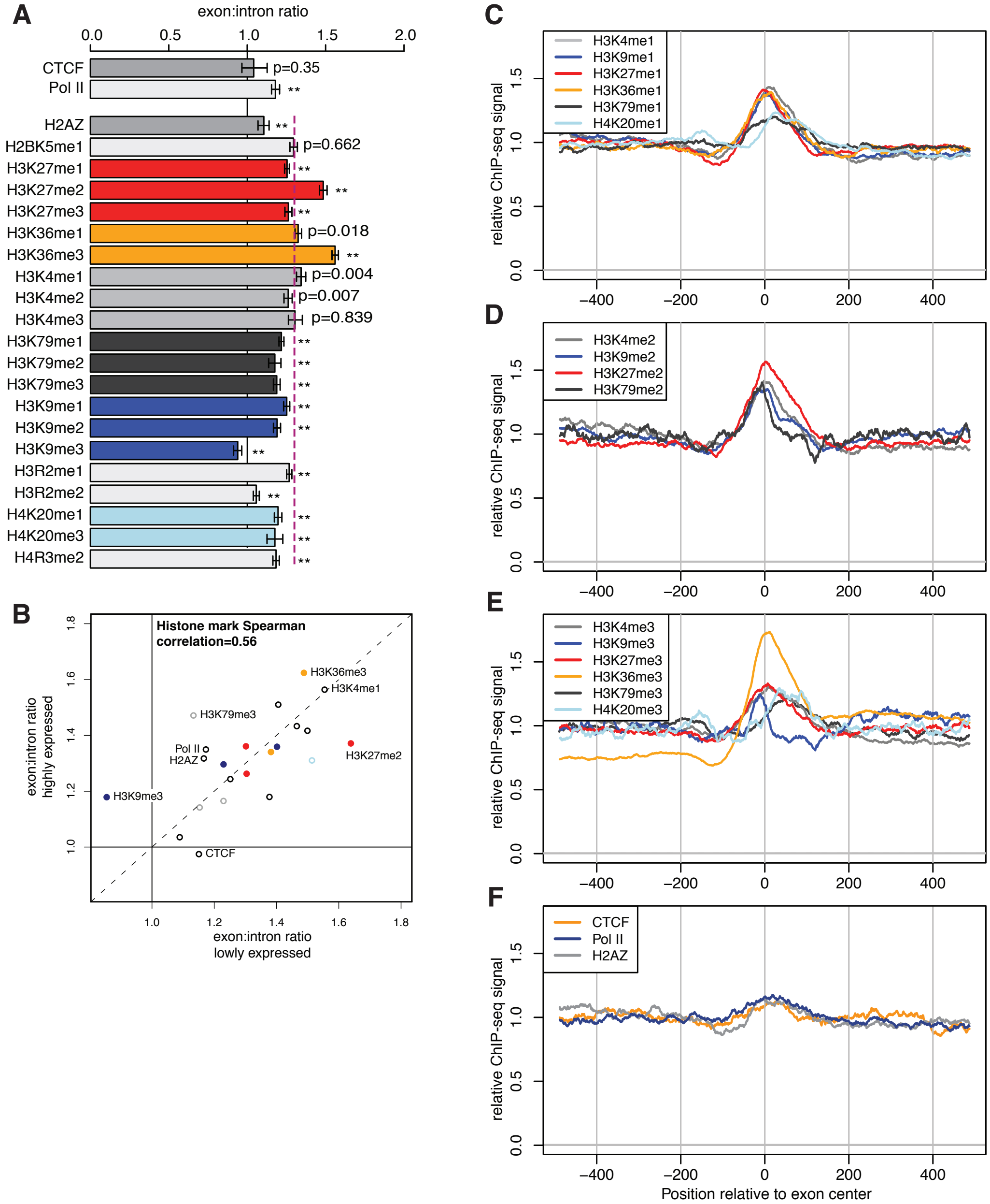


Figure 2: Exon-Biased Distribution of Specific Histone H3 Methylation Marks

A ChIP enrichment for exons, relative to flanking intronic regions (see the Experimental Procedures), compared to 1.0 (CTCF and Pol II) or histone overall average of 1.3 (purple dashed line). Error bars are 95% confidence intervals (resampling). $^{**}p < 0.01$ after correction for multiple testing (resample test, Bonferroni corrected). B Histone marks are similarly enriched in highly and lowly expressed genes. Profiles centered on exons for C monomethyl histone marks, D dimethyl histone marks, and E trimethyl histone marks and Pol II, H2AZ, and the negative control CTCF F. C–F are normalized to average library ChIP signal across the displayed region.

sition relative to the TSS. An overall increase in nucleosome enrichment at larger distances from the TSS was observed (Figure S2, Appendix A). The H3K4me3 mark showed characteristic enrichment for both exons and introns located near the TSS, but exon:intron ratios for this mark and for other position-biased marks generally increased with distance from the TSS (Figure S3).

The pronounced enrichment of these marks on exons suggested potential connections between RNA processing and histone methylation. For example, cotranscriptional recognition of exons at the RNA level might in some way influence methylation of specific histone residues in exon-associated nucleosomes or vice versa. Comparing histone marks in subsets of genes that were either expressed or not expressed in human T cells (based on mRNA microarray data), we observed that most histone marks exhibited similar levels of enrichment in exons independent of transcriptional activity (Figure 2B, Figure S4). These data suggested the possibility that differential marking of exons may not require transcription and RNA processing but may contribute to recognition and even definition of exons at the RNA level, e.g., through direct recognition of histone marks by RNA-processing factors or by factors that modify or interact with RNA-processing factors. It is also possible that the observed differential marking of exons was established at an earlier stage in cellular differentiation during which these genes were expressed. In contrast, a few marks, most notably H3K27me2, were significantly less exon-enriched in genes with high expression in human T cells. One mark, H3K9me3, was unusual in being underrepresented rather than overrepresented in exons (Figure 2A), suggesting that this repression-associated mark might have a different relationship to RNA processing than other marks.

ChIP-Seq reads for RNA Pol II were marginally enriched on exons, with somewhat higher enrichment observed in highly expressed genes (Figure 2), but enrichment was not significant in the data from Schones and coworkers. Pol II enrichment, if it occurs, could result from slowing of the polymerase due to the presence of increased nucleosome density or specific histone marks, or to recognition of splicing-related motifs in the nascent transcript by splicing factors associated with the Pol II CTD.

Isolated Exons Have Stronger Nucleosome Enrichment The specificity of exon recognition by the pre-mRNA splicing machinery is not completely understood (Wang et al. 2004). While the core splice site motifs and known splicing-regulatory elements located in exons and introns play central roles in splicing specificity, these motifs do not appear sufficient to define exon locations with high accuracy. The insufficiency of known motifs is particularly acute for mammalian genes with long, multikilobase introns, where more information is required to distinguish authentic exons and splice sites from the larger pool of decoys (Wang et al. 2004; Lim and Burge 2001). Notably, nucleosome enrichment was significantly greater for “isolated” exons flanked by long introns compared to “clustered” exons flanked by short introns, with both lower intronic nucleosome density and a sharper peak of exonic density observed for isolated exons (Figure 3).

A subset of histone methylation marks also showed significantly higher enrichment in isolated exons, including both of the marks most highly enriched globally in exons – H3K27me2 and H3K36me3 – as well as H3K4me3, H3K27me1, and H3K36me1, but not the insulator element CTCF (Figure 3A). Since the information requirements for accurate splicing of longer transcripts containing isolated exons are intrinsically higher, the increased enrichment of nucleosomes and of specific exon-associated histone marks on isolated exons represents a source of information encoded in the chromatin that would be particularly useful for ensuring accurate pre-mRNA splicing if it could be read out by the splicing machinery. Of course, the potential of marks that are extremely rare on actively expressed genes to contribute to the overall specificity of pre-mRNA splicing

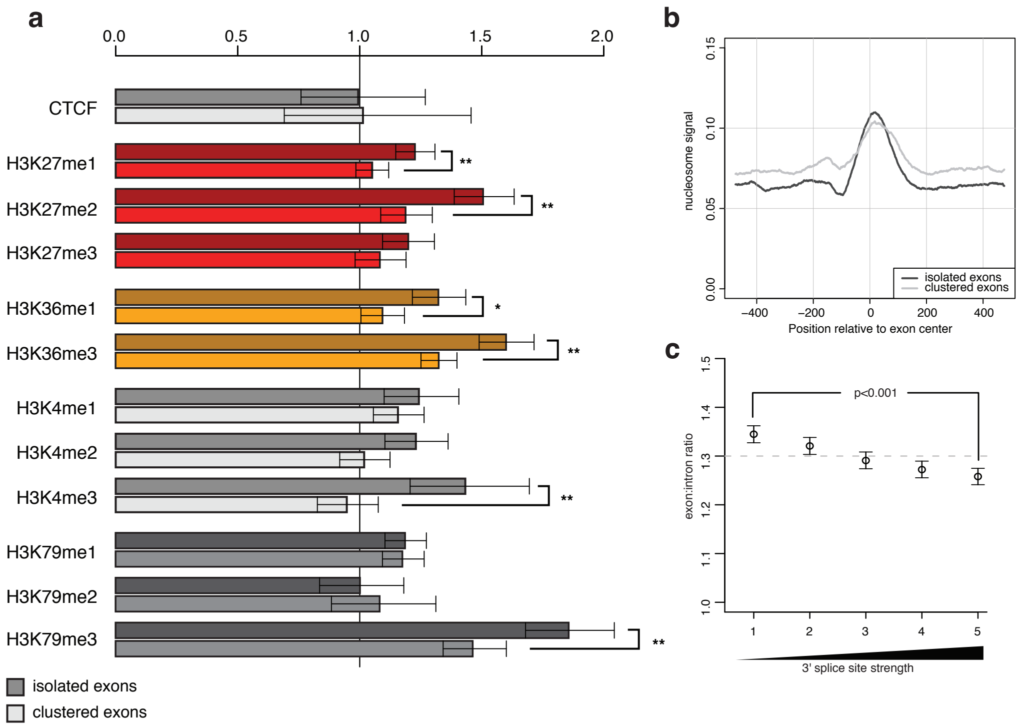


Figure 3: Increased Exonic Bias of Specific Histone H3 Methylation Marks in Exons with Long Flanking Introns or Weaker 3′ss Motifs

(A) Exon enrichment, relative to flanking introns for isolated exons (flanking introns > 5 kb, top bar of each pair) and clustered exons (flanking introns between 0.5 and 1.0 kb). Error bars are 95% confidence intervals. $*p < 0.05$ and $**p < 0.01$ after Bonferroni correction for multiple testing (resample test).
(B) Nucleosome signal profile for exons with short and with long flanking introns.
(C) Nucleosome enrichment on exons is inversely correlated with 3′ splice site strength.

is less than for similarly exon-enriched marks that are abundant in expressed genes. Recognition of one of these enriched marks, H3K4me3, is known to facilitate pre-mRNA splicing, likely mediated through the CHD1 protein, which interacts both with H3K4me3 and with components of the spliceosome (Sims et al. 2007). Interestingly, H3K4me3 enrichment on isolated exons was significantly more pronounced in highly expressed genes, although considerable variance was observed when comparing highly and lowly expressed isolated exons with clustered exons (Figure S5). Previously, lower density of H3K36me3 was reported in alternative exons relative to constitutive exons (Kolasinska-Zwierz et al. 2009). However, in our analyses using larger data sets of alternative exons, significant differences in the density of histone marks in alternative relative to constitutively spliced exons were not detected (Experimental Procedures).

Weak Splice Site Exons Have Stronger Nucleosome Enrichment Sequence features that enhance recognition of exons, including both ESEs and intronic splicing enhancers (ISEs), are common in and adjacent to constitutively spliced exons, and are particularly enriched when core splice site motifs are weaker, i.e., have below-average match to the consensus (Xiao et al. 2009; Murray et al. 2008; Fairbrother et al. 2002). Considering the relationship between splice site strength and nucleosome density, a significant negative correlation between 3′ss strength and exonic nucleosome enrichment was observed (Figure 3C; $p < 0.001$, comparing strongest and weakest splice site strength bins, bootstrap sampling test). The inverse correlation with 3′ss strength persisted after controlling for splice site distance to the TSS (Figure S6), flanking intron length, and exonic oligonucleotide composition, indicating that the association is largely independent of these variables. An inverse relationship was also observed between exonic nucleosome enrichment and 5′ss strength, though this relationship was less pronounced than for the 3′ss (Figure S7). This inverse relationship was also apparent for H3K36me3 and H3K27me2, but not for H3K79me3 (Figure S8). Thus, as for isolated versus clustered exons, a more pronounced nucleosome enrichment signal was observed for the subset of exons expected to have the greatest requirements for splicing enhancement.

Nucleosome Locations Enhance Splicing Simulation Accuracy The patterns of nucleosome enrichment on exons observed above suggested the hypothesis that nucleosome positions might contribute to recognition of exons in pre-mRNA splicing. Under this hypothesis, inclusion of nucleosome position information should improve the accuracy of algorithms that seek to simulate splicing specificity, such as ExonScan (Wang et al. 2004). For this purpose, log-odds scores were derived for specific ranges of exonic nucleosome density in a training set of 1000 genes based on the nucleosome data of Schones and coworkers (Schones et al. 2008). Application of this scoring model using empirical nucleosome densities in a separate set of 12,800 genes yielded modest but highly significant improvements in the prediction of exon locations (Table 1). This improvement occurred whether nucleosome scoring was incorporated into models involving scoring of 5′ss and 3′ss motifs only or using the full model that included also scoring of ESEs, ESSs, and ISEs. The latter result indicated that exonic nucleosome density provides additional information useful for exon recognition beyond that present in known splicing motifs.

Polyadenylation Sites Are Strongly Depleted of Nucleosomes Previous work has suggested connections between transcript termination, chromatin structure, and histone modification (Lian et al. 2008). Additionally, a nucleosome-depleted region has been observed near the PAS in yeast (Mavrich et al. 2008). We observed a sharp dip in nucleosome signal around human PAS, extending roughly 100 bp upstream and downstream of the canonical polyadenylation signal 6-mer, AATAAA (Figure 1H) (Nielsen 2008). Differences in nucleosome-binding affinity have been reported for distinct genomic sequences, and in particular, poly(dA:dT) stretches have low nucleosome affinity as a result of their resistance to curvature (Peckham et al. 2007; Satchwell et al. 1986; Drew and Travers 1985). Nucleosome density plots centered at control AATAAA 6-mers in intergenic regions supported the idea that this 6-mer by itself has a nucleosome-positioning effect, with a dip in nucleosome density observed at the AATAAA sequence flanked by increased nucleosome density 100 bp upstream and downstream (Figure 1H). Controls based on other common variants of the poly(A) signal 6-mer yielded similar patterns (data not shown). However, authentic PAS differed from the controls in that the reduction in nucleosome density near the 6-mer was much stronger – stronger even than the "nucleosome-free" region observed near the TSS (Figure 1G) – and differed from the TSS distribution in that clear phasing of adjacent nucleosomes was not observed. These differences may result in part from additional sequence effects of the U-rich downstream sequence element (DSE) and/or other regulatory elements of cleavage and polyadenylation (Hu et al. 2005). Alternatively, it is conceivable that the differences could result from the presence of nucleosome-excluding DNA-binding proteins if such factors commonly bound near the PAS. Both high- and low-expressed genes exhibited pronounced nucleosome depletion near the PAS, with only moderately weaker depletion in inactive genes (Nielsen 2008), suggesting that the primary mechanisms responsible for PAS-associated nucleosome depletion are not dependent on expression.

Higher Downstream Nucleosome Affinity Is Associated with Higher PAS Usage Several thousand human genes express mRNAs with multiple distinct $3'$ untranslated regions (UTRs) through regulated usage of "tandem PAS," i.e., distinct PAS located at some distance apart without intervening splicing (Wang et al. 2008). To investigate the possibility that PAS recognition and nucleosome positioning might be functionally related, the individual PAS in such pairs were designated as high usage or low usage based on available transcript data (Figure S9). Strikingly, high-usage sites displayed a significantly stronger reduction in nucleosome density immediately surrounding the PAS, and stronger nucleosome enrichment from approximately +75 to +375 downstream of the PAS ($p < 10^{-10}$ and $p < 10^{-7}$, respectively; Figure 4A). These differences were evident even after controlling for the strength of core poly(A) sequence elements that function at the RNA level. To assess the potential contributions of intrinsic nucleosome affinity to the observed biases in nucleosome positioning relative to alternative explanations such as chromatin remodeling, a sequence-based model of nucleosome affinity was developed (Experimental Procedures). This model yielded a distribution of nucleosome affinity scores (NASs) that qualitatively matched the observed distribution of nucleosome density around TSSs (Figure S10). When applied to regions around tandem PAS, this model predicted a somewhat more pronounced dip in nucleosome affinity around high-usage PAS than around low-usage sites, and significantly stronger nucleosome affinity downstream of high-usage than low-usage PAS ($p < 10^{-23}$; Figure 4B). These observations, matching the ChIP-Seq data in both aspects, indicated that sequences surrounding high-usage PAS differ from those near low-usage PAS in their inherent nucleosome affinity.

2.3 Discussion

Here we have shown that the major sites of pre-mRNA processing in human genes, including both exons and the PAS, differ substantially from background levels of nucleosome density. Furthermore, more highly used alternative PAS had both higher downstream nucleosome density and higher intrinsic nucleosome affinity than less highly used alternative sites. These differences suggest that nucleosome positioning might directly influence PAS usage, e.g., through effects on the kinetics of polymerase elongation in the vicinity of the PAS, or mediated through interactions between nucleosome-associated proteins and the cleavage and polyadenylation machinery, components of which are associated with Pol II (Nag et al. 2007). This possibility could be tested by inserting well-characterized nucleosome-positioning elements near PAS and assessing the effects on PAS activity. It is also possible that sequence elements not included in standard core PAS scoring influence both PAS usage and nucleosome affinity. The largely expression-independent depletion of nucleosomes near the PAS does not support the alternative interpretation that components of the cleavage and polyadenylation machinery commonly alter nucleosome positions. The density of histone mark data was too low in the vicinity of the PAS to be informative about whether or not histones near sites of cleavage and poladenylation exhibit a distinctive modification signature, but the biased distribution of histone marks observed on exons motivates investigation of this possibility. In any event, these data indicate that differences in empirical nucleosome density and/or in NAS have significant potential to predict PAS usage and alternative $3'$UTR expression.

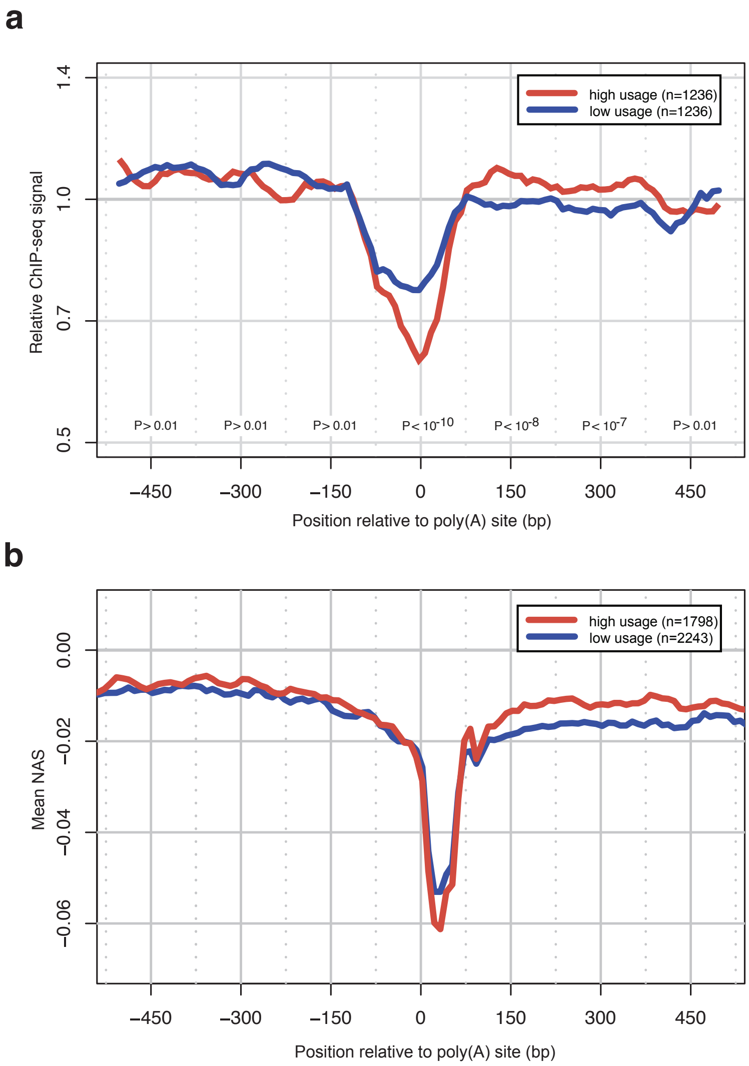


Figure 4: Nucleosome Depletion and Downstream Nucleosome Enrichment at High-Usage PAS

(A) Mean nucleosome density around human PAS of low (blue) or high (red) usage, normalized to average ChIP signal.

(B) Mean NAS for positions around human PAS of low or high usage. Wilcoxon rank sum test p values shown for 150 bp windows centered on indicated positions.

Enrichment of nucleosomes and specific histone marks on exons has been noted in three papers published very recently (Andersson et al. 2009; Tilgner et al. 2009; Schwartz et al. 2009). While some of these works noted that the sequence composition of exons is biased in a direction that tends to favor nucleosome occupancy, our analysis of ECRs in introns and intergenic regions demonstrates not only that exonic composition favors nucleosome occupancy but that the biases in oligonucleotide content of exons are sufficient to account for the magnitude of nucleosomal enrichment observed on exons (Figure 1). The importance of this finding is that it supports models in which the biased DNA sequence composition positions nucleosomes on exons (where they could potentially modulate splicing activity) independently of transcription or RNA processing. However, this observation does not preclude the existence of additional nucleosome-positioning constraints for subsets of exons, particularly those exons with weak splice sites or long flanking introns.

Schwartz et al. (2009) and Tilgner et al. (2009) suggest that the previously discovered H3K36me3 enrichment on exons (Kolasinska-Zwierz et al. 2009) might be explained by preferential exonic nucleosome positioning. Here we observed that exonic enrichment of H3K36me3, as well as H3K27me2, significantly exceeds the global enrichment of nucleosomes on exons. It will be interesting to repeat these analyses once additional high-throughput sequencing data become available, as additional trends may emerge once these comparisons can confidently be performed on an individual exon basis.

Here splicing simulation algorithms were used to demonstrate that empirical nucleosome density significantly improves the accuracy of exon identification. The increase in accuracy was observed when scoring only splice site sequences. But, interestingly, the increase was also observed when known ESE, ESS, and ISE sequences were scored as well. This observation thus provides direct evidence that nucleosome positioning contains information not present in known cis-acting RNA elements involved in splicing.

Two types of models (not mutually exclusive) could plausibly account for the observed improvements in splicing simulation resulting from nucleosome scoring. First, the set of exonic motifs that have ESE activity at the RNA level might (coincidentally) also have high inherent nucleosome affinity at the DNA level. Under this scenario, in order to account for the improvement in accuracy observed relative to splicing models that include scoring of known ESEs, the set of ESE sequences with high nucleosome affinities would need to include a number of ESEs that have not been previously described. Second, nucleosomes might directly influence splicing, e.g., mediated through effects of nucleosomes on the kinetics of Pol II transcription or through interactions between nucleosome-associated or nucleosome-modifying proteins on the one hand and RNA splicing factors on the other (Moore and Proudfoot 2009). This possibility could be tested through assessment of effects on splicing following manipulation of nucleosome positions in the vicinity of exons.

Tilgner and coworkers noted that exons with strong splice sites show the least nucleosome enrichment (Tilgner et al. 2009). Our results show that this inverse relationship persists even after controlling for exonic composition biases (as well as other factors; see the Experimental Procedures), suggesting the existence of additional influences on nucleosome positions. Several interesting possibilities could explain this result. First, intronic sequences may exist which help modulate nucleosome density in the region of exons, particularly those with weak splice sites. Second, sequences not fully captured in our Markov model of exonic nucleotide content might serve to recruit chromatin remodeling factors. The SWI/SNF chromatin remodeling complex has been reported to regulate alternative splicing (Batsché et al. 2006), although the fact that its chromatin remodeling activity appears dispensable for this regulation complicates discussion of a potential role in marking exons with weak splice sites.

Nucleosome density was inversely correlated not only with splice site strength but also with proximity to neighboring exons. This is the sort of pattern that would be expected if nucleosome occupancy enhanced exon recognition in splicing. Most vertebrate exons are recognized by exon definition, involving recognition of pairs of splice sites across exons, a mechanism that is favored by the presence of long introns (Robberson et al. 1990). Thus, nucleosome enrichment on exons might specifically facilitate recognition of exons by exon-definition mechanisms, perhaps by influencing the activity of SR proteins associated with the CTD of Pol II (Das et al. 2007). Because splicing can occur independently of transcription in vitro, chromatin is clearly not essential for splicing. However, transcription-coupled splicing occurs far more efficiently (Das et al. 2006), and our results suggest the chromatin structure itself may contribute to these differences.

Previous research has shown that specific changes at the chromatin level can locally affect splicing factor recruitment (Loomis et al. 2009) and splicing regulation (Alló et al. 2009; Tyagi et al. 2009; Schor et al. 2009; Batsché et al. 2006). Beyond connections to nucleosome positioning, several histone modifications were observed to differ from background nucleosome levels in exons, raising the intriguing possibility that these or other modifications directly or indirectly regulate splicing on a global scale. The depth of ChIP-Seq data presently available for individual histone marks did not seem sufficient to rigorously test potential contributions of these marks to splicing by splicing simulation analyses; this issue could be explored through manipulations of histones or histone-modifying enzymes. Histone modifications represent a reversible but stable form of chemical marking that could potentially be used either to enhance the fidelity of splicing or to toggle between distinct patterns of alternative splicing, e.g., in a program of cellular differentiation. Involvement of a long-lasting mark such as histone modification in splicing control could help in situations where long-term maintenance of expression of a specific alternative isoform might be desirable, e.g., in the context of immune memory or definition of cellular identity (Wojtowicz et al. 2007).

Because chromatin structure impacts mutation rates, the biased distribution of nucleosomes relative to exons has important evolutionary implications. Recently, a pattern has been observed in which positions with higher nucleosome occupancy had higher rates of substitutions but lower rates of insertions and deletions than adjacent positions with lower nucleosome density in the Japanese killifish (Sasaki et al. 2009). Thus, the association of nucleosomes with exons is expected to exert a protective effect on coding regions, lowering the rate of potentially reading frame-disrupting insertions/deletions relative to less disruptive substitution mutations.

2.4 Experimental Procedures

ChIP-Seq Data Sets We analyzed two previously published ChIP-Seq data sets: histone methylation marks in human T cells (Barski et al. 2007) and nucleosome-positioning data in human T cells (Schones et al. 2008). We chose only internal exons and ensured flanking introns were at least 500 bp long. For a given genomic position, we calculated the read coverage as the number of reads mapping upstream (on the + strand) at −73 bp and downstream (on the − strand) at +73 bp, corresponding to average nucleosome dyad positions. Because reads were only mapped to unique positions in the genome, we computed densities as a ratio of reads per unique genomic position. To reduce the impact of potential PCR amplification biases, the read count for any specific read sequence was truncated at 10. Nucleosome density was smoothed using a sliding window of size 25 or 50 bp.

Exon Analyses Certain nucleotides were overrepresented in the first few bases at the $5'$ ends of sequencing reads (see Figure S1). These biases are likely to result primarily from aspects of MNase digestion (Johnson et al. 2006) or other technical factors. To control for this technical bias, read counts were normalized as follows. Overrepresentation of each $5'$ pentamer in a library was estimated as the ratio of the number of occurrences at the $5'$ ends of all reads to the average number of occurrences of that pentamer at positions 15–25 downstream, and read counts were normalized accordingly. This control moderated the sharp peaks at the $3'$ss and $5'$ss but had little overall effect on the nucleosome densities around exons. Results were largely unchanged when reads that mapped to the most homogeneous positions around $5'$ and $3'$ splice sites, including the conserved $5'$ splice site GT and $3'$ splice site AG dinucleotides (Figure 2), were removed.

ECRs were derived from intronic regions lacking exons within 2 kb or intergenic regions with no cDNA/EST coverage that were at least 1 kb from the nearest gene annotation. We randomly chose 20,000 5′ss and 20,000 3′ss and matched 9-mer sequences from those splice sites to intronic or intergenic regions to define decoy sites. To define pseudoexons based on nucleotide content, we generated a fifth-order Markov model to score exonic versus intronic sequence composition and identified intergenic regions that closely matched authentic exons in length and 6-mer composition. ECRs defined based on exonic 5-mer, 4-mer, 3-mer, or even 2-mer composition exhibited peaks of nucleosome density qualitatively similar to those observed in Figure 1E (data not shown). This observation indicates that the nucleotide and dinucleotide content of exons is sufficient to explain, at least qualitatively, why nucleosomes are biased toward exons. The distribution of histone marks in exons was explored by comparing read densities within exons to read densities in the flanking introns. The entire exon was included as well as the most proximal 10 bp of the upstream and downstream introns. Intronic densities were calculated from the regions 200–300 bp upstream of the 3′ss and 200–300 bp downstream of the 5′ss. The choice of both upstream and downstream intronic regions helped control for changes in density of some histone marks along the length of transcripts. Average read densities were determined as the total read counts across 69,000 exons divided by the total number of unique positions. Confidence intervals and p values were produced by bootstrap sampling. In Figure 2B, genes were ranked by microarray expression signal (resting T cell data from Schones and coworkers). The top 10% and bottom 10% were defined as highly and lowly expressed genes, respectively.

Internal exons with both flanking introns of size between 500 and 1000 bp were defined as clustered exons, and those with both flanking introns of size at least 5 kb were defined as isolated exons. We analyzed a like number of isolated and clustered exons, sampled to match exon length between the two sets.

Splice site strength was scored using the maximum entropy-based log-odds scoring method (Yeo and Burge 2004), and all 5′ss and 3′ss scores were required to be nonnegative. Exons were divided into five equally sized bins based on 3′ss score, and exons were sampled from each bin to match flanking intron size and average exonic nucleotide composition. Exon:intron ratios, confidence intervals, and p values were calculated as in Figure 2.

Alternative Splicing Analysis Nearly 600 sets of adjacent exon triples were identified, where the first and third exons are constitutively spliced and the middle exon is skipped in a subset of ESTs. Similar to Kolasinska-Zwierz and coworkers (Kolasinska-Zwierz et al. 2009), we calculated read densities for the central skipped exons, normalized to the read densities of the adjacent constitutive exons. No significant difference was observed for any histone mark when compared to a like number of control triples, each consisting of three adjacent, constitutively spliced exons, matched for length and oligonucleotide composition (data not shown).

Splicing Simulation Analyses The ExonScan algorithm (Wang et al. 2004) was modified to perform exon predictions using nucleosome density information. A training set of 1000 randomly chosen genes was used to estimate log-odds scores distinguishing correctly and incorrectly predicted exons based on their nucleosome densities. This model was then applied to a set of 12,585 known genes with no evidence of alternative splicing or alternative overlapping transcripts in RefSeq (Pruitt et al. 2005). As there is a significant amount of noise in the exonic nucleosome read counts (because of their small average size), the nucleosome-scoring model was applied only to exons with at least 50 mappable genomic positions. To estimate the significance of improvements in exon prediction based on nucleosome densities, we compared accuracy when nucleosome density was scored normally to simulations in which the scores of nucleosome density in random genic regions of the same length were assigned to exons. Receiver operator characteristics were calculated using the R package ROCR (http://www.r-project.org/).

Poly(A) Analyses Genome-wide sequence alignments of available cDNAs and ESTs were obtained from the University of Santa Cruz Genome Browser Database. Uniquely mapping cDNAs and ESTs were filtered for evidence of a nongenomically derived poly(A) tail and a canonical or variant poly(A) signal (Beaudoing and Gautheret 2001) in the −1 to −40 region upstream of the aligned poly(A) site (Figure S9). The resulting set was then mapped to a comprehensive and nonredundant set of RefSeq transcripts (Pruitt et al. 2005) and clustered to create a database of polyadenylation sites. Sites with usage were defined as those supported by greater than 70% of the gene's mapped polyadenylated ESTs, whereas low-usage sites were defined as those having less than 30% of the supporting ESTs.

Weight matrix models of core poly(A) motifs described by Hu and coworkers (Hu et al. 2005) were obtained as a part of their PolyA_SVM distribution, http://exon.umdnj.edu/polya_svm/. The output of polya_svm.pl run in matching-element mode was parsed to obtain scores for each poly(A) cis-element. The core poly(A) motif score was then reported as the sum of the score for the CUE2 element, corresponding to the poly(A) signal, and the average score for the CDE1-CDE4 elements, corresponding to the U-rich downstream signals.

Nucleosome Affinity Scores A total of 84 million Illumina read starts, representing 75% of the perfectly and uniquely mapping reads in the Barski and coworkers (Barski et al. 2007) data set, were chosen at random for the nucleosome training set. An equally sized background set was obtained by randomly sampling a position within 500 bp of each of the read starts in the nucleosome training set (excluding sites mapped by other read starts in the nucleosome training set). Using these data, a fifth-order Markov model was trained for every position n in the nucleosome-occupied region (or control region), such that we obtained $P(X_n = x|X_{n-1} = x_{n-1}, \ldots, X_{n-5} = x_{n-5})$ for every $n = 1, \ldots, 146$, and for every combination of $x, x_{n-1}, x_{n-2}, \ldots =$ A, C, G, T. Due to the aforementioned $5'$ nucleotide bias in the sequencing data, positions 115 were subsequently excluded from the model. NASs were calculated as the $\log_2$ ratio of P(seq|nucleosome model) to P(seq|background model). Scores were plotted at a 73 bp offset to reflect the center of the corresponding nucleosome.

2.5 References

Alló, M, V Buggiano, JP Fededa, E Petrillo, I Schor, M de la Mata, E Agirre, M Plass, E Eyras, SA Elela, R Klinck, B Chabot, and AR Kornblihtt (July 2009). "Control of alternative splicing through siRNA-mediated transcriptional gene silencing." In: *Nat Struct Mol Biol* 16.7, pp. 717–24. DOI: 10.1038/nsmb.1620 (cit. on p. 43).

Andersson, R, S Enroth, A Rada-Iglesias, C Wadelius, and J Komorowski (Oct. 2009). "Nucleosomes are well positioned in exons and carry characteristic histone modifications." In: *Genome Res* 19.10, pp. 1732–41. DOI: 10.1101/gr.092353.109 (cit. on p. 41).

Baldi, P, S Brunak, Y Chauvin, and A Krogh (Nov. 1996). "Naturally occurring nucleosome positioning signals in human exons and introns." In: *J Mol Biol* 263.4, pp. 503–10. DOI: 10.1006/jmbi.1996.0592 (cit. on p. 34).

Barski, A, S Cuddapah, K Cui, TY Roh, DE Schones, Z Wang, G Wei, I Chepelev, and K Zhao (May 2007). "High-resolution profiling of histone methylations in the human genome." In: *Cell* 129.4, pp. 823–37. DOI: 10.1016/j.cell.2007.05.009 (cit. on pp. 34, 35, 43, 45).

Batsché, E, M Yaniv, and C Muchardt (Jan. 2006). "The human SWI/SNF subunit Brm is a regulator of alternative splicing." In: *Nat Struct Mol Biol* 13.1, pp. 22–9. DOI: 10.1038/nsmb1030 (cit. on pp. 42, 43).

Beaudoing, E and D Gautheret (Sept. 2001). "Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data." In: *Genome Res* 11.9, pp. 1520–6. DOI: 10.1101/gr.190501 (cit. on p. 45).

Burge, C and S Karlin (Apr. 1997). "Prediction of complete gene structures in human genomic DNA." In: *J Mol Biol* 268.1, pp. 78–94. DOI: 10.1006/jmbi.1997.0951 (cit. on p. 34).

Das, R, K Dufu, B Romney, M Feldt, M Elenko, and R Reed (May 2006). "Functional coupling of RNAP II transcription to spliceosome assembly." In: *Genes Dev* 20.9, pp. 1100–9. DOI: 10.1101/gad.1397406 (cit. on p. 42).

Das, R, J Yu, Z Zhang, MP Gygi, AR Krainer, SP Gygi, and R Reed (June 2007). "SR proteins function in coupling RNAP II transcription to pre-mRNA splicing." In: *Mol Cell* 26.6, pp. 867–81. DOI: 10.1016/j.molcel.2007.05.036 (cit. on p. 42).

de la Mata, M, CR Alonso, S Kadener, JP Fededa, M Blaustein, F Pelisch, P Cramer, D Bentley, and AR Kornblihtt (Aug. 2003). "A slow RNA polymerase II affects alternative splicing in vivo." In: *Mol Cell* 12.2, pp. 525–32 (cit. on p. 33).

Denisov, DA, ES Shpigelman, and EN Trifonov (Dec. 1997). "Protective nucleosome centering at splice sites as suggested by sequence-directed mapping of the nucleosomes." In: *Gene* 205.1-2, pp. 145–9 (cit. on p. 34).

Drew, HR and AA Travers (Dec. 1985). "DNA bending and its relation to nucleosome positioning." In: *J Mol Biol* 186.4, pp. 773–90 (cit. on p. 39).

Fairbrother, WG, F Yeh, PA Sharp, and CB Burge (Aug. 2002). "Predictive identification of exonic splicing enhancers in human genes." In: *Science* 297.5583, pp. 1007–13. DOI: 10.1126/science.1073774 (cit. on p. 38).

Görnemann, J, KM Kotovic, K Hujer, and KM Neugebauer (July 2005). "Cotranscriptional spliceosome assembly occurs in a stepwise fashion and requires the cap binding complex." In: *Mol Cell* 19.1, pp. 53–63. DOI: 10.1016/j.molcel.2005.05.007 (cit. on p. 33).

Howe, KJ, CM Kane, and M Ares (Aug. 2003). "Perturbation of transcription elongation influences the fidelity of internal exon inclusion in Saccharomyces cerevisiae." In: *RNA* 9.8, pp. 993–1006 (cit. on p. 33).

Hu, J, CS Lutz, J Wilusz, and B Tian (Oct. 2005). "Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation." In: *RNA* 11.10, pp. 1485–93. DOI: 10.1261/rna.2107305 (cit. on pp. 39, 45).

Johnson, SM, FJ Tan, HL McCullough, DP Riordan, and AZ Fire (Dec. 2006). "Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin." In: *Genome Res* 16.12, pp. 1505–16. DOI: 10.1101/gr.5560806 (cit. on p. 43).

Kaplan, N, IK Moore, Y Fondufe-Mittendorf, AJ Gossett, D Tillo, Y Field, EM LeProust, TR Hughes, JD Lieb, J Widom, and E Segal (Mar. 2009). "The DNA-encoded nucleosome organization of a eukaryotic genome." In: *Nature* 458.7236, pp. 362–6. DOI: 10.1038/nature07667 (cit. on p. 34).

Kolasinska-Zwierz, P, T Down, I Latorre, T Liu, XS Liu, and J Ahringer (Mar. 2009). "Differential chromatin marking of introns and expressed exons by H3K36me3." In: *Nat Genet* 41.3, pp. 376–81. DOI: 10.1038/ng.322 (cit. on pp. 33, 35, 38, 42, 44).

Kornblihtt, AR, M de la Mata, JP Fededa, MJ Munoz, and G Nogues (Oct. 2004). "Multiple links between transcription and splicing." In: *RNA* 10.10, pp. 1489–98. DOI: 10.1261/rna.7100104 (cit. on p. 33).

Lian, Z, A Karpikov, J Lian, MC Mahajan, S Hartman, M Gerstein, M Snyder, and SM Weissman (Aug. 2008). "A genomic analysis of RNA polymerase II modification and chromatin architecture related to 3' end RNA polyadenylation." In: *Genome Res* 18.8, pp. 1224–37. DOI: 10.1101/gr.075804.107 (cit. on p. 39).

Li, B, M Carey, and JL Workman (Feb. 2007). "The role of chromatin during transcription." In: *Cell* 128.4, pp. 707–19. DOI: 10.1016/j.cell.2007.01.015 (cit. on p. 33).

Lim, LP and CB Burge (Sept. 2001). "A computational analysis of sequence features involved in recognition of short introns." In: *Proc Natl Acad Sci USA* 98.20, pp. 11193–8. DOI: 10.1073/pnas.201407298 (cit. on p. 37).

Lin, S, G Coutinho-Mansfield, D Wang, S Pandit, and XD Fu (Aug. 2008). "The splicing factor SC35 has an active role in transcriptional elongation." In: *Nat Struct Mol Biol* 15.8, pp. 819–26. DOI: 10.1038/nsmb.1461 (cit. on p. 33).

Loomis, RJ, Y Naoe, JB Parker, V Savic, MR Bozovsky, T Macfarlan, JL Manley, and D Chakravarti (Feb. 2009). "Chromatin binding of SRp20 and ASF/SF2 and dissociation from mitotic chromosomes is modulated by histone H3 serine 10 phosphorylation." In: *Mol Cell* 33.4, pp. 450–61. DOI: 10.1016/j.molcel.2009.02.003 (cit. on pp. 33, 43).

Mavrich, T, IP Ioshikhes, BJ Venters, C Jiang, LP Tomsho, J Qi, SC Schuster, I Albert, and BF Pugh (July 2008). "A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome." In: *Genome Res* 18.7, pp. 1073–83. DOI: 10.1101/gr.078261.108 (cit. on p. 39).

Moore, MJ and NJ Proudfoot (Feb. 2009). "Pre-mRNA processing reaches back to transcription and ahead to translation." In: *Cell* 136.4, pp. 688–700. DOI: 10.1016/j.cell.2009.02.001 (cit. on pp. 33, 42).

Murray, JI, RB Voelker, KL Henscheid, MB Warf, and JA Berglund (Jan. 2008). "Identification of motifs that function in the splicing of non-canonical introns." In: *Genome Biol* 9.6, R97. DOI: 10.1186/gb-2008-9-6-r97 (cit. on p. 38).

Nag, A, K Narsinh, and HG Martinson (July 2007). "The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase." In: *Nat Struct Mol Biol* 14.7, pp. 662–9. DOI: 10.1038/nsmb1253 (cit. on p. 40).

Nielsen, C (2008). "Mammalian gene regulation through the 3′ UTR." PhD thesis. Massachusetts Institute of Technology (cit. on pp. 39, 40).

Peckham, HE, RE Thurman, Y Fu, JA Stamatoyannopoulos, WS Noble, K Struhl, and Z Weng (Aug. 2007). "Nucleosome positioning signals in genomic DNA." In: *Genome Res* 17.8, pp. 1170–7. DOI: 10.1101/gr.6101007 (cit. on p. 39).

Peterson, ML, S Bertolino, and F Davis (Aug. 2002). "An RNA polymerase pause site is associated with the immunoglobulin mus poly(A) site." In: *Mol Cell Biol* 22.15, pp. 5606–15 (cit. on p. 33).

Pruitt, KD, T Tatusova, and DR Maglott (Jan. 2005). "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." In: *Nucleic Acids Res* 33.Database issue, pp. D501–4. DOI: 10.1093/nar/gki025 (cit. on pp. 44, 45).

Robberson, BL, GJ Cote, and SM Berget (Jan. 1990). "Exon definition may facilitate splice site selection in RNAs with multiple exons." In: *Mol Cell Biol* 10.1, pp. 84–94 (cit. on p. 42).

Sasaki, S, CC Mello, A Shimada, Y Nakatani, SI Hashimoto, M Ogawa, K Matsushima, SG Gu, M Kasahara, B Ahsan, A Sasaki, T Saito, Y Suzuki, S Sugano, Y Kohara, H Takeda, A Fire, and S Morishita (Jan. 2009). "Chromatin-associated periodicity in genetic variation downstream of transcriptional start sites." In: *Science* 323.5912, pp. 401–4. DOI: 10.1126/science.1163183 (cit. on pp. 34, 43).

Satchwell, SC, HR Drew, and AA Travers (Oct. 1986). "Sequence periodicities in chicken nucleosome core DNA." In: *J Mol Biol* 191.4, pp. 659–75 (cit. on p. 39).

Schones, DE, K Cui, S Cuddapah, TY Roh, A Barski, Z Wang, G Wei, and K Zhao (Mar. 2008). "Dynamic regulation of nucleosome positioning in the human genome." In: *Cell* 132.5, pp. 887–98. DOI: 10.1016/j.cell.2008.02.022 (cit. on pp. 34, 39, 43).

Schor, IE, N Rascovan, F Pelisch, M Alló, and AR Kornblihtt (Mar. 2009). "Neuronal cell depolarization induces intragenic chromatin modifications affecting NCAM alternative splicing." In: *Proc Natl Acad Sci USA* 106.11, pp. 4325–30. DOI: 10.1073/pnas.0810666106 (cit. on pp. 33, 43).

Schwartz, S, E Meshorer, and G Ast (Sept. 2009). "Chromatin organization marks exon-intron structure." In: *Nat Struct Mol Biol* 16.9, pp. 990–5. DOI: 10.1038/nsmb.1659 (cit. on pp. 41, 42).

Sims, RJ, S Millhouse, CF Chen, BA Lewis, H Erdjument-Bromage, P Tempst, JL Manley, and D Reinberg (Nov. 2007). "Recognition of trimethylated histone H3 lysine 4 facilitates the recruitment of transcription postinitiation factors and pre-mRNA splicing." In: *Mol Cell* 28.4, pp. 665–76. DOI: 10.1016/j.molcel.2007.11.010 (cit. on pp. 33, 34, 38).

Tilgner, H, C Nikolaou, S Althammer, M Sammeth, M Beato, J Valcárcel, and R Guigó (Sept. 2009). "Nucleosome positioning as a determinant of exon recognition." In: *Nat Struct Mol Biol* 16.9, pp. 996–1001. DOI: 10.1038/nsmb.1658 (cit. on pp. 41, 42).

Tyagi, A, J Ryme, D Brodin, AKO Farrants, and N Visa (May 2009). "SWI/SNF associates with nascent pre-mRNPs and regulates alternative pre-mRNA processing." In: *PLoS Genet* 5.5, e1000470. DOI: 10.1371/journal.pgen.1000470 (cit. on p. 43).

Wang, ET, R Sandberg, S Luo, I Khrebtukova, L Zhang, C Mayr, SF Kingsmore, GP Schroth, and CB Burge (Nov. 2008). "Alternative isoform regulation in human tissue transcriptomes." In: *Nature* 456.7221, pp. 470–6. DOI: 10.1038/nature07509 (cit. on p. 40).

Wang, Z, ME Rolish, G Yeo, V Tung, M Mawson, and CB Burge (Dec. 2004). “Systematic identification and analysis of exonic splicing silencers.” In: *Cell* 119.6, pp. 831–45. DOI: 10.1016/j.cell.2004.11.010 (cit. on pp. 33, 37, 39, 44).

Wojtowicz, WM, W Wu, I Andre, B Qian, D Baker, and SL Zipursky (Sept. 2007). “A vast repertoire of Dscam binding specificities arises from modular interactions of variable Ig domains.” In: *Cell* 130.6, pp. 1134–45. DOI: 10.1016/j.cell.2007.08.026 (cit. on p. 43).

Xiao, X, Z Wang, M Jang, R Nutiu, ET Wang, and CB Burge (Oct. 2009). “Splice site strength-dependent activity and genetic buffering by poly-G runs.” In: *Nat Struct Mol Biol* 16.10, pp. 1094–100. DOI: 10.1038/nsmb.1661 (cit. on p. 38).

Yeo, G and CB Burge (Jan. 2004). “Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals.” In: *J Comput Biol* 11.2-3, pp. 377–94. DOI: 10.1089/1066527041410418 (cit. on p. 44).

Chapter 3

Global regulation of mRNA stability through alternative cleavage and polyadenylation

Abstract

While coordinated regulation of alternative cleavage and polyadenylation is associated with cellular proliferation rate, cellular differentiation and oncogenic transformation, little is known about the genome-wide consequences of regulated $3'$ end formation. We produced isoform-specific mRNA half-life estimates using transcriptional shut-off followed by high-throughput sequencing of $3'$ ends using 3P-Seq. Proximal tandem $3'$ untranslated region (UTR) isoforms were significantly more stable than distal isoforms, in part because of the destabilizing effects of microRNAs and PUF-binding proteins. $3'$ UTR sequence conservation was highest for the least stable genes, supporting a model where destabilizing elements play a key biological role in post-transcriptional regulation of mRNA abundance.

Contents

Acknowledgments I could not have completed this work without the guidance in wet lab experimentation provided by lab mates in both the Bartel and Burge labs. In particular, I am grateful to Calvin Jan for teaching me the ins and outs of his 3P-Seq method; I've analyzed several of his unpublished libraries in the course of this project. I'd also like to thank Vincent Butty for sharing his RNA-Seq data with me.

3.1 Introduction

Post-transcriptional gene regulation is important in determining the pool of messenger RNA available for translation into protein. Most of this regulation is likely to be mediated through cis motifs embedded within the mRNA, and hence modulation of the mature mRNA sequence through alternative splicing or alternative cleavage and polyadenylation are key to understanding post-transcriptional gene regulation. Mature mRNAs are composed of three DNA-encoded parts, the $5'$ untranslated region ($5'$ UTR), the open reading frame (ORF), and the $3'$ UTR. The $5'$ UTR tends to be short because the ribosome must scan its length before initiating translation, and the open reading frame is tightly constrained by the amino acid sequence it must code for. In contrast, $3'$ UTRs are unrestricted in length and sequence (aside from encoding the cleavage and polyadenylation signal), and are therefore central to defining the post-transcriptional regulation of an mRNA.

MicroRNAs are an important class of post-transcriptional regulatory RNA. MicroRNAs recruit a protein silencing complex to mRNAs based on sequence complementarity to the mRNA, downregulating the targeted mRNA and inhibiting protein translation (Filipowicz et al. 2008). When microRNA target sequences occur in the ORF, they mediate a very slight downregulation of the targeted genes, in comparison to a robust effect of sites found in the $3'$ UTR (Grimson et al. 2007). In fact, even sites immediately downstream of the stop codon were reduced in effectiveness, leading to a model where the ribosome is able to displace regulatory trans-factors from within the ORF and the so-called ribosome shadow ~15 bp downstream of the stop codon (Grimson et al. 2007). This emphasizes the importance of $3'$ UTRs in post-transcriptional gene regulation, not only for microRNAs, but presumably for other less well-studied trans-factors which would also be displaced by the massive translation complex. Recent work on the AU-rich element binding protein HuR also suggests this regulatory factor binds target mRNAs outside the coding sequence (Mukherjee et al. 2011; Lebedeva et al. 2011). Most known binding sites of the PUF family of regulatory proteins are also in the $3'$ UTR (Quenault et al. 2011). Finally, cytoplasmic mRNA localization is most frequently mediated through cis regulatory elements found in the $3'$ UTR (Andreassi and Riccio 2009).

Current models of translation show a circularization of cytoplasmic mRNAs, bringing the $3'$ UTR and poly(A) tail in close physical proximity to the $5'$ end of the message (Sonenberg and Hinnebusch 2009). mRNA circularization would therefore bring $3'$ UTR-bound trans factors near to the site of translation initiation, as well as deadenylation and decapping – these latter two are steps generally required for mRNA degradation.

Given the prominent role $3'$ UTRs play in regulating gene stability, translation and localization, it is unsurprising that $3'$ UTR isoform choice is itself highly regulated. Analysis of expressed sequence tags (ESTs) suggested a majority of mammalian genes contain multiple cleavage and polyadenylation sites (Tian et al. 2005), either in "tandem UTRs" where choice of an earlier site by the cleavage and polyadenylation machinery precludes use of later sites; or in alternative last exons, where splicing patterns determine which cleavage site is used.

Sandberg et al. (2008) discovered a coordinated shift towards shorter tandem UTR isoforms upon cell proliferation. This pattern appeared to generalize across tissues: the more highly proliferative a tissue, the shorter the average genome-wide $3'$ UTR length (Sandberg et al. 2008). The correlation between $3'$ UTR length and cell proliferation was shown to hold true across mouse embryonic development (Ji et al. 2009). Additionally, oncogenic transformation also appears to sometimes favor $3'$ UTR shortening, even after taking into account cell proliferation (Mayr and Bartel 2009), although genome-wide follow-up work has suggested that transformation only results in UTR shortening in some cell lines (Shepard et al. 2011).

The functional relevance of tandem UTR shortening remains unclear. One previous study suggested genes with particularly long 3′ UTRs (> 1kb) had relatively short half-lives (Yang et al. 2003), although another higher resolution study was unable to show a significant genome-wide relationship between 3′ UTR length and stability (Sharova et al. 2009). However, these genomic surveys were limited to gene level stability measurements, and relied on published genome annotations of 3′ UTRs, leaving open the possibility that the observed effects were due to other causative attributes merely correlated with 3′ UTR length. Low-throughput work has supported the destabilizing effect of longer UTRs. Upon fusing short or long 3′ UTRs to the luciferase open reading frame, reporters showed increased protein production from the short 3′ UTR isoforms for a handful of constructs (Mayr and Bartel 2009; Sandberg et al. 2008). At least some of this downregulation is likely due to mRNA destabilization of the longer isoform, as assayed by half-life measurements performed on three genes across a handful of cell lines (Mayr and Bartel 2009). Whether shortened 3′ UTRs globally lead to upregulation of mRNA stability and protein production is still an open question.

One of these studies included an experiment to directly assay the effect of 3′ UTR length on stability, rather than the potential contribution of destabilizing cis motifs in the longer 3′ UTR. Mayr and Bartel (2009) compared the stabilities of the short and long IGF2BP1 3′ UTR isoforms to an artificial 3′ UTR with the extension region sequence reverse complemented (the rc construct). The short isoform reporter had a significantly longer half life than the long isoform across a number of cell types, but the rc reporter showed an intermediate stability. As the rc extension region was not expected to contain the same presumably repressive sequence motifs such as microRNA targets, this experiment was supportive of a direct role for 3′ UTR length in determining gene stability. However, because the short and rc constructs differed by several kb in length, the authors concluded it is likely that the rc construct does randomly include some stability-altering sequence motifs and hence it is still unknown whether 3′ UTR length is important in determining isoform stability.

A number of methods have been recently developed to annotate or quantitate cleavage and polyadenylation events globally using high-throughput sequencing (Fu et al. 2011; Shepard et al. 2011; Jan et al. 2011). Most methods involve d(T)-priming off the poly(A) tail, which can lead to significant internal priming artifacts caused by hybridization to A-rich regions in the middle of mRNAs (Jan et al. 2011). The 3P-Seq method by Jan et al. (2011) is to date the most specific high-throughput method for annotating 3′ UTRs, as it uses only ligations in the library preparation, such that when a portion of the poly(A) tail is sequenced, it is an unambiguous marker of a true poly(A) site (Figure 1A). The 3P-Seq method preferentially includes a 4–6 bp portion of the poly(A) tail at the 3′ end of its reads though partial digestion of the poly(A) tail using RNase H followed by ligation of a sequencing adapter. Those reads including a non-genomic poly(A) stretch at their 3′ end can be used for confident annotation of cleavage and polyadenylation sites.

We explored the role of 3′ UTR length in mRNA stability using 3P-Seq to globally annotate UTR isoforms in mouse fibroblast cells. Our results show that 3P-Seq can be used to quantitate UTR isoforms, and we used the actinomycin D drug and 3P-Seq to estimate isoform-level half-lives. By directly comparing the stability of tandem UTRs, we were able to exclude potential effects caused by transcription, splicing, 5′ UTRs or the coding sequence. Our results confirm that shorter tandem 3′ UTR isoforms are globally more stable than the longer isoforms. Finally, we characterized the contributions of several classes of 3′ UTR regulatory motifs to this differential stability.

3.2 Results

Reannotation of murine poly(A) sites using 3P-Seq In order to characterize cleavage and polyadenylation sites in mouse NIH 3T3 cells, we prepared 3P-Seq libraries from total RNA (Figure 1). We sequenced 25 million raw reads, of which 9.7 million reads (comprising 937,673 unique sequences) mapped to the mouse genome.

As previously published (Jan et al. 2011), 3P-Seq reads had an average of 4 A's at their 3′ ends (Figure 1B, red curve), a remnant of the incomplete RNase H digestion of the poly(A) tail. Importantly, 6.6 million reads (68%) had at least 1 non-genome-matching A at the 3′ (Figure 1B, blue curve). These untemplated A's are a key indicator that the read stemmed from a true cleavage and polyadenylation site and were not sequencing artifacts from within the body of a transcript.

Over 80% of the reads mapped to annotated 3′ UTRs (Figure 1C). While some of the reads mapping to introns (2.2% of all mapped reads; Figure 1C) may be due to sequencing of internal mRNA fragments, this percentage is relatively unchanged after filtering for reads with at least 1 untemplated 3′ A's (1.5% of filtered reads) further supporting previous research suggesting alternative cleavage and polyadenylation events occur frequently in introns (Tian et al. 2007). Reads mapping to coding exons make up 2.1% of all mapped reads whereas this fraction accounts for only 0.1% of reads with non-genomic end A's, suggesting most of these reads are background. Altogether, these results support the highly specific enrichment of 3P-Seq for reads overlapping poly(A) sites.

Using the 3T3 3P-Seq data, we were able to de novo annotate mouse poly(A) sites. Using an approach similar to Jan et al. (2011), we clustered reads within and downstream of annotated transcripts. For each read, the putative cleavage and poly(A) site was inferred by removing all A's at the 3′ end. Clusters were centered on the most highly supported poly(A) site. To ensure that clusters were true poly(A) sites, we required at least 10% of overlapping reads to contain non-genomic 3′ end A's. For the complete decision tree used to annotate poly(A) sites, see the Methods. Using this approach, we annotated 33,590 mouse poly(A) sites, an average of 2.2 per gene expressed in our control mouse 3P-Seq libraries.

Most 3P-Seq clusters mapped near RefSeq annotated transcript 3′ ends (Figure 1D), and there were more upstream of annotated sites than downstream, supporting the notion that the most highly used 3′ UTR isoforms (which are also the most likely to be annotated in RefSeq) are the last poly(A) sites in each gene (Tian et al. 2005). The mean 3′ UTR length was 1.2kb, although the distribution of UTR lengths peaks at 150bp and has a long tail (Figure 1E). Clusters up to 3.5kb downstream of the last RefSeq annotated poly(A) were presumed to belong to the annotated gene. Figure 2 shows an example of the 3P-Seq reads used to annotate poly(A) sites.

To establish that the 3P-Seq method can accurately quantitate gene expression, we compared 3P-Seq read counts to RNA-Seq data from the same cell type (Figure 1F; V. Butty and C. Burge, personal communication). There is no gold standard for global 3′ UTR isoform quantitation as RNA-Seq is for gene expression quantitation. To gain an understanding of our strength in assessing 3′ UTR isoform abundance, we quantified gene expression using only 3P-Seq reads overlapping with a 3P-Seq cluster. This allowed us to use all reads, not only those with untemplated 3′ A's (as was done in Jan et al. (2011)), but this filter should remove the vast majority of internal mRNA fragment reads. We then totaled all 3P-Seq reads within all clusters for each gene, and compared this value to the RNA-Seq RPKM, or reads per exonic kilobase per million mapped reads. The Spearman correlation of 0.77 gave us confidence to use 3P-Seq reads as a quantitative method for assessing relative isoform-level abundance.

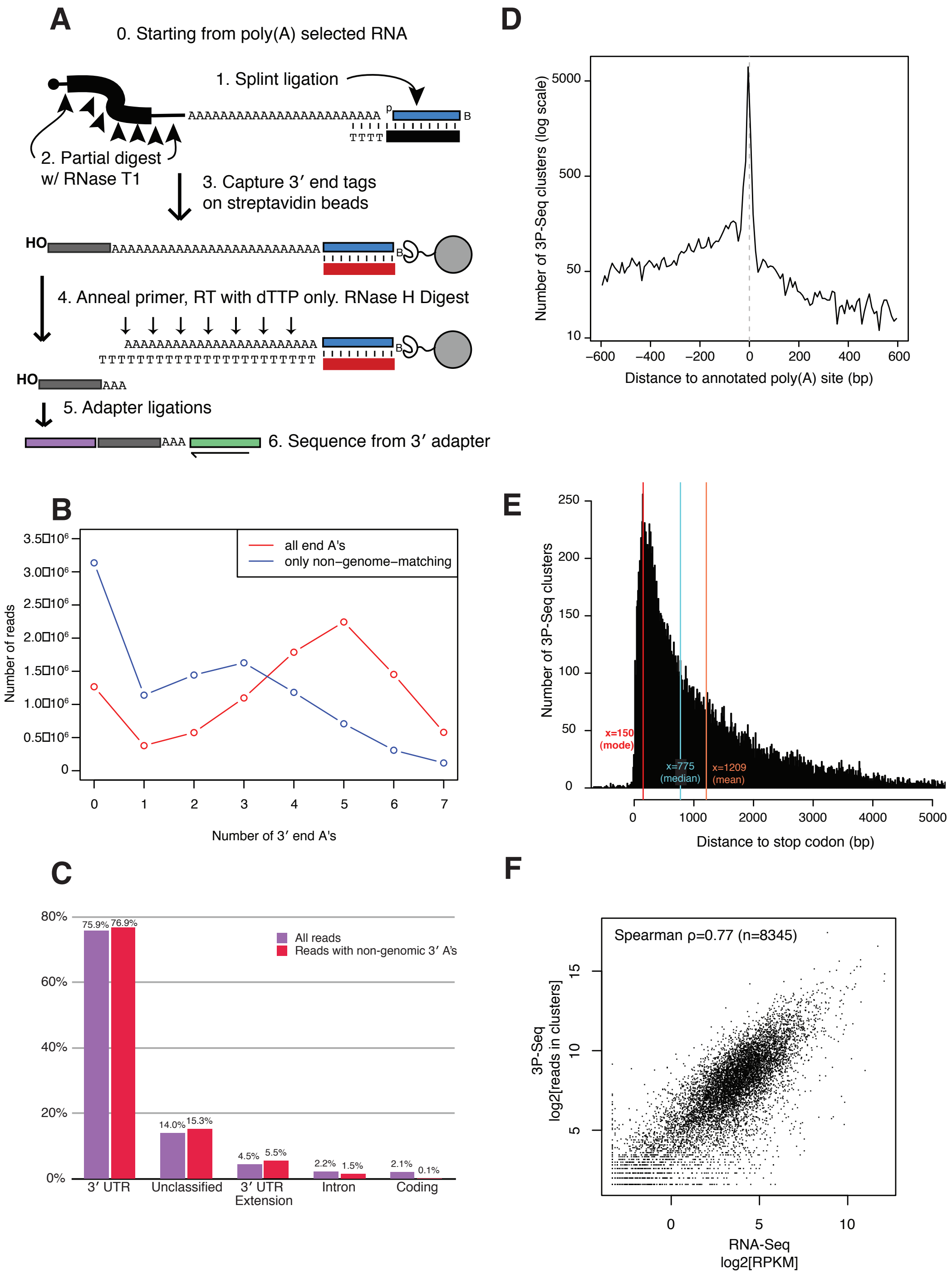


Figure 1: Overview of 3P-Seq method and results.

(A) Overview of 3P-Seq. Modified from Jan et al. (2011).

(B) The majority of 3P-Seq reads have at least one A at its 3′ end (red curve), and of these 3′ A's, many are non-genome matching (blue curve).

(C) Most reads map within, or immediately downstream of, annotated 3′ UTRs. All mapped reads (purple) were filtered based on the presence of at least one untemplated 3′ end A (red). Other categories (5′ UTR, repeat elements, etc) accounted for less than 1.5% of all reads.

(D) Most 3P-Seq clusters are close to ref-seq annotated poly(A) sites (note the y-axis is on a log scale).

(E) Distribution of 3′ UTR lengths. Negative lengths occur from premature cleavage events upstream of an annotated stop codon.

(F) 3P-Seq correlates well with RNA-Seq in mouse 3T3 cells. For 3P-Seq, only reads within clusters were used to quantify gene expression.

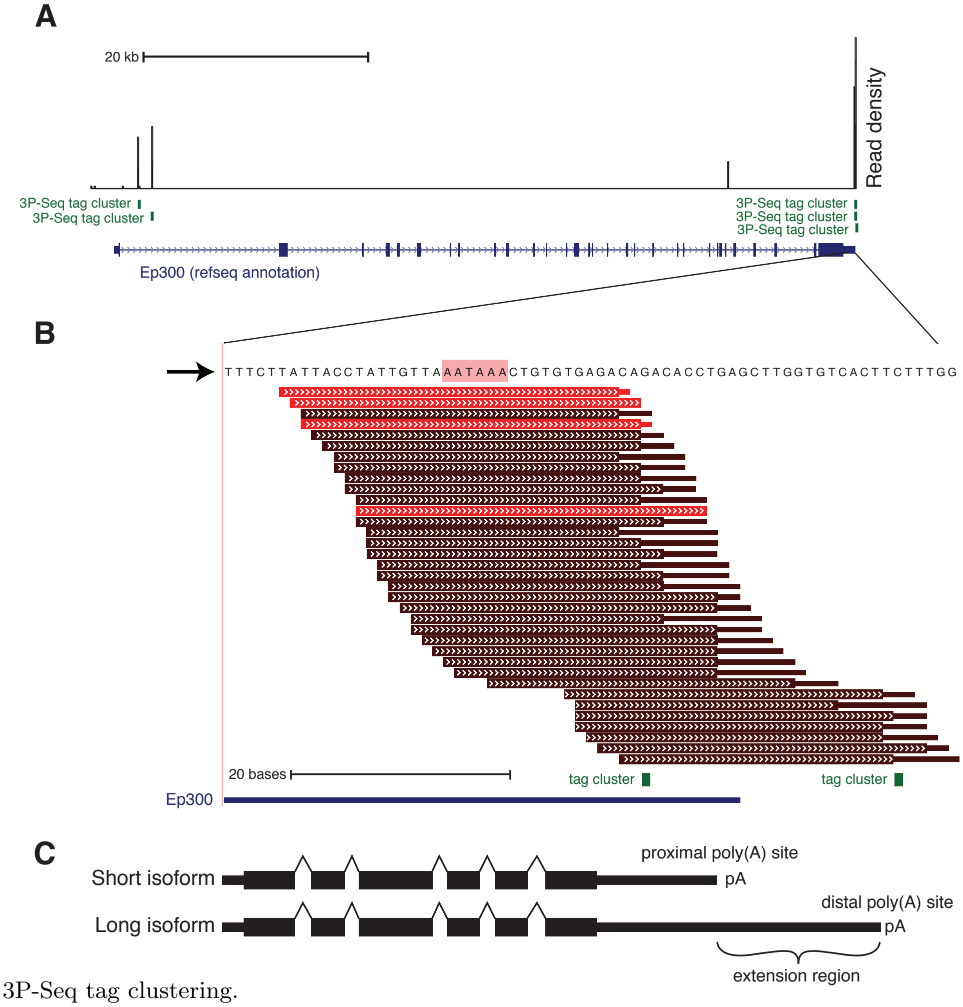


Figure 2: 3P-Seq tag clustering.

(A) 3P-Seq tag clusters for a representative gene, P300. 3P-Seq annotated poly(A) sites are green vertical hashes labeled as tag clusters. Note several premature poly(A) sites in the first intron, and three sites downstream of the stop codon.
(B) Zoomed in view of P300 3′ UTR shows reads used to call tag clusters. Thick portions of each read indicate genome-matching portion, and thin portion indicates length of 3′ end A's. The putative poly(A) site is where the thick and thin portions meet. Dark red reads contain non-genome matching 3′ end A's, and light red reads do not. There is one canonical poly(A) signal motif (highlighted in red) just upstream of the first tag cluster. When tag clusters occurred within 100bp, downstream analyses would use only the most highly expressed cluster.
(C) Diagram of tandem 3′ UTRs.

Isoform level half-life quantitation Previous research indicated that shorter 3′ UTR isoforms were more stable than longer isoforms for three human genes. To ascertain whether this trend holds genome-wide, we treated 3T3 cells for 8 hours with actinomycin D to block transcription and prepared a 3P-Seq library from these cells. The fold-change between control and actinomycin D-treated cells should be proportional to the stability of the isoform – the least stable isoforms should decrease in abundance the quickest, and therefore have the most negative log fold change. Gene level half-life estimates correlated well with those from a previously published transcription shutoff experiment quantified by micro-array, despite using a different cell type than the mouse embryonic stem cells used in the published study (Figure 4; Spearman correlation=0.68, n=6656)(Sharova et al. 2009). Because we did not have data for early time points after

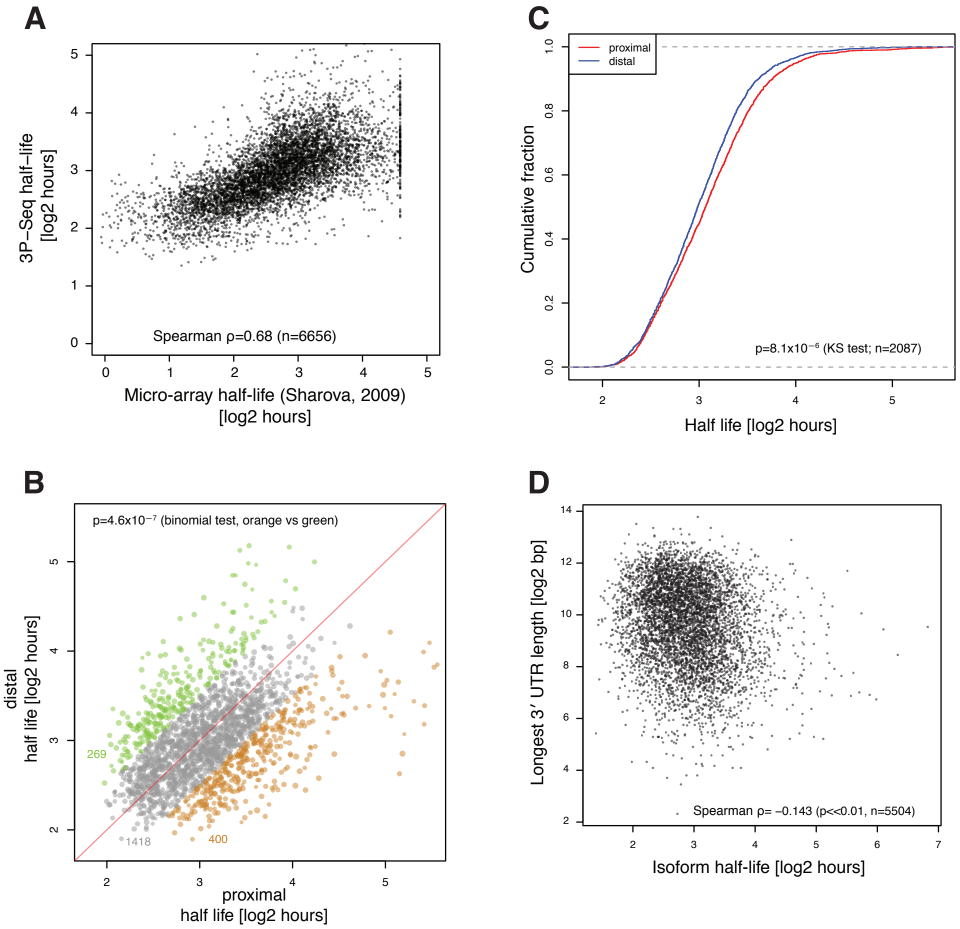


Figure 3: 3′ UTR isoform-level half-life quantitation using actinomycin D followed by 3P-Seq

(A) Comparison of gene-level half-life estimates in mouse 3T3 cells by 3P-Seq (y-axis) and in mouse ES cells by micro-array (y-axis). Values on the x-axis were truncated at 24 hours by the authors of Sharova et al. (2009) because of concerns that longer half-life estimates were inaccurate.
(B) Stability of long and short tandem 3′ UTR isoforms. Orange points show genes with the proximal isoform at least 1.4-fold more stable than the distal isoform; and conversely for the green points.
(C) Cumulative distribution of the same UTRs show in panel B.

(D) Dependence of stability on 3′ UTR length. Shown are the length and half-lives for the longest isoform for each gene.

transcription blockage, we were unable to quantify very unstable transcripts.

3′ UTR shortening increases mRNA half-life We calculated half-lives for genes with multiple tandem poly(A) sites (Figure 3B, 3C). For each such gene, we chose the top two most highly expressed isoforms separated by at least 100 bp. There was a highly significant shift towards longer half-lives for the shorter isoforms (p=4.6×10^{-7}, binomial test). This confirmed that globally, 3′ UTR shortening leads to more stable isoforms on average. However, there were a considerable number of genes for which the distal isoform was at least 1.4-fold more stable than the proximal isoform (green points, figure 3B).

We investigated the effect of UTR length on stability across genes. For each gene, we chose the distal most tag cluster to represent the final poly(A) site. We observed a very slight but highly significant negative correlation between 3′ UTR length and stability (Spearman correlation = -0.14, p$\ll 0.01$; Figure 3D). Without a complete catalog of stability-influencing motifs, it would be premature to conclude that this relationship does or does not derive from a direct effect of 3′ UTR length on stability, rather than sequence determinants. However, given the large variance in stabilities and very slight correlation between length and half-life, it seems likely that 3′ UTR length is at most a minor determinant of gene stability.

MicroRNAs contribute to destabilization of distal isoform Most previous explorations of 3′ UTR stability-related motifs have been conducted at the gene level, but may be confounded by the presence of other motifs within the 5′ UTR or ORF, as well as differences at the transcriptional or splicing level that might differentially affect stability depending on the gene. Our 3P-Seq stability experiment allowed us to isolate the effect of specific 3′ UTR regions on half-life by comparing tandem UTR isoforms. Because the only difference between the short and long isoforms is the extension region (see Figure 2C), the presence or absence of certain motifs within these extension regions can be compared to the differences in stability between the isoforms. This makes the assumption that the short and long isoforms share the same splicing patterns; for a first approximation, this is likely to be a good assumption to make, although it might be prudent for more nuanced analyses to remove genes with significant alternative isoforms present in 3T3 cells.

We found microRNA target sites in extension regions, restricting ourselves to the top five most highly expressed microRNAs in mouse 3T3 cells (Rissland et al. in press). We predicted functional sites with TargetScan (Friedman et al. 2009) using only 3′ UTRs contained within the RefSeq annotations used to build the TargetScan databases, and restricting predictions to conserved 7mer target sites. For comparison, we chose tandem UTRs without microRNA target sites in the extension region. Because of the correlation between 3′ UTR length and stability, and because the microRNA-containing extension regions were significantly longer than the average (data not shown), we matched extension region length between our microRNA-targets and the controls. The distribution of log fold changes between the short and long isoforms for the control UTRs was centered near zero (mean log fold change=0.013), meaning the matched non-microRNA containing extension regions did not exert a considerably negative effect on gene stability (Figure 4A). However, this distribution was significantly shifted to the right (p=0.013, KS test) for the microRNA-containing extension regions (mean log fold change=0.167), confirming that microRNAs are at least partly responsible for the destabilizing effect of 3′ UTR lengthening. A similar but non-significant shift was observed when using only the presence of the (non-conserved) microRNA 7mer seed match as prediction of microRNA targeting, although the difference in means was less (0.03 in the control vs 0.10 with the microRNA target site).

Other stability-influencing motifs While microRNA target sites are among the most readily identified among stability-influencing motifs, we explored the effects of other known regulatory systems. We were unable to observe a significant difference in stability based on presence of AU-rich elements, but given the variety of ARE-binding proteins and their potentially antagonistic effects, as well as the degenerate nature of the ARE motif, this is unsurprising. We were able to observe a significant (p=1.3×10^{-4}, KS test) difference between extension regions containing the PUF motif and their matched controls, with a slightly smaller difference than that observed for the microRNA-containing extensions (mean log fold change=0.04 in control vs 0.14 containing the PUF motif).

Least stable 3′ UTRs are the most conserved To more comprehensively assess the contribution of sequence determinants to mRNA

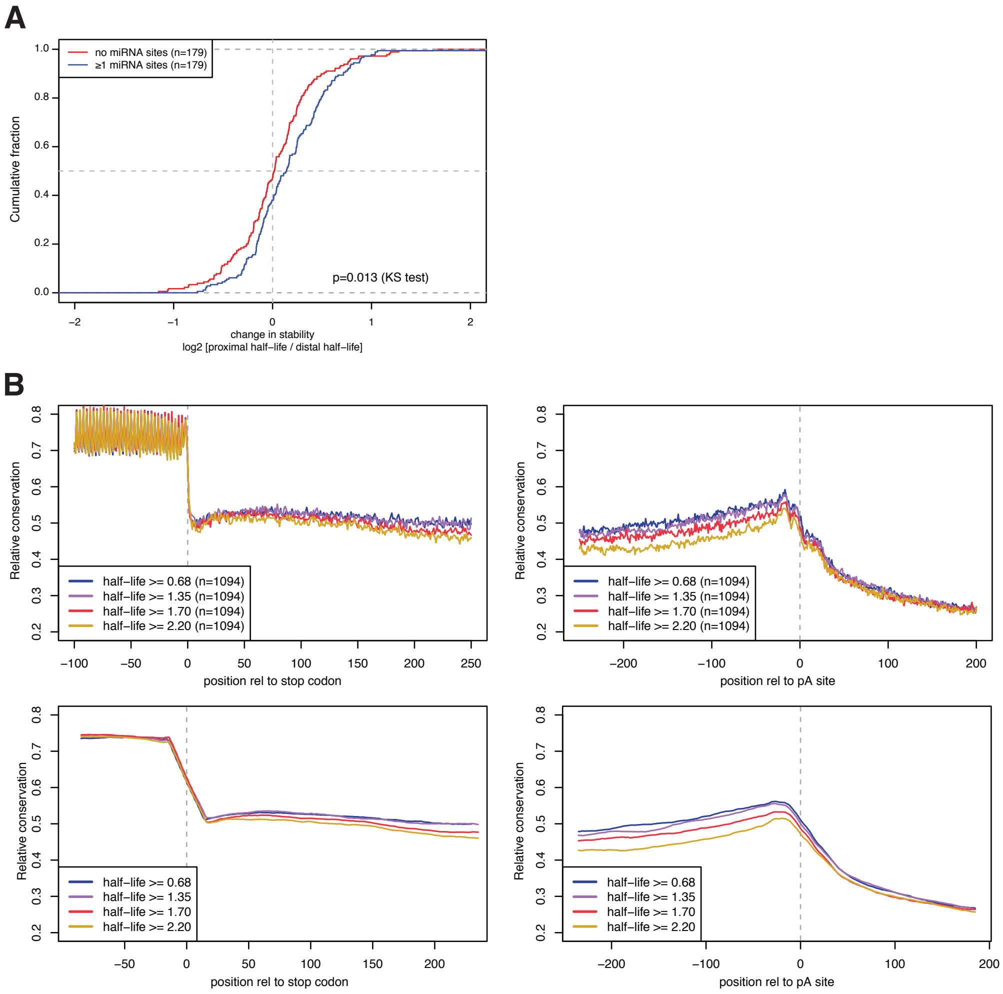


Figure 4: Sequence determinants of tandem UTR differential stability.

(A) MicroRNAs destabilize long tandem 3′ UTR isoforms. Curves compare distributions of log fold change in half-life between proximal isoform and distal isoform, so if the proximal isoform is more stable than the distal isoform, it would fall to the right of zero. Genes with a 7mer microRNA target site for miR-29 or let-7 (top miRs expressed in 3T3 cells) in the extension region are plotted in blue, and matched tandem UTRs without a microRNA target site are plotted in red.
(B) Per-base conservation for genes binned by stability of their most highly expressed 3′ UTR isoform. Conservation was calculated relative to the number of species, out of 13 vertebrates, which were alignable to the given 3′ UTR (see Methods).

half-life, we binned genes based on the half-life of their most highly expressed 3′ UTR isoform, and plotted conservation around the stop codon and the poly(A) site for these four bins (Figure 4B). As expected, conservation was high within the coding sequence and showed a three nucleotide periodicity. A dip in conservation occurs immediately downstream of the stop codon, and the middle of the 3′ UTR is the least conserved region, as has been noted in relation to microRNAs (Gaidatzis et al. 2007; Grimson et al. 2007). Overall UTR conservation peaks at the location of the poly(A) signal, approximately 20 bp upstream of the cleavage site.

We found a marked separation between the most stable genes (plotted in light orange), which had the least conservation on average, and the least stable genes (plotted in dark blue), which were on average the most conserved. This separation was not evident within the coding sequence (see smoothed curves, bottom panels of figure 4B), and became greatest toward the 3′ end of the UTR, nearest to the poly(A) signal. The conservation patterns for the different bins once again converge downstream of the poly(A) site. This pattern is striking, and strongly supports sequence motifs as the most important factor in determining differential stability of 3′ UTR isoforms.

3.3 Discussion

We have shown here that shortened 3′ UTR isoforms are globally more stable in large part due to the loss of destabilizing sequence elements harbored in the extension region of the long isoform. We conclude that 3′ UTR shortening as cells enter proliferative states (Sandberg et al. 2008) or lengthening as development proceeds (Ji et al. 2009) enable the cell to coordinately upregulate or downregulate, respectively, large numbers of genes.

We performed the first, to our knowledge, 3′ UTR isoform-specific stability measurements. 3P-Seq provided quantitative data as evidenced by good correlations with RNA-Seq data and gene-level half-life data. However, the observed correlation (Spearman correlation=0.77) with an RNA-Seq experiment performed in the same cell type shows that there is still room for improvement in the use of 3P-Seq for isoform quantitation. 3P-Seq reads all fall within ~20 bp of each consensus poly(A) site, yielding a short sequence space to produce reads from. This would magnify the effects of any sequence-specific biases in the library preparation methods, and mRNA secondary structure could also bias the libraries. In support of such biases, RNA-Seq signal often spikes randomly across the length of the mRNA. RNA-Seq, in contrast, averages reads over the entire constitutive portions of each gene, usually at least several hundred base pairs if not several kb, producing a highly accurate measure of gene expression.

The 3P-Seq protocol involves numerous purifications which may reduce the quantitative nature of the library preparations. We found that the simpler although less specific 3′ end library preparation method PAS-Seq (Shepard et al. 2011) showed a similar correlation with HeLa RNA-Seq as 3P-Seq even after controlling for the total read number of each end tag library (data not shown). This suggests that there may be an inherit limitation to quantitating 3′ UTR isoforms.

The shift towards increased stability of the proximal 3′ UTR isoform is highly statistically significant, but interestingly, we observed a large variance in the differential stability between short and long isoforms. A considerable number of genes show the opposite pattern, where the distal isoform is in fact more stable than the proximal isoform. We conclude that the majority of stability regulating cis motifs are destabilizing in nature, but that stabilizing motifs are also prevalent in the genome. This is interesting, as most well-characterized 3′ regulatory sequences are known to have destabilizing roles, including microRNAs, PUF-binding sites and AU-rich elements, and at least some published reports of stabilization caused by these motifs may involve de-repression by competitive binding to destabilizing elements (Barreau et al. 2005).

Because we compared long and short isoforms within a single cell type, stabilizing effects cannot be due to differences in levels of trans factors present, and hence the stabilization of long isoforms we observed must be due to sequence motifs directly increasing mRNA stability. Indeed, it is likely that destabilized long isoforms include both stabilizing and destabilizing elements, although on balance the destabilizing elements are either more prevalent or more potent. Given higher-resolution isoform-specific half-life data, it would be possible to predict which elements are important in stabilizing 3′ UTRs, and an eventual goal should be to build a quantitative model that can predict the combined effects of multiple stabilizing and destabilizing motifs in a tissue-specific manner.

Our conservation results argue that the destabilizing elements are globally more important biologically than the stabilizing elements. The most unstable 3′ UTRs were the most highly conserved, and conservation patterns anticorrelated with the stability of the message, especially closest to the poly(A) site. Future analyses could focus on the tandem UTRs where the long isoform has a greater half-life than the short isoform to tease out the biological relevance of the stabilizing elements.

Previous work has shown that microRNA binding sites closest to the edges of the 3′ UTR are most effective, and in support of this being a general phenomenon, we find peaks of conservation at either end of the UTR (Gaidatzis et al. 2007; Grimson et al. 2007). Follow-up work on AU-rich elements and PUF-binding sites could test whether these elements are also most effective at the 5′ and 3′ ends of the UTR. If this were generally true, it may imply additional looping of the mRNA to bring the stop codon into proximity with the poly(A) tail and the mRNA cap.

3.4 Methods

3P-Seq Library Preparation RNA was TRIzol-extracted from low-confluency dividing 3T3 cells. For the degradation library, cells were treated with 10μM actinomycin D, then harvested after 8 hours of incubation. 3P-Seq libraries were prepared as previously (Jan et al. 2011) using 60μg of RNA each.

Annotation of mammalian poly(A) sites using 3P-Seq data Base-calling from Illumina libraries was improved by using nucleotides 11–15 of each read for cluster-calling. This allows the Illumina software to distinguish close-by clusters from one another even if they begin with numerous identical bases (in this case, T's from the poly(A) tail). After removing the 3′ 3 bp (the 3 mRNA-proximal bp), 33 bp reads were reverse complemented to match the mRNA strand and mapped to the mouse mm9 genome using a perfect 25mer seed match.

To call 3P-Seq read clusters, overlapping transcripts on the same strand were first grouped together, and extended an arbitrary 3.5kb downstream or a minimum of 500 bp upstream of the next downstream gene on the same strand. An iterative process was used to call clusters. First, the poly(A) site with the most supporting reads (ie those reads with 3′ end at the given position, after stripping at least 1 3′ end A) was chosen. Second, reads with 3′ ends (after stripping A's) within the region 20 bp to either side of this position were analyzed. Third, to include the cluster as a bona fide poly(A) site, the following criteria were required:

1. At least one overlapping read had 4 or more 3′ end A's

2. At least 1% of all reads with end A's mapping to the gene overlapped the poly(A) site region
3. At least two reads with non-genomic 3′ end A's must overlap the poly(A) site region and at least one read must have 2 or more untemplated A's (unless there were more than 3 reads with 1 non-genomic 3′ end A).

If the poly(A) site region passed those criteria, it was annotated as a true poly(A) site cluster. All reads with a putative poly(A) site (inferred from the 3′ end of the read after stripping A's) within 20 bp were included in the cluster, and the remaining reads were used to iteratively call further clusters for the same gene.

Isoform quantitation Isoforms were quantitated using all reads with 3′ ends overlapping the 41 bp region around each poly(A) site cluster. For half-life analyses, only isoforms with at least 10 such reads in the control condition were used, and when isoforms were within 100 bp of one another, the most highly expressed (in the control library) isoform was chosen and the other one excluded. For analysis of proximal vs distal tandem UTR isoforms, only the top two most highly expressed isoforms were chosen.

Cis regulatory motif analyses Predicted microRNA target sites were downloaded from TargetScan (Friedman et al. 2009). We used the conserved 7mer predicted sites. We also compiled a list of UTR isoforms containing the 7mer motifs, without further prediction of efficacy (eg without regard to conservation). The following motifs were used to predict AU-rich elements: `UUAUUUAWW` and `WWWUAUUUAUWWW`, where `W` is either `U` or `A` (Barreau et al. 2005). The Pum2 PUF motif `UGUANAUA` was used, where `N` could be any nucleotide (Hafner et al. 2010).

Conservation 30-way multiz (Blanchette et al. 2004) vertebrate alignments were downloaded from the UCSC Genome Browser (Kent et al. 2002). The following species from this alignment were used because of reasonably good assembly quality: mouse, human, rat, chimpanzee, rhesus, opossum, cat, cow, chicken, guinea pig, orangutan, dog, frog (*Xenopus tropicalis*). For each 3′ UTR, the corresponding region was extracted from the multiz alignment. Conservation for each position in the alignment was calculated as the number of species with sequence matching the mouse sequence, divided by the number of organisms which were aligned to the 3′ UTR.

3.5 References

Andreassi, C and A Riccio (Sept. 2009). "To localize or not to localize: mRNA fate is in 3′UTR ends." In: *Trends Cell Biol* 19.9, pp. 465–74. DOI: `10.1016/j.tcb.2009.06.001` (cit. on p. 53).

Barreau, C, L Paillard, and HB Osborne (2005). "AU-rich elements and associated factors: are there unifying principles?" In: *Nucleic Acids Res* 33.22, pp. 7138–50. DOI: `10.1093/nar/gki1012` (cit. on pp. 62, 63).

Blanchette, M, WJ Kent, C Riemer, L Elnitski, AFA Smit, KM Roskin, R Baertsch, K Rosenbloom, H Clawson, ED Green, D Haussler, and W Miller (Apr. 2004). "Aligning multiple genomic sequences with the threaded blockset aligner." In: *Genome Res* 14.4, pp. 708–15. DOI: `10.1101/gr.1933104` (cit. on p. 63).

Filipowicz, W, SN Bhattacharyya, and N Sonenberg (Feb. 2008). "Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight?" In: *Nat Rev Genet* 9.2, pp. 102–14. DOI: `10.1038/nrg2290` (cit. on p. 53).

Friedman, RC, KKH Farh, CB Burge, and DP Bartel (Jan. 2009). “Most mammalian mRNAs are conserved targets of microRNAs.” In: *Genome Res* 19.1, pp. 92–105. DOI: 10.1101/gr.082701.108 (cit. on pp. 59, 63).

Fu, Y, Y Sun, Y Li, J Li, X Rao, C Chen, and A Xu (May 2011). “Differential genome-wide profiling of tandem 3′ UTRs among human breast cancer and normal cells by high-throughput sequencing.” In: *Genome Res* 21.5, pp. 741–7. DOI: 10.1101/gr.115295.110 (cit. on p. 54).

Gaidatzis, D, E van Nimwegen, J Hausser, and M Zavolan (2007). “Inference of miRNA targets using evolutionary conservation and pathway analysis.” In: *BMC Bioinformatics* 8, p. 69. DOI: 10.1186/1471-2105-8-69 (cit. on pp. 61, 62).

Grimson, A, KKH Farh, WK Johnston, P Garrett-Engele, LP Lim, and DP Bartel (July 2007). “MicroRNA targeting specificity in mammals: determinants beyond seed pairing.” In: *Mol Cell* 27.1, pp. 91–105. DOI: 10.1016/j.molcel.2007.06.017 (cit. on pp. 53, 61, 62).

Hafner, M, M Landthaler, L Burger, M Khorshid, J Hausser, P Berninger, A Rothballer, M Ascano Jr, AC Jungkamp, M Munschauer, A Ulrich, GS Wardle, S Dewell, M Zavolan, and T Tuschl (Apr. 2010). “Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.” In: *Cell* 141.1, pp. 129–41. DOI: 10.1016/j.cell.2010.03.009 (cit. on p. 63).

Jan, CH, RC Friedman, JG Ruby, and DP Bartel (Jan. 2011). “Formation, regulation and evolution of Caenorhabditis elegans 3′UTRs.” In: *Nature* 469.7328, pp. 97–101. DOI: 10.1038/nature09616 (cit. on pp. 54–56, 62).

Ji, Z, JY Lee, Z Pan, B Jiang, and B Tian (Apr. 2009). “Progressive lengthening of 3′ untranslated regions of mRNAs by alternative polyadenylation during mouse embryonic development.” eng. In: *Proc Natl Acad Sci USA* 106.17, pp. 7028–33. DOI: 10.1073/pnas.0900028106 (cit. on pp. 53, 61).

Kent, WJ, CW Sugnet, TS Furey, KM Roskin, TH Pringle, AM Zahler, and D Haussler (June 2002). “The human genome browser at UCSC.” In: *Genome Res* 12.6, pp. 996–1006. DOI: 10.1101/gr.229102.ArticlepublishedonlinebeforeprintinMay2002 (cit. on p. 63).

Lebedeva, S, M Jens, K Theil, B Schwanhäusser, M Selbach, M Landthaler, and N Rajewsky (Aug. 2011). “Transcriptome-wide Analysis of Regulatory Interactions of the RNA-Binding Protein HuR.” In: *Mol Cell* 43.3, pp. 340–52. DOI: 10.1016/j.molcel.2011.06.008 (cit. on p. 53).

Mayr, C and DP Bartel (Aug. 2009). “Widespread shortening of 3′UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.” In: *Cell* 138.4, pp. 673–84. DOI: 10.1016/j.cell.2009.06.016 (cit. on pp. 53, 54).

Mukherjee, N, DL Corcoran, JD Nusbaum, DW Reid, S Georgiev, M Hafner, M Ascano Jr, T Tuschl, U Ohler, and JD Keene (Aug. 2011). “Integrative Regulatory Mapping Indicates that the RNA-Binding Protein HuR Couples Pre-mRNA Processing and mRNA Stability.” In: *Mol Cell* 43.3, pp. 327–39. DOI: 10.1016/j.molcel.2011.06.007 (cit. on p. 53).

Quenault, T, T Lithgow, and A Traven (Feb. 2011). “PUF proteins: repression, activation and mRNA localization.” In: *Trends Cell Biol* 21.2, pp. 104–12. DOI: 10.1016/j.tcb.2010.09.013 (cit. on p. 53).

Rissland, OS, SJ Hong, and DP Bartel (in press). "MicroRNA Destabilization Enables Dynamic Regulation of the miR-16 Family in Response to Cell Cycle Changes." In: *Mol Cell* (cit. on p. 59).

Sandberg, R, JR Neilson, A Sarma, PA Sharp, and CB Burge (June 2008). "Proliferating cells express mRNAs with shortened 3′ untranslated regions and fewer microRNA target sites." In: *Science* 320.5883, pp. 1643–7. DOI: `10.1126/science.1155390` (cit. on pp. 53, 54, 61).

Sharova, LV, AA Sharov, T Nedorezov, Y Piao, N Shaik, and MSH Ko (Feb. 2009). "Database for mRNA half-life of 19 977 genes obtained by DNA microarray analysis of pluripotent and differentiating mouse embryonic stem cells." In: *DNA Res* 16.1, pp. 45–58. DOI: `10.1093/dnares/dsn030` (cit. on pp. 54, 57, 58).

Shepard, PJ, EA Choi, J Lu, LA Flanagan, KJ Hertel, and Y Shi (Apr. 2011). "Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq." In: *RNA* 17.4, pp. 761–72. DOI: `10.1261/rna.2581711` (cit. on pp. 54, 61).

Sonenberg, N and AG Hinnebusch (Feb. 2009). "Regulation of translation initiation in eukaryotes: mechanisms and biological targets." In: *Cell* 136.4, pp. 731–45. DOI: `10.1016/j.cell.2009.01.042` (cit. on p. 53).

Tian, B, J Hu, H Zhang, and CS Lutz (Jan. 2005). "A large-scale analysis of mRNA polyadenylation of human and mouse genes." eng. In: *Nucleic Acids Res* 33.1. - look at refs: 13 (poly-A tail length), 14, 15, pp. 201–12. DOI: `10.1093/nar/gki158` (cit. on pp. 53, 55).

Tian, B, Z Pan, and JY Lee (Feb. 2007). "Widespread mRNA polyadenylation events in introns indicate dynamic interplay between polyadenylation and splicing." eng. In: *Genome Res* 17.2, pp. 156–65. DOI: `10.1101/gr.5532707` (cit. on p. 55).

Yang, E, E van Nimwegen, M Zavolan, N Rajewsky, M Schroeder, M Magnasco, and JE Darnell Jr (Aug. 2003). "Decay rates of human mRNAs: correlation with functional characteristics and sequence attributes." In: *Genome Res* 13.8, pp. 1863–72. DOI: `10.1101/gr.1272403` (cit. on p. 54).

Chapter 4

TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway

Marc Bühler*, **Noah Spies***, **David P Bartel and Danesh Moazed**

Abstract

In the fission yeast Schizosaccharomyces pombe, the RNA interference (RNAi) machinery is required to generate small interfering RNAs (siRNAs) that mediate heterochromatic gene silencing. Efficient silencing also requires the TRAMP complex, which contains the noncanonical Cid14 poly(A) polymerase and targets aberrant RNAs for degradation. Here we use high-throughput sequencing to analyze Argonaute-associated small RNAs (sRNAs) in both the presence and absence of Cid14. Most sRNAs in fission yeast start with a 5′ uracil, and we argue these are loaded most efficiently into Argonaute. In wild-type cells most sRNAs match to repeated regions of the genome, whereas in cid14 cells the sRNA profile changes to include major new classes of sRNAs originating from ribosomal RNAs and a tRNA. Thus, Cid14 prevents certain abundant RNAs from becoming substrates for the RNAi machinery, thereby freeing the RNAi machinery to act on its proper targets.

*These authors contributed equally to this project

Contents

This chapter has been published as:

Bühler, M, N Spies, DP Bartel, and D Moazed (Oct. 2008). “TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway.” In: *Nat Struct Mol Biol* 15.10, pp. 1015–23. DOI: 10.1038/nsmb.1481.

Author contributions M.B. performed the experimental work with yeast; N.S. performed the computational analysis of sequencing reads; All authors contributed to the design of the study and preparation of the manuscript.

Acknowledgments We thank W. Johnston and S. Buker for reagents and members of the Moazed and Bartel laboratories for helpful discussions. We also thank H. Grosshans for comments on the manuscript.

4.1 Introduction

RNAi is a conserved silencing mechanism that is triggered by double-stranded RNA (dsRNA) (Hannon 2002; Fire et al. 1998). Silencing is mediated by small interfering RNAs (siRNAs) of about 22 nucleotides (nt) in size, which are produced from the long dsRNA by the Dicer RNase (Bernstein et al. 2001; Elbashir et al. 2001; Zamore et al. 2000; Hammond et al. 2000; Hamilton and Baulcombe 1999). siRNAs guide Argonaute proteins to complementary nucleic acids where they promote the inactivation of the homologous sequences. In some systems, efficient RNAi requires synthesis of dsRNA by an RNA-directed RNA polymerase (RdRP) (Sijen et al. 2001; Baulcombe 2004). Besides their role in post-transcriptional gene silencing (PTGS), siRNAs have also been implicated in regulation at the DNA and chromatin levels in plants and some fungi (Bühler and Moazed 2007; Zaratiegui et al. 2007).

The role of siRNAs in gene regulation at the chromatin level has been well studied in fission yeast, whose genome encodes a single gene each for Argonaute, Dicer and RdRP: ago1+, dcr1+ and rdp1+, respectively. At centromeres, deletion of any of these genes results in a loss of gene silencing, reduced histone H3 lysine 9 (H3K9) methylation and Swi6 (the homolog of heterochromatin protein-1 (HP1)) localization, all of which are conserved molecular markers of heterochromatin (Volpe et al. 2002). Ago1 is found in the RNA-induced transcriptional silencing (RITS) and Argonaute siRNA chaperone (ARC) complexes (Verdel et al. 2004; Buker et al. 2007). Early sequencing of small RNAs from fission yeast revealed heterochromatic siRNAs that match centromeric repeats (Reinhart and Bartel 2002). In addition, 1,300 siRNAs isolated from the RITS complex, using a tag on its Chp1 subunit, have been reported (Cam et al. 2005). These RITS-associated siRNAs are 20–22-nt long and map to repeat elements embedded in heterochromatic regions, the ribosomal DNA (rDNA) array, intergenic regions, mRNAs, tRNAs, subtelomeric and silent mating-type regions (Cam et al. 2005). Generation of these siRNAs requires Dicer, Argonaute and Rdp1 (Verdel et al. 2004; Motamedi et al. 2004; Bühler et al. 2006).

The RNAi pathway is essential for high levels of H3K9 methylation and gene silencing at fission yeast centromeres, but it is dispensable at other heterochromatic loci such as telomeres or the silent mating-type loci (Volpe et al. 2002; Sadaie et al. 2004). Although heterochromatin has long been thought to be transcriptionally inactive, recent observations in fission yeast show that heterochromatic domains are transcribed to some degree (Volpe et al. 2002; Bühler et al. 2006; Bühler et al. 2007; Chen et al. 2008). However, the resulting heterochromatic transcripts are rapidly turned over by a mechanism called co-transcriptional gene silencing (CTGS) (Bühler et al. 2006; Bühler and Moazed 2007). Although possibly mediated by the RNAi pathway at centromeres, CTGS at other fission yeast heterochromatic regions depends on a specialized polyadenylation complex referred to as the TRAMP (Trf4-Air1/Air2-Mtr4 polyadenylation) complex (LaCava et al. 2005), most likely targeting heterochromatic transcripts for degradation by the exosome (Bühler et al. 2007).

The role of TRAMP in exosome-mediated degradation of aberrant RNAs was first described in budding yeast (LaCava et al. 2005). Homologs of the budding yeast TRAMP subunits Trf4/5, Air1/2 and Mtr4 are found in the fission yeast TRAMP complex (Bühler et al. 2007). The S. pombe homolog of the budding yeast Trf4/5 poly(A) polymerases is Cid14, a member of the Cid1 family of noncanonical poly(A) polymerases (Stevenson and Norbury 2006). Cid14 is required for polyadenylation of ribosomal RNAs (rRNAs) and proper chromosome segregation (Win et al. 2006). In addition to its role in rRNA biogenesis and CTGS, deletion of cid14+ results in a dramatic decrease in centromeric siRNA levels, suggesting a role for Cid14 in siRNA biogenesis or stabilization (Bühler et al. 2007).

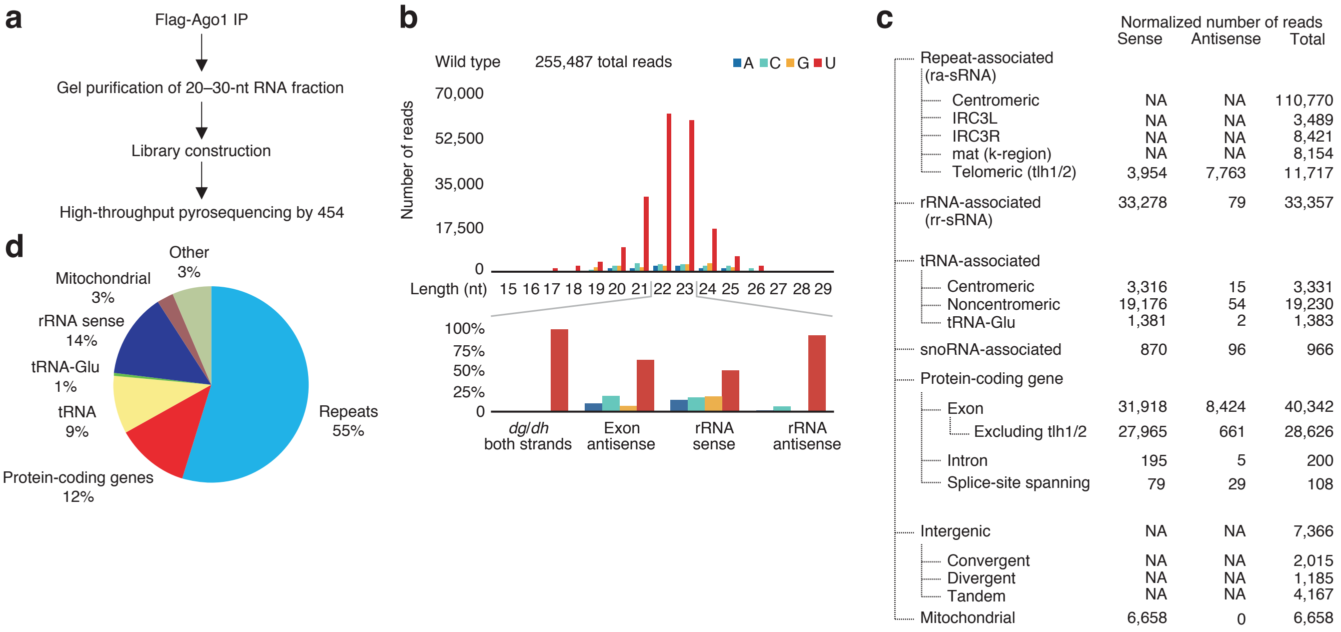
To better understand the role of Cid14 in accumulation of centromeric siRNAs, we used high-throughput sequencing to examine Ago1-associated small RNAs in wild-type and cid14 fission yeast cells. Most of the small RNAs recovered by an Ago1 pull-down start with a 5′ U and are 22 nt or 23 nt long. In wild-type cells, most Ago1-associated small RNAs correspond to repetitive DNA elements found at the centromeres. Other Ago1-associated small RNAs match the sequences of tRNAs, small nucleolar RNAs (snoRNAs), rDNA and intergenic regions. The small RNA profile changes dramatically in cid14 cells. Consistent with previous findings (Bühler et al. 2007), the levels of centromeric siRNAs are reduced in cid14 cells, whereas the levels of other small RNAs increase dramatically. The most prominent new class of small RNAs in cid14 cells includes those that match tRNA-Glu and ribosomal RNA sequences, which are normally substrates of TRAMP (Wyers et al. 2005; Vanácová et al. 2005; Win et al. 2006; LaCava et al. 2005). These findings indicate that Cid14 acts as a negative regulator of siRNA biogenesis by competing with the RNAi machinery for substrates.

4.2 Results

Ago1-associated small RNAs To obtain a more comprehensive view of the S. pombe siRNA profile and to better understand the connection between Cid14 and RNAi20, we generated small RNA libraries from affinity-purified Flag-tagged Ago1. These libraries should contain siRNAs from RITS, ARC and free or other possible Ago1 complexes. We then subjected the libraries to high-throughput pyrosequencing (Margulies et al. 2005) (Fig. 1a). Analysis of 200,000 sequences showed that most of the Ago1-associated small RNAs derived from wild-type cells were 22 nt or 23 nt long (Fig. 1b) and that most matched repetitive elements in the genome (55%, Fig. 1c,d). Other small RNAs matched annotated sequences of rDNA, tRNAs, snoRNAs, intergenic regions, introns, exons and mitochondrial DNA (Fig. 1c,d). We classified Ago1-associated small RNAs as siRNAs and sRNAs. The term 'siRNA' was used when there was evidence that Dcr1 generated the small RNA. In other cases, Ago1 seemed to be associated with small RNAs that corresponded to abundant cellular RNAs and derived from mostly the sense strand, and thus seemed to be generated primarily by non-Dcr1 degradation processes. For example, no reads were antisense to mitochondrial genes, suggesting that all mitochondrial reads were fragments of normal transcripts. To distinguish this set of small RNAs, we refer to them throughout this work as 'sRNAs'. Although some sRNAs, such as antisense gene-specific sRNAs, might be produced by Dcr1 and could have physiological roles, a larger fraction seemed to be degradation products that may nonspecifically associate with overexpressed Flag-Ago1. In this paper, we focus on the small RNA populations that either derived from centromeric repeat sequences or showed a shift in their abundance in cid14 cells.

General properties of Ago1-associated siRNAs Consistent with previous reports (Cam et al. 2005; Bühler et al. 2007), siRNAs corresponding to the centromeric dg and dh repeats were present in the Ago-associated small RNA pool, with similar numbers matching the forward and reverse strands. Most siRNAs in plants also derive from both DNA strands (Rajagopalan et al. 2006), whereas those in Caenorhabditis elegans are predominantly antisense to mRNAs (Ruby et al. 2006; Ambros et al. 2003). In S. pombe, the origin from both strands probably reflects RNA polymerase II (RNA Pol II) transcription in both directions, which gives rise to forward and reverse transcripts that are then converted to dsRNA by the RNA-directed RNA polymerase complex (RDRC).

As observed for some classes of Argonaute-



	Normalized number of reads		
	Sense	Antisense	Total
Repeat-associated (ra-sRNA)			
Centromeric	NA	NA	110,770
IRC3L	NA	NA	3,489
IRC3R	NA	NA	8,421
mat (k-region)	NA	NA	8,154
Telomeric (tlh1/2)	3,954	7,763	11,717
rRNA-associated (rr-sRNA)	33,278	79	33,357
tRNA-associated			
Centromeric	3,316	15	3,331
Noncentromeric	19,176	54	19,230
tRNA-Glu	1,381	2	1,383
snoRNA-associated	870	96	966
Protein-coding gene			
Exon	31,918	8,424	40,342
Excluding tlh1/2	27,965	661	28,626
Intron	195	5	200
Splice-site spanning	79	29	108
Intergenic	NA	NA	7,366
Convergent	NA	NA	2,015
Divergent	NA	NA	1,185
Tandem	NA	NA	4,167
Mitochondrial	6,658	0	6,658

Figure 1: Profiling of Ago1-associated small RNAs from wild-type cells

(A) Ago1-associated RNA was isolated and 20–30-nt RNAs were PAGE purified. Small RNA libraries suitable for 454 deep sequencing were generated as described previously3. IP, immunoprecipitation.
(B) Size distribution and the 5′-most nucleotide of Ago1-associated small RNAs.

(C) Classification of Ago1-associated small RNAs isolated from wild-type cells into mitochondrial, repeat-associated, gene-associated, rRNA-associated, tRNA-associated and snoRNA-associated small RNAs. If possible, the orientation of the small RNA with respect to its target is indicated. NA, not applicable.
(D) Pie chart illustrating percentages for the individual small RNA classes relative to the total number of small RNAs sequenced from wild-type cells.

associated sRNAs in other lineages (Ruby et al. 2006; Lau et al. 2001; Reinhart et al. 2002; Aravin et al. 2003; Aravin et al. 2006; Lau et al. 2006; Girard et al. 2006; Grivna et al. 2006; Watanabe et al. 2006), a large majority ($> 98\%$) of the siRNAs corresponding to centromeric dg and dh repeats started with a 5′ U (Fig. 1b). Previous projects using the same methods for library construction and sequencing revealed classes of siRNAs that started predominantly with 5′ guanosine and another that started mostly with 5′ adenosine, thereby indicating that our method does not artifactually favor the sequencing of RNAs with 5′ U (Rajagopalan et al. 2006; Ruby et al. 2006). Processing of the double-stranded RNA was also unlikely to explain most of this extreme bias for RNAs with 5′ U, because Dicer cleavage is thought to occur sequentially in 22–23-nt intervals, and the genome does not encode uracil at such regularly spaced intervals. Nonetheless, processing preferences could contribute to this bias, and we uncovered some evidence that they do contribute to a small degree.

Because the siRNAs were predominantly a near-equal mixture of 22-mers and 23-mers, a reasonable proposal would be that Dicer has some leeway in choosing the precise cleavage site and that sequence context might influence the choice of whether to cleave to produce 22-nt siRNAs or to cleave at the next base pair to produce 23-nt siRNAs. Therefore, we examined all 16 dinucleotide possibilities at positions 23 and 24, counting from the 5′ end of each sequenced siRNA (Supplementary Table 1 found in Appendix B). As would be expected if Dicer prefers to cleave before a uracil and thereby preferentially generates a downstream siRNA beginning with uracil, we observed a propensity toward 22-mers when the nucleotide at position 23 was a uracil. Other notable biases suggested that Dicer prefers to cleave at sites that avoid creating an siRNA beginning with G. However, all of these propensities were modest, generally less than three-fold, indicating that preferential siRNA processing contributes relatively little to the striking preference for 5′ U. Having ruled out a more-than-modest effect of preferential processing, we conclude that preferential stability of siRNAs beginning with U explains most of the bias for a 5′ U. This preferential stability could be at different levels, including preferential stability before encountering Ago1 or preferential stability after loading into Ago1. A reasonable hypothesis is that the much higher stability of 5′ U siRNAs arises primarily from a strong preference of Ago1 for loading siRNAs beginning with a 5′ U, and that those siRNAs that Ago1 rejects because they do not begin with a U are rapidly degraded.

To investigate 5′ nucleotide preferences in more depth, we considered the inferred siRNA duplexes corresponding to sequenced 23-mers deriving from the centromeric dg/dh repeats. (The choice of 23-mers over 22-mers stemmed from the notion that these longer siRNAs were less likely to be degradation intermediates of longer siRNAs.) When considering the influence of the 5′ nucleotide on siRNA loading and stability, six classes of duplexes that each involved siRNAs with different 5′ nucleotides were informative (Table 1). Regardless of the duplex under consideration, a consistent hierarchy was observed in the sequenced reads, with 5′ U $\gg$ A $>$ C $>$ G.

The $>$ 100-fold bias in reads from the strand beginning with a 5′ U was consistent with the idea that one of the two siRNA strands, the passenger strand, was discarded during loading (Buker et al. 2007; Rand et al. 2005; Matranga et al. 2005), probably after cleavage of the passenger strand by the inherent slicer activity of Ago1 (Buker et al. 2007; Irvine et al. 2006). Moreover, this bias showed that nearly all of the siRNAs that were sequenced were already single stranded, which indicated that in fission yeast the siRNA duplex is transient when compared to the loaded single strand. Furthermore, the predicted pairing asymmetry (Schwarz et al. 2003) had no correlation with the most frequently sequenced strand of these duplexes (Supplementary Table 2 and Supplementary Results, in Appendix B), as has been reported for endogenous siRNAs of plants (Rajagopalan et al. 2006).

Having ruled out pairing asymmetry as a factor influencing strand choice, we examined whether strand choice might be influenced by the identity of the 5′ nucleotide. As mentioned earlier, one hypothesis for explaining the abundance of siRNAs beginning with U is that Ago1 has a strong preference for loading siRNAs beginning with a 5′ U, and that those siRNAs that Ago1 rejects because they do not begin with a U are rapidly degraded. The alternative hypothesis is that siRNAs are loaded equally efficiently regardless of their 5′ nucleotide, and those siRNAs beginning with G, C and A are much less stable after loading than those are those beginning with U. Examination of the reads matching centromeric dg/dh repeats indicated that 5′ U siRNAs were more likely to be associated with Ago1 if they were paired originally to a 5′ A siRNA than if they were paired with another 5′ U siRNA. Because the model positing differential post-loading stabilities cannot explain this observation, but the model positing differential loading can explain it, we conclude that at least part of the 5′ U bias is due to the preferential loading of siRNAs beginning with U (Supplementary Results and Supplementary Fig. 1).

The cells used for the isolation of Ago1-associated small RNAs in our experiments contained a ura4+ transgene inserted into the outer centromeric repeats on the right arm of chromosome 1 (otr1R ::ura4+)(Allshire et al. 1994). We sequenced 249 siRNAs (20–25 nt) that corresponded to ura4+ sequences (Fig. 2a and Supplementary Table 3). Like the cen siRNAs, ura4+ siRNAs showed a preference for uracil at their 5′

terminus, but, unlike the cen siRNAs and consistent with previous results (Bühler et al. 2007), ura4+ siRNAs showed a five-fold preference for the sense strand (206 sense, 43 antisense; Fig. 2a–c and Supplementary Table 3).

More than 90% of the antisense reads corresponding to coding exons matched tlh1 and tlh2 (Fig. 1c), which are subtelomeric genes classified as ‘repeat associated’. The remaining 661 reads antisense to protein-coding exons were distributed among 341 genes, usually in far lower numbers than those of sense reads, although for adh1+ the numbers were roughly equal (Fig. 2d and Supplementary Table 3).

Small RNA profile changes in cid14 cells
In addition to its role in rRNA biogenesis and CTGS, Cid14 has been proposed to be involved in siRNA generation, as deletion of cid14+ results in a dramatic decrease in centromeric repeat-associated siRNA levels (Bühler et al. 2007). We found that deletion of cid14+ had no effect on the size distribution of siRNAs (Fig. 3). Consistent with previous findings, we observed a marked (five-fold) decrease in the fraction of reads mapping to centromeric repeats in cid14 cells (compare Fig. 1c,d with Fig. 3a,c). However, other classes of Ago1-associated small RNAs spanning many regions across all three chromosomes increased disproportionately (Fig. 3d), the most prominent among them being small RNAs antisense to rRNA, which increased by 274-fold (compare Fig. 1c with Fig. 3a,c). Both rRNAs and tRNAs have previously been shown to be targets for processing or degradation by the TRAMP and exosome pathway (Wyers et al. 2005; Vanácová et al. 2005; Win et al. 2006; LaCava et al. 2005). In contrast, the fraction of reads from gene-specific sense and antisense sRNAs were similar in wild-type and cid14 cells (increases of 1.4-fold and 1.1-fold, respectively). Our observations suggest that, in cells lacking Cid14, accumulated rRNAs become substrates for the RNAi pathway and give rise to siRNAs.

Internal repeat elements flank the centromeric repeat regions of chromosome 3 (Internal repeat centromere 3, IRC3R, Fig. 4a) and coincide precisely with a sharp decrease in H3K9 methylation and Swi6 levels (Cam et al. 2005). Therefore, they have been proposed to serve as boundary elements, similar to tRNA genes (Scott et al. 2006), that prohibit spreading of heterochromatin to euchromatic regions surrounding centromeres. H3K9 methylation levels at fission yeast centromeres, including dg/dh repeats and IRC elements (Cam et al. 2005), are reduced substantially in cells lacking siRNAs (dcr1), and RNAi has an essential role in the proper assembly of heterochromatin at these repeat elements. Consistent with previous results, in cid14 cells, centromeric siRNA levels were reduced by about 20-fold, but the levels of H3K9 methylation at the centromeric dg/dh repeats were unaffected (Bühler et al. 2007) (Fig. 3 and Fig. 4b). This reduction was far greater at IRC sequences, where we observed a nearly complete loss of siRNAs in cid14 cells (Fig. 4a).

To determine the possible contribution of RNAi, heterochromatin and the TRAMP and exosome pathways to the regulation of IRC transcripts, we used quantitative reverse-transcription PCR (RT-PCR) to analyze IRC3R transcript levels in cells that carried deletions or mutations in an essential gene in each pathway. IRC3R transcript levels were unaffected in clr4 and dcr1 cells (Fig. 4c), suggesting that these transcripts were not silenced by RNAi-mediated heterochromatin formation. In contrast, IRC3R transcript levels increased five- to seven-fold in cid14, mtr4-1 and dis3-54 mutant cells, indicating that IRC3R is a substrate of the TRAMP (Cid14/Mtr4) and the exosome (Dis3) pathways (Fig. 4c). Furthermore, deletion of rrp6+, a subunit of the nuclear exosome, did not affect IRC3R transcript levels, suggesting that degradation occurs in the cytoplasm rather than in the nucleus. We next asked whether H3K9me levels at the IRC on the right arm of chromosome 3 (IRC3R) were affected in cid14 cells. Unexpectedly, we did not detect any difference in H3K9 methylation levels between

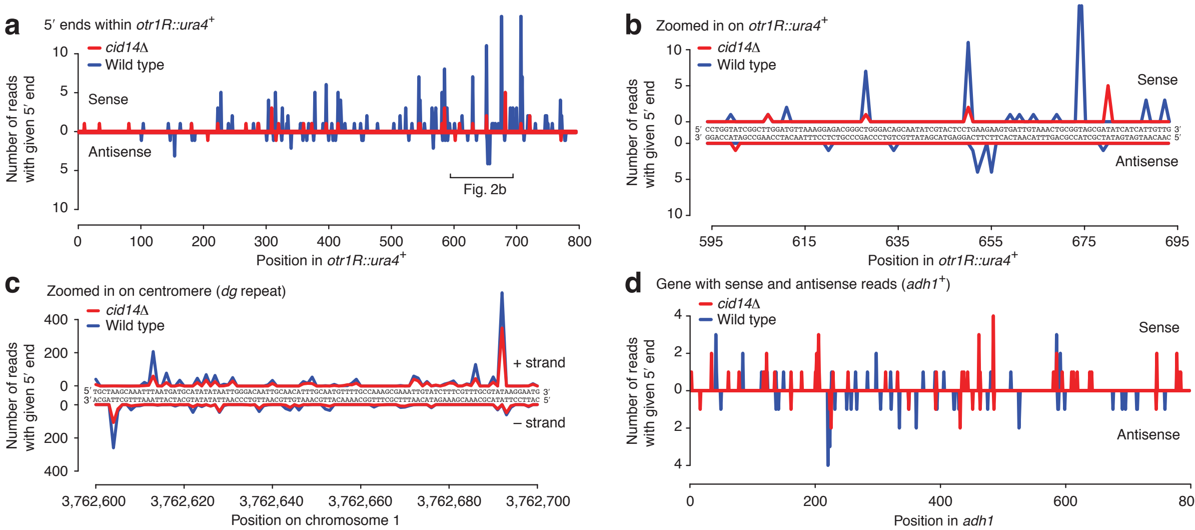


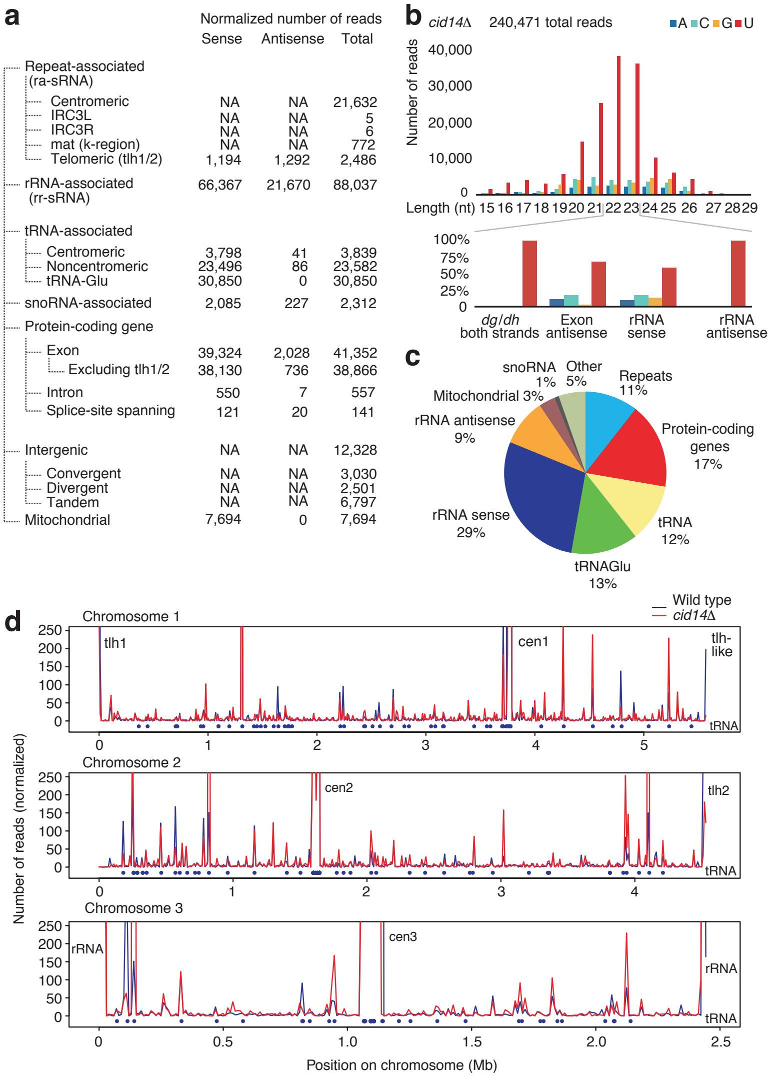
Figure 2: Distribution of reads mapping to genomic loci.

(A) Distribution of siRNAs at a ura4+ transgene inserted into the outermost centromeric repeats on the right arm of chromosome 1 (otr1R ::ura4+). ura4+ small RNAs show a five-fold preference for the sense strand, and only one of the two strands is found in Ago1. Peaks indicate the number of ura4+ reads with 5′ ends at each genomic position.
(B) Zoomed in version of a. Note that nearly all of the reads start with a T (U).

(C) Distribution of siRNAs at a centromeric dg repeat. For any given position, generally only one of the two centromeric siRNA strands, starting with a T (U), is present in Ago1.
(D) Distribution of sRNAs at the endogenous adh1+ gene.

wild-type and cid14 cells (Fig. 4b). In contrast, H3K9 methylation has been shown to be absent at IRCs in dcr1 cells (Cam et al. 2005). Together, these observations suggest that the spreading of H3K9 methylation into the IRC regions can occur independently of siRNAs but may be lost in dcr1 cells because of defects in RNAi-mediated nucleation of heterochromatin at the dg/dh repeats.

The sRNAs corresponding to the 5′ end of tRNA-Glu formed the third largest class of small RNAs found in the cid14 library (Fig. 3c). Whereas these RNAs were sequenced 1,381 times in wild-type cells, they were sequenced 30,850 times in cid14 cells (Fig. 3a,c and Fig. 4d). They also were clearly much more abundant than any other sRNAs mapping to tRNAs. Consistent with the sequencing data, the tRNA-Glu sRNA was specifically detected on northern blots of Ago1-associated RNAs from cid14 cells, but not from wild-type cells (Fig. 4e). The larger tRNA fragments present in Flag-Ago1 preparations were background RNAs, because they were also recovered from an untagged Ago1 strain (Fig. 4f). In contrast, tRNA-Glu sRNA was present only in Flag-Ago1 pull-downs (Fig. 4f). However, the tRNA-Glu sRNA was not generated by Dcr1 or Rdp1 (Fig. 4e). This observation is consistent with the idea that abundant small RNAs, which are in the size range of siRNAs, can associate with Ago1. However, the physiological significance of this association remains to be determined. In particular, chromatin immunoprecipitation (ChIP) experiments indicated that there was no increase in histone H3K9 methylation at the tRNA-Glu locus in cid14 cells (Fig. 4g). Sense sRNAs loaded onto Ago1 may be unable to initiate H3K9 methylation because they cannot base pair with sense nascent tRNA-Glu transcripts. The propensity of this sRNA to associate with Ago1 might stem in part from its 5′ U, although 10 of the other 69 unique tRNAs also begin with U.



	Normalized number of reads		
	Sense	Antisense	Total
Repeat-associated (ra-sRNA)			
Centromeric	NA	NA	21,632
IRC3L	NA	NA	5
IRC3R	NA	NA	6
mat (k-region)	NA	NA	772
Telomeric (tlh1/2)	1,194	1,292	2,486
rRNA-associated (rr-sRNA)	66,367	21,670	88,037
tRNA-associated			
Centromeric	3,798	41	3,839
Noncentromeric	23,496	86	23,582
tRNA-Glu	30,850	0	30,850
snoRNA-associated	2,085	227	2,312
Protein-coding gene			
Exon	39,324	2,028	41,352
Excluding tlh1/2	38,130	736	38,866
Intron	550	7	557
Splice-site spanning	121	20	141
Intergenic	NA	NA	12,328
Convergent	NA	NA	3,030
Divergent	NA	NA	2,501
Tandem	NA	NA	6,797
Mitochondrial	7,694	0	7,694

Figure 3: Profiling of Ago1-associated small RNAs from cid14 cells. Small RNA libraries suitable for 454 deep sequencing were generated as for wild-type cells.

(A) Classification of Ago1-associated small RNAs isolated from cid14 cells into the same classes as shown in Figure 1.

(B) Size distribution and indication of the 5′-most nucleotide of small RNAs.

(C) Pie chart illustrating percentages for the individual small RNA classes relative to the total amount of small RNAs sequenced from cid14 cells.

(D) Chromosomal distribution profiles of Ago1-associated small RNAs isolated from wild-type (blue) and cid14(red) cells. Blue bullets indicate the location of tRNA genes.

Ribosomal RNAs give rise to antisense siRNAs in cid14 cells Small RNAs mapping to rDNA were identified previously and represented about 30% of the total number of sequences in the collection of 1,300 RITS-associated small RNAs (Cam et al. 2005). However, fragments of the abundant rRNAs are present in nearly all small RNA sequence libraries, and it had remained unclear whether these rRNA-associated small RNAs were produced by the RNAi pathway or were degradation products. We observed that in wild-type cells small RNAs corresponding to rRNAs were mainly of the sense orientation (Fig. 5a,b), and furthermore, were generated independently of the RNAi pathway (Fig. 5c), suggesting that they may be rRNA

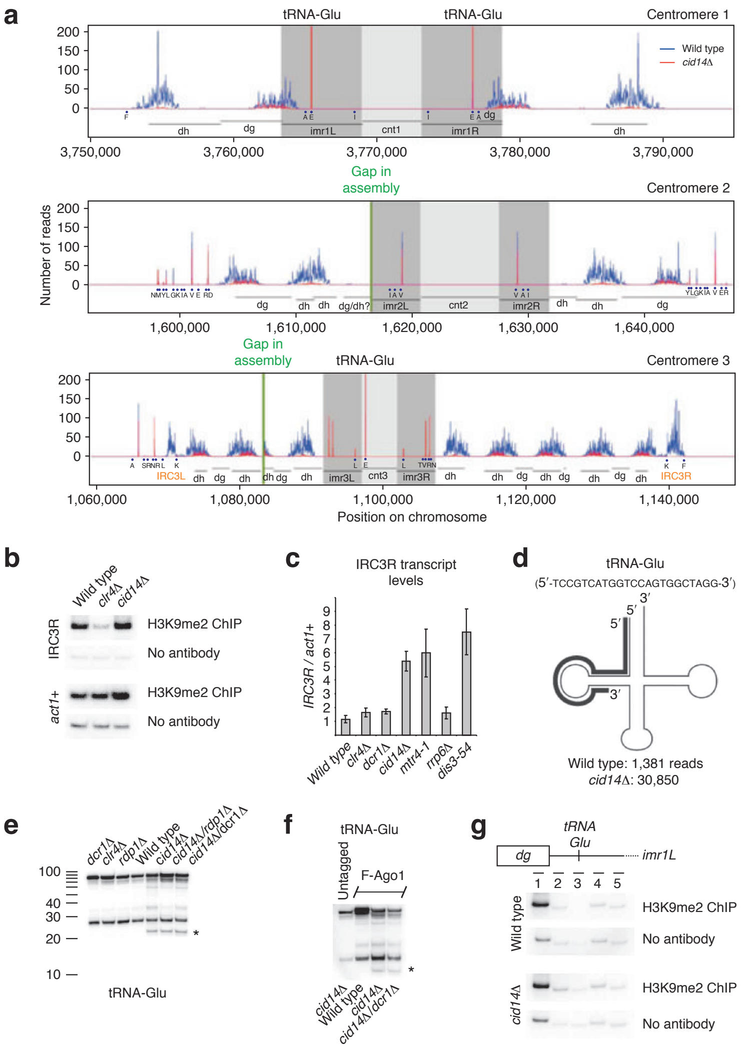


Figure 4: Small RNAs generated from centromeres in wild-type and cid14 cells.

(A) siRNA distribution at centromeres in wild-type (blue) and cid14 (red) cells. IRC3-L/R, unique inverted repeats flanking both the left and right sides of centromere 3 (Sijen et al. 2001); blue bullets, tRNA genes in single letter amino acid code. Three identical tRNA-Glu genes are found in centromeric heterochromatin, as well as three noncentromeric genes with identical sequence. Because all reads come from regions of perfect identity, it is ambiguous from which tRNA-Glu locus or loci these reads originate.
(B) Quantitative RT-PCR was performed to determine IRC transcript levels in various mutant backgrounds as indicated on the x-axes. H3K9me2, dimethylated H3K9.
(C) ChIP experiment showing that H3K9me2 in cid14 cells, where siRNAs are absent, is not affected at IRC3R. DNA from ChIP reactions with or without an antibody against H3K9me2 was used for PCR with primers to amplify the indicated sequences. Error bars are s.d.
(D) Cloverleaf schematic of tRNA-Glu. Bold line represents the most prevalent Ago1-associated small RNA (5′-`TCCGTCATGGTCCAGTGGCTAGG`-3′), which matches the tRNA-Glu 5′ end and D-loop.
(E) Northern blot of Ago1-associated RNAs demonstrating that the tRNA-Glu sRNA (indicated with an asterisk) was specifically detected from cid14 cells, but not from wild-type cells, in a dcr1- and rdp1-independent manner.
(F) Larger tRNA fragments are background contaminating RNAs, because they were also recovered from an untagged Ago1 strain.
(G) ChIP experiment showing that H3K9me2 around the tRNA-Glu genes found in centromere 1 is not different in wild-type and cid14 cells. DNA from ChIP reactions with or without an antibody against H3K9me2 was used for PCR with primers to amplify imr fragments 1–5.

degradation products that nonspecifically associate with Ago1. In cid14 cells, we observed a dramatic increase in small RNAs of the opposite orientation (antisense). Unlike the sense-strand small RNAs, antisense ribosomal small RNAs required Rdp1 and Dcr1 for their biogenesis (Fig. 5c). These antisense ribosomal small RNAs were therefore classified with confidence as siRNAs (rr-siRNA). Ribosomal RNA genes are transcribed as a unit by RNA polymerase I, and the completed transcript is rapidly processed to form the mature 18S, 5.8S and 28S rRNAs (Good et al. 1997) (Fig. 5a). Notably, antisense rr-siRNAs were more or less equally distributed along the 18S and 5.8S rRNAs, whereas most of the antisense 28S rr-siRNAs mapped to the $3'$ end (Fig. 5b). Together, these observations suggested that, in cid14 cells, rRNAs become substrates for dsRNA synthesis by the RDRC complex and processing into siRNAs by Dcr1, thereby suggesting competition between the components of the RNAi machinery and possible degradation or processing initiated by the TRAMP complex. The RNAi pathway is required for H3K9 methylation and silencing of foreign promoters inserted within the rDNA repeats (Cam et al. 2005), suggesting that the low levels of rr-siRNAs observed in cid14+ cells are functional. The dramatic increase in rr-siRNA levels in cid14 cells is likely to increase the efficiency of rDNA silencing and rDNA H3K9 methylation. Our efforts to unambiguously determine the role of Cid14 in regulation of rDNA H3K9 methylation were unsuccessful, probably because of the previously described variations in rDNA copy number in cid14 cells (Wang et al. 2008).

Deletion of Clr4 gives rise to antisense rr-siRNAs In addition to components of the RNAi pathway, the Clr4 H3K9 methyltransferase and its associated factors are required for centromeric siRNA generation in fission yeast (Verdel et al. 2004; Motamedi et al. 2004; Bühler et al. 2006; Hong et al. 2005; Li et al. 2005). The requirement for Clr4 in both H3K9 methylation and siRNA generation has been suggested to indicate a chromatin-dependent step in recruitment of RITS and RDRC to their target transcripts (Motamedi et al. 2004; Verdel and Moazed 2005). Here we found detectable levels of antisense rr-siRNAs in Ago1 pull-downs from clr4 cells (Fig. 5c). These observations suggest that rRNAs can become targets for the RNAi machinery when the components of the RNAi pathway are released from centromeres as a result of the lack of H3K9 methylation in clr4 cells, thus allowing them to access rRNAs that would usually be processed by the TRAMP pathway. Furthermore, the high rRNA abundance is likely to overcome the requirement for H3K9 methylation–dependent recruitment of RDRC, allowing siRNA generation on rRNA substrates.

4.3 Discussion

Our results provide a more comprehensive picture of Ago1-associated small RNAs in fission yeast and reveal new insights into their biogenesis and genomic distribution. Furthermore, our analysis of sRNAs in wild-type and cid14 cells revealed a previously unsuspected role for the RNA surveillance pathway involving the TRAMP complex in regulation of genomic siRNA distribution through removing entire classes of RNAs that have the potential to enter the sRNA pathways.

Specific siRNA features The vast majority of Ago1-associated sRNAs contained U at the $5'$ position. This preference for $5'$ U was mostly attributed to much higher stability of the $5'$ U siRNAs, which reflects a marked loading preference for those siRNAs beginning with U. Although the relationship between $5'$ nucleotide composition and biogenesis, loading and stability have not been teased apart in most other systems, this preference for a $5'$ U in loading might be conserved in a large subset of Argonaute and Piwi family pro-

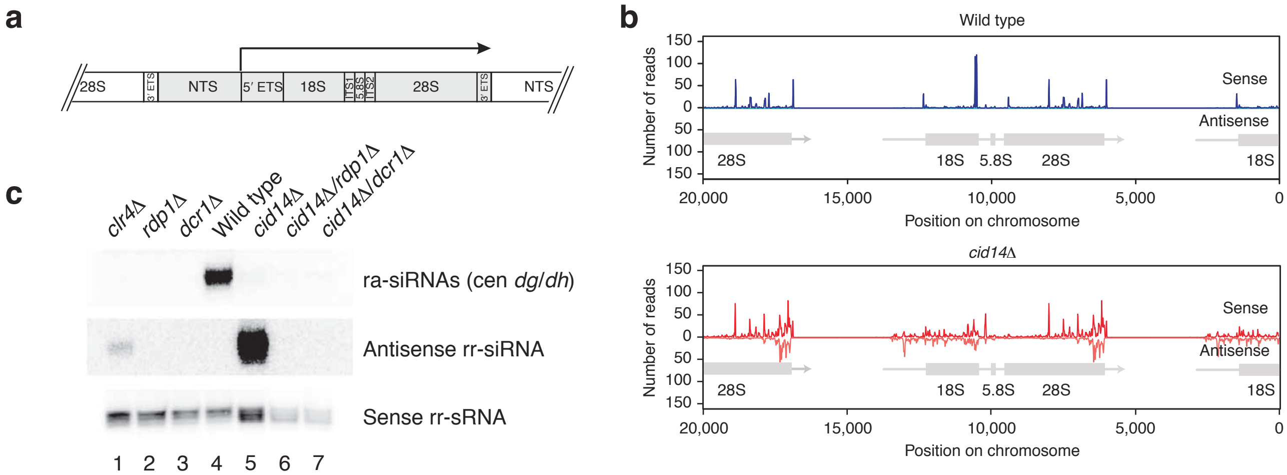


Figure 5: Ribosomal RNAs give rise to antisense siRNAs (rr-siRNAs) in cid14 cells.

(A) Structure of the S. pombe rDNA unit (Good et al. 1997). The long precursor RNA indicated by the arrow is rapidly processed to form the mature 18S, 5.8S and 28S rRNAs through removal of the 5′ and 3′ external transcribed spacers (ETS) and the internal transcribed spacers (ITS) 1 and 2. The nontranscribed spacer (NTS) separates the different rRNA units at the rDNA locus.
(B) Antisense rr-siRNAs are produced only in cid14cells. Antisense rr-siRNAs are more or less equally distributed along the 18S and 5.8S rRNAs, whereas most of the antisense 28S rr-siRNAs map to the 3′ end.
(C) Antisense rr-siRNA biogenesis strictly depends on Rdp1 and Dcr1, but not Clr4. Northern blot was performed with Ago1-associated RNAs isolated from different genetic backgrounds as indicated. The same blot was consecutively hybridized with probes specific for either centromeric dg/dh repeat–associated siRNAs (ra-siRNAs), antisense rr-siRNAs or sense rr-sRNAs.

teins. U is the preferred 5′ nucleotide of miRNAs of animals and plants (Lau et al. 2001; Reinhart et al. 2002), piRNAs of flies (Aravin et al. 2003) and mammals(Aravin et al. 2006; Lau et al. 2006; Girard et al. 2006; Grivna et al. 2006; Watanabe et al. 2006) and 21U-RNAs of worms (Ruby et al. 2006), although G is the preferred 5′ nucleotide of endogenous siRNAs of worms (Ambros et al. 2003) and A is the preferred 5′ nucleotide of the most populated class of endogenous siRNAs in plants (Rajagopalan et al. 2006).

Role of Cid14 in regulation of siRNA distribution Members of the family of noncanonical poly(A) polymerases that includes Cid14 seem to have central roles in surveillance mechanisms that monitor RNA quality. These enzymes are involved in rRNA processing, tRNA processing, snoRNA processing and the interferon response (Stevenson and Norbury 2006; Justesen et al. 2000). Furthermore, members of this family have been implicated in RNAi and siRNA biogenesis in C. elegans, S. pombe and Tetrahymena thermophila (Motamedi et al. 2004; Bühler et al. 2007),(Chen et al. 2005; Lee and Collins 2007). They are therefore likely to have a broad and ancient role in coordination of endogenous RNA quality control and the recognition of aberrant and foreign RNAs.

In addition to Cid14, another member of the fission yeast family of noncanonical poly(A) polymerases, Cid12, has previously been implicated in siRNA biogenesis (Motamedi et al. 2004). Whereas in cells lacking Cid12 cen siRNAs are absent (Motamedi et al. 2004), cen siRNA levels in cid14 cells are dramatically reduced (Bühler et al. 2007). Our results provide an explanation for this reduction in cen siRNA levels. Cid14 is a subunit of the TRAMP polyadenylation complex, which is involved in recognition and targeting of aberrant RNAs for exosomal degradation (LaCava et al. 2005; Vanácová et al. 2005). Recognition is thought to involve polyadenylation of aberrant 3′ ends by Trf4 in S. cerevisiae and Cid14 in S. pombe. Notably, Cid12 is a stable component of the RDRC complex, which is required for RNAi-mediated heterochromatin formation (Motamedi et al. 2004). Together with our present observations on the specific appearance of antisense rRNA siRNAs (rr-siRNAs) in cid14 cells, these results suggest a model for the regulation of siRNA levels from different genomic regions that involves competition between the TRAMP and RDRC complexes for RNA substrates, mediated by the two poly(A) polymerase proteins Cid12 and Cid14. In this model, Cid12 and Cid14 would have preferences for different substrates but could also act on noncanonical substrates. For example, Cid14 would normally promote the targeting of rRNA precursors or the tRNA-Glu fragment for exosomal degradation or processing. In the absence of Cid14, such precursors accumulate and become targets for RDRC, recruit RDRC away from centromeric transcripts, and thus give rise to rr-siRNAs with a concomitant decrease in cen siRNAs (Fig. 6). In support of this competition model, we also observe the emergence of antisense rr-siRNAs in clr4 cells. Clr4 is required for efficient cen siRNA generation and for the physical association of RDRC with RITS and centromeric transcripts, and localization of RDRC to centromeric DNA repeats (Motamedi et al. 2004). The release of RDRC (Cid12) from heterochromatic regions probably allows RDRC to more effectively compete for abundant rRNA precursors, even in the presence of a functional TRAMP complex.

A second possible level of competition could arise from the preference of Ago1 for small RNAs with a 5′ U, independently of RDRC and Dcr1. In this case, any small RNA with a 5′ U not degraded by TRAMP and exosome would have the potential to load onto and therefore sequester Ago1. Presumably, those sRNAs that resemble Dcr1 products in being double stranded with 2-nt 3′ overhangs would have the benefit of preferential loading into Ago1, but even single-stranded sRNAs would have some ability to be loaded into, or at least associated with, Ago1. As a result, aberrant RNAs may directly interfere with Ago1 function at centromeres, providing another possible explanation for reduced cen siRNA levels in cid14 cells. In support of this model, we find that Ago1 is associated with massive amounts of an sRNA, starting with 5′ U and matching sense to tRNA-Glu, in cid14 cells. Although this sRNA may not be functional, its sheer abundance in ago1-associated small RNAs (14% of total reads) suggests that it may directly interfere with Ago1 function at centromeres, contributing to the reduced cen siRNA levels in cid14 cells.

Gene-specific sRNAs A substantial portion (28,000, 13%) of the Ago1-associated sRNAs in this study map to genes and intergenic regions (Figs. 1 and 3, and Supplementary Table 3). In particular, we note that the sRNAs that map to intergenic regions account for a large fraction (22%) of this class. Although intergenic regions are not expected to be as highly transcribed as annotated genes, they are transcribed to some extent. A recent study suggests that extensive read-through transcription occurs at convergent gene pairs in the G1 phase of the cell cycle, giving rise to overlapping sense and antisense transcripts (Gullerova and Proudfoot 2008). Such overlapping transcripts are proposed to create a dsRNA substrate for siRNA generation by Dicer, which then leads to RITS recruitment and transient heterochromatin formation (Gullerova and Proudfoot 2008). However, Ago1-bound sRNAs do not preferentially correspond to convergent gene pairs, suggesting that siRNAs resulting from overlapping transcripts in these regions may be too rare in asynchronous fission yeast cultures to be represented above the level of background Ago1-bound gene-specific sRNAs. Finally, we note that global analyses of H3K9 methylation and RNA levels show that, for most S. pombe genes, neither H3K9 methylation nor RNA levels

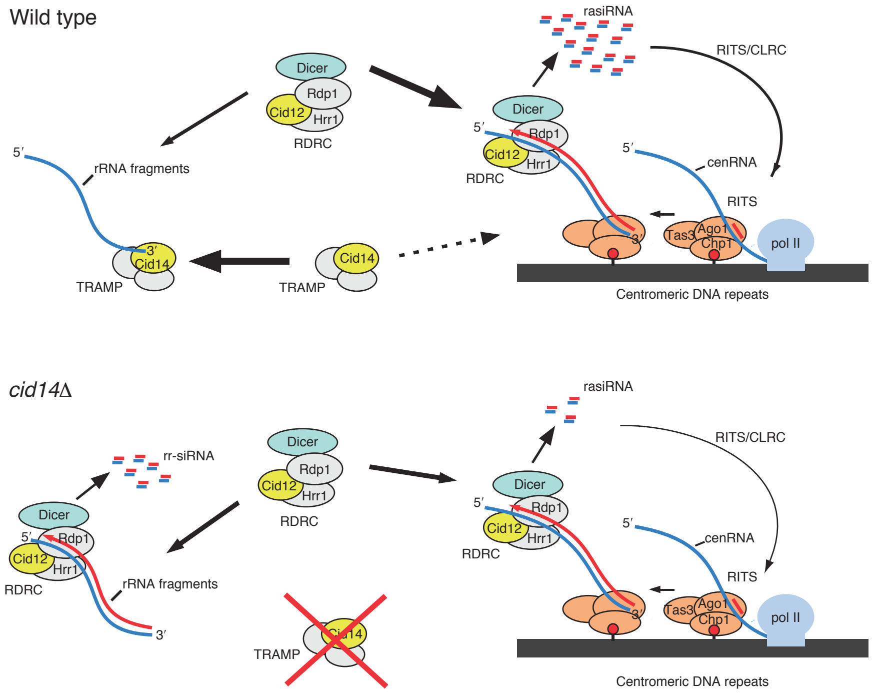


Figure 6: Model for competition between the RNAi and the Cid14–TRAMP RNA surveillance pathways. In wild-type cells, RDRC and Dicer are recruited to centromeric repeats by the RITS complex, which is tethered to chromatin via siRNA-dependent base-pairing interactions with noncoding centromeric RNA (cenRNA) and association with H3K9 methylated nucleosomes (red lollipops). This results in dsRNA synthesis and the generation of repeat-associated siRNAs (rasiRNAs), which mediate further RITS recruitment coupled to H3K9 methylation by the Clr4-containing CLRC methyltransferase complex. The TRAMP complex targets rRNA fragments for exosomal degradation. In cid14cells, rRNA fragments accumulate and become substrates for RDRC and Dicer. This titrates RDRC and Dicer away from cenRNA, resulting in the generation of rRNA-siRNAs (rr-siRNAs) and a reduction in rasiRNAs.

change substantially in RNAi mutants (Cam et al. 2005; Hansen et al. 2005). These observations suggest that the sRNAs identified in our study may act at the post-transcriptional level, but the functional relevance of the gene-specific sRNAs, if any, remains speculative and requires further investigation.

In conclusion, eukaryotes have evolved elaborate surveillance mechanisms to monitor the quality of the transcriptome. These mechanisms often involve the degradation of aberrant RNAs that lack proper processing signals. Translation-dependent mechanisms such as nonsense-mediated mRNA decay act in the cytoplasm to control the quality of open reading frames and thereby prevent the production of potentially malfunctioning proteins. The surveillance system also recognizes and degrades other types of aberrant transcripts, some of which lack the potential to be translated into protein. As we show in this study, such aberrant RNAs may have deleterious effects by interfering with the generation of endogenous siRNAs or serving as templates to generate new siRNAs with the potential to silence genetic information.

4.4 Methods

Fission yeast strains and plasmids The plasmid pREP1-3Flag-Ago1 was described previously (Buker et al. 2007). Schizosaccharomyces pombe strains used in this study are described in Supplementary Table 4 and were grown at 30°C in YEA medium (yeast extract supplemented with adenine). If transformed with pREP1-3Flag-Ago1, cells were grown at 30°C in EMMC–leu+his medium.

Generation of small RNA libraries for 454 deep sequencing Ago1-associated RNA was isolated as described previously (Bühler et al. 2007) and 20–30-nt RNAs were PAGE purified. The eluted small RNAs were cloned based upon the preactivated, adenylated linkering method described previously (Lau et al. 2001) using a mutant T4 RNA ligase (Rnl21–249)(Ho et al. 2004). Single-stranded DNA suitable to go directly into the emulsion PCR step of 454 pyrosequencing was generated as described previously (Margulies et al. 2005).

In silico analysis of sequencing data We selected 454 reads with matches to the terminal 9 nt of the $5'$ linker and the first 9 nt of the $3'$ linker, which resulted in a total of 349,477 wild-type reads and 315,701 reads in cid14. Next, we mapped reads of size 15–29 nt to the S. pombe genome, requiring a perfect match to the genome. This yielded 255,487 reads (73%) in wild-type and 240,471 reads (76%) in cid14, which we analyzed in this paper. We used the genome and annotations that were current as of 18 July 2007, available from The S. pombe Genome Project (`http://www.sanger.ac.uk/Projects/S_pombe/`). Unless otherwise noted, all read counts were normalized by the number of times the read perfectly matched the genome. The mating-type K-region was obtained from PubMed (U57841).

DNA oligonucleotides Sequences of the DNA oligonucleotides used in this study are described in Supplementary Table 5.

Northern blot analysis Ago1-associated RNAs were recovered from Flag-purified Flag-Ago1 protein and analyzed by northern blot as described previously (Bühler et al. 2007). To detect centromeric siRNAs (cen dg/dh), a mixture of oligonucleotides complementary to the siRNAs sequenced by Reinhart and Bartel (2002) were $5'$ end labeled. Sense ribosomal small RNAs (rsRNAs), antisense ribosomal siRNAs (rsiRNAs) and tRNA-Glu sRNAs were detected with labeled DNA oligonucleotides rsi1-10, rsi11-18 and mb512, respectively.

Chromatin immunoprecipitation ChIP was performed with the antibody ab1220 (abcam) as described previously (Bühler et al. 2006). Primers to amplify IRC3R and actin were mb510/511 and mb90/91, respectively. Primers to amplify dh/imr1R sequences 1–5 surrounding the tRNA-Glu gene were DM566/567, mb527/528, mb521/522, mb523/524 and mb525/526, respectively.

Accession codes Gene Expression Omnibus: small RNA sequencing data were deposited with the accession number GSE12416.

4.5 References

Allshire, RC, JP Javerzat, NJ Redhead, and G Cranston (Jan. 1994). “Position effect variegation at fission yeast centromeres.” In: *Cell* 76.1, pp. 157–69 (cit. on p. 72).

Ambros, V, RC Lee, A Lavanway, PT Williams, and D Jewell (May 2003). “MicroRNAs and other tiny endogenous RNAs in C. elegans.” In: *Curr Biol* 13.10, pp. 807–18 (cit. on pp. 70, 78).

Aravin, AA, M Lagos-Quintana, A Yalcin, M Zavolan, D Marks, B Snyder, T Gaasterland, J Meyer, and T Tuschl (Aug. 2003). “The small RNA profile during Drosophila melanogaster development.” In: *Dev Cell* 5.2, pp. 337–50 (cit. on pp. 71, 78).

Aravin, A, D Gaidatzis, S Pfeffer, M Lagos-Quintana, P Landgraf, N Iovino, P Morris, MJ Brownstein, S Kuramochi-Miyagawa, T Nakano, M Chien, JJ Russo, J Ju, R Sheridan, C Sander, M Zavolan, and T Tuschl (July 2006). "A novel class of small RNAs bind to MILI protein in mouse testes." In: *Nature* 442.7099, pp. 203–7. DOI: `10.1038/nature04916` (cit. on pp. 71, 78).

Baulcombe, D (Sept. 2004). "RNA silencing in plants." In: *Nature* 431.7006, pp. 356–63. DOI: `10.1038/nature02874` (cit. on p. 69).

Bernstein, E, AA Caudy, SM Hammond, and GJ Hannon (Jan. 2001). "Role for a bidentate ribonuclease in the initiation step of RNA interference." In: *Nature* 409.6818, pp. 363–6. DOI: `10.1038/35053110` (cit. on p. 69).

Bühler, M and D Moazed (Nov. 2007). "Transcription and RNAi in heterochromatic gene silencing." In: *Nat Struct Mol Biol* 14.11, pp. 1041–8. DOI: `10.1038/nsmb1315` (cit. on p. 69).

Bühler, M, A Verdel, and D Moazed (June 2006). "Tethering RITS to a nascent transcript initiates RNAi- and heterochromatin-dependent gene silencing." In: *Cell* 125.5, pp. 873–86. DOI: `10.1016/j.cell.2006.04.025` (cit. on pp. 69, 77, 81).

Bühler, M, W Haas, SP Gygi, and D Moazed (May 2007). "RNAi-dependent and -independent RNA turnover mechanisms contribute to heterochromatic gene silencing." In: *Cell* 129.4, pp. 707–21. DOI: `10.1016/j.cell.2007.03.038` (cit. on pp. 69, 70, 73, 78, 81).

Buker, SM, T Iida, M Bühler, J Villén, SP Gygi, JI Nakayama, and D Moazed (Mar. 2007). "Two different Argonaute complexes are required for siRNA generation and heterochromatin assembly in fission yeast." In: *Nat Struct Mol Biol* 14.3, pp. 200–7. DOI: `10.1038/nsmb1211` (cit. on pp. 69, 72, 80).

Cam, HP, T Sugiyama, ES Chen, X Chen, PC FitzGerald, and SIS Grewal (Aug. 2005). "Comprehensive analysis of heterochromatin- and RNAi-mediated epigenetic control of the fission yeast genome." In: *Nat Genet* 37.8, pp. 809–19. DOI: `10.1038/ng1602` (cit. on pp. 69, 70, 73–75, 77, 80).

Chen, CCG, MJ Simard, H Tabara, DR Brownell, JA McCollough, and CC Mello (Feb. 2005). "A member of the polymerase beta nucleotidyltransferase superfamily is required for RNA interference in C. elegans." In: *Curr Biol* 15.4, pp. 378–83. DOI: `10.1016/j.cub.2005.01.009` (cit. on p. 78).

Chen, ES, K Zhang, E Nicolas, HP Cam, M Zofall, and SIS Grewal (Feb. 2008). "Cell cycle control of centromeric repeat transcription and heterochromatin assembly." In: *Nature* 451.7179, pp. 734–7. DOI: `10.1038/nature06561` (cit. on p. 69).

Elbashir, SM, W Lendeckel, and T Tuschl (Jan. 2001). "RNA interference is mediated by 21- and 22-nucleotide RNAs." In: *Genes Dev* 15.2, pp. 188–200 (cit. on p. 69).

Fire, A, S Xu, MK Montgomery, SA Kostas, SE Driver, and CC Mello (Feb. 1998). "Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans." In: *Nature* 391.6669, pp. 806–11. DOI: `10.1038/35888` (cit. on p. 69).

Girard, A, R Sachidanandam, GJ Hannon, and MA Carmell (July 2006). "A germline-specific class of small RNAs binds mammalian Piwi proteins." In: *Nature* 442.7099, pp. 199–202. DOI: `10.1038/nature04917` (cit. on pp. 71, 78).

Good, L, RV Intine, and RN Nazar (July 1997). “The ribosomal-RNA-processing pathway in Schizosaccharomyces pombe.” In: *Eur J Biochem* 247.1, pp. 314–21 (cit. on pp. 77, 78).

Grivna, ST, E Beyret, Z Wang, and H Lin (July 2006). “A novel class of small RNAs in mouse spermatogenic cells.” In: *Genes Dev* 20.13, pp. 1709–14. DOI: 10.1101/gad.1434406 (cit. on pp. 71, 78).

Gullerova, M and NJ Proudfoot (Mar. 2008). “Cohesin complex promotes transcriptional termination between convergent genes in S. pombe.” In: *Cell* 132.6, pp. 983–95. DOI: 10.1016/j.cell.2008.02.040 (cit. on p. 79).

Hamilton, AJ and DC Baulcombe (Oct. 1999). “A species of small antisense RNA in posttranscriptional gene silencing in plants.” In: *Science* 286.5441, pp. 950–2 (cit. on p. 69).

Hammond, SM, E Bernstein, D Beach, and GJ Hannon (Mar. 2000). “An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells.” In: *Nature* 404.6775, pp. 293–6. DOI: 10.1038/35005107 (cit. on p. 69).

Hannon, GJ (July 2002). “RNA interference.” In: *Nature* 418.6894, pp. 244–51. DOI: 10.1038/418244a (cit. on p. 69).

Hansen, KR, G Burns, J Mata, TA Volpe, RA Martienssen, J Bähler, and G Thon (Jan. 2005). “Global effects on gene expression in fission yeast by silencing and RNA interference machineries.” In: *Mol Cell Biol* 25.2, pp. 590–601. DOI: 10.1128/MCB.25.2.590-601.2005 (cit. on p. 80).

Ho, CK, LK Wang, CD Lima, and S Shuman (Feb. 2004). “Structure and mechanism of RNA ligase.” In: *Structure* 12.2, pp. 327–39. DOI: 10.1016/j.str.2004.01.011 (cit. on p. 81).

Hong, EJE, J Villén, EL Gerace, SP Gygi, and D Moazed (2005). “A cullin E3 ubiquitin ligase complex associates with Rik1 and the Clr4 histone H3-K9 methyltransferase and is required for RNAi-mediated heterochromatin formation.” In: *RNA Biol* 2.3, pp. 106–11 (cit. on p. 77).

Irvine, DV, M Zaratiegui, NH Tolia, DB Goto, DH Chitwood, MW Vaughn, L Joshua-Tor, and RA Martienssen (Aug. 2006). “Argonaute slicing is required for heterochromatic silencing and spreading.” In: *Science* 313.5790, pp. 1134–7. DOI: 10.1126/science.1128813 (cit. on p. 72).

Justesen, J, R Hartmann, and NO Kjeldgaard (Oct. 2000). “Gene structure and function of the 2’-5’-oligoadenylate synthetase family.” In: *Cell Mol Life Sci* 57.11, pp. 1593–612 (cit. on p. 78).

LaCava, J, J Houseley, C Saveanu, E Petfalski, E Thompson, A Jacquier, and D Tollervey (June 2005). “RNA degradation by the exosome is promoted by a nuclear polyadenylation complex.” In: *Cell* 121.5, pp. 713–24. DOI: 10.1016/j.cell.2005.04.029 (cit. on pp. 69, 70, 73, 79).

Lau, NC, LP Lim, EG Weinstein, and DP Bartel (Oct. 2001). “An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans.” In: *Science* 294.5543, pp. 858–62. DOI: 10.1126/science.1065062 (cit. on pp. 71, 78, 81).

Lau, NC, AG Seto, J Kim, S Kuramochi-Miyagawa, T Nakano, DP Bartel, and RE Kingston (July 2006). “Characterization of the piRNA complex from rat testes.” In: *Science* 313.5785, pp. 363–7. DOI: 10.1126/science.1130164 (cit. on pp. 71, 78).

Lee, SR and K Collins (July 2007). “Physical and functional coupling of RNA-dependent RNA polymerase and Dicer in the biogenesis of endogenous siRNAs.” In: *Nat Struct Mol Biol* 14.7, pp. 604–10. DOI: 10.1038/nsmb1262 (cit. on p. 78).

Li, F, DB Goto, M Zaratiegui, X Tang, R Martienssen, and WZ Cande (Aug. 2005). “Two novel proteins, dos1 and dos2, interact with rik1 to regulate heterochromatic RNA interference and histone modification.” In: *Curr Biol* 15.16, pp. 1448–57. DOI: 10.1016/j.cub.2005.07.021 (cit. on p. 77).

Margulies, M et al. (Sept. 2005). “Genome sequencing in microfabricated high-density picolitre reactors.” In: *Nature* 437.7057, pp. 376–80. DOI: 10.1038/nature03959 (cit. on pp. 70, 81).

Matranga, C, Y Tomari, C Shin, DP Bartel, and PD Zamore (Nov. 2005). “Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes.” In: *Cell* 123.4, pp. 607–20. DOI: 10.1016/j.cell.2005.08.044 (cit. on p. 72).

Motamedi, MR, A Verdel, SU Colmenares, SA Gerber, SP Gygi, and D Moazed (Dec. 2004). “Two RNAi complexes, RITS and RDRC, physically interact and localize to noncoding centromeric RNAs.” In: *Cell* 119.6, pp. 789–802. DOI: 10.1016/j.cell.2004.11.034 (cit. on pp. 69, 77–79).

Rajagopalan, R, H Vaucheret, J Trejo, and DP Bartel (Dec. 2006). “A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana.” In: *Genes Dev* 20.24, pp. 3407–25. DOI: 10.1101/gad.1476406 (cit. on pp. 70–72, 78).

Rand, TA, S Petersen, F Du, and X Wang (Nov. 2005). “Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation.” In: *Cell* 123.4, pp. 621–9. DOI: 10.1016/j.cell.2005.10.020 (cit. on p. 72).

Reinhart, BJ and DP Bartel (Sept. 2002). “Small RNAs correspond to centromere heterochromatic repeats.” In: *Science* 297.5588, p. 1831. DOI: 10.1126/science.1077183 (cit. on pp. 69, 81).

Reinhart, BJ, EG Weinstein, MW Rhoades, B Bartel, and DP Bartel (July 2002). “MicroRNAs in plants.” In: *Genes Dev* 16.13, pp. 1616–26. DOI: 10.1101/gad.1004402 (cit. on pp. 71, 78).

Ruby, JG, C Jan, C Player, MJ Axtell, W Lee, C Nusbaum, H Ge, and DP Bartel (Dec. 2006). “Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans.” In: *Cell* 127.6, pp. 1193–207. DOI: 10.1016/j.cell.2006.10.040 (cit. on pp. 70, 71, 78).

Sadaie, M, T Iida, T Urano, and JI Nakayama (Oct. 2004). “A chromodomain protein, Chp1, is required for the establishment of heterochromatin in fission yeast.” In: *EMBO J* 23.19, pp. 3825–35. DOI: 10.1038/sj.emboj.7600401 (cit. on p. 69).

Schwarz, DS, G Hutvágner, T Du, Z Xu, N Aronin, and PD Zamore (Oct. 2003). “Asymmetry in the assembly of the RNAi enzyme complex.” In: *Cell* 115.2, pp. 199–208 (cit. on p. 72).

Scott, KC, SL Merrett, and HF Willard (Jan. 2006). “A heterochromatin barrier partitions the fission yeast centromere into discrete chromatin domains.” In: *Curr Biol* 16.2, pp. 119–29. DOI: 10.1016/j.cub.2005.11.065 (cit. on p. 73).

Sijen, T, J Fleenor, F Simmer, KL Thijssen, S Parrish, L Timmons, RH Plasterk, and A Fire (Nov. 2001). “On the role of RNA amplification in dsRNA-triggered gene silencing.” In: *Cell* 107.4, pp. 465–76 (cit. on pp. 69, 76).

Stevenson, AL and CJ Norbury (Oct. 2006). "The Cid1 family of non-canonical poly(A) polymerases." In: *Yeast* 23.13, pp. 991–1000. DOI: 10.1002/yea.1408 (cit. on pp. 69, 78).

Vanácová, S, J Wolf, G Martin, D Blank, S Dettwiler, A Friedlein, H Langen, G Keith, and W Keller (June 2005). "A new yeast poly(A) polymerase complex involved in RNA quality control." In: *PLoS Biol* 3.6, e189. DOI: 10.1371/journal.pbio.0030189 (cit. on pp. 70, 73, 79).

Verdel, A and D Moazed (Oct. 2005). "RNAi-directed assembly of heterochromatin in fission yeast." In: *FEBS Lett* 579.26, pp. 5872–8. DOI: 10.1016/j.febslet.2005.08.083 (cit. on p. 77).

Verdel, A, S Jia, S Gerber, T Sugiyama, S Gygi, SIS Grewal, and D Moazed (Jan. 2004). "RNAi-mediated targeting of heterochromatin by the RITS complex." In: *Science* 303.5658, pp. 672–6. DOI: 10.1126/science.1093686 (cit. on pp. 69, 77).

Volpe, TA, C Kidner, IM Hall, G Teng, SIS Grewal, and RA Martienssen (Sept. 2002). "Regulation of heterochromatic silencing and histone H3 lysine-9 methylation by RNAi." In: *Science* 297.5588, pp. 1833–7. DOI: 10.1126/science.1074973 (cit. on p. 69).

Wang, SW, AL Stevenson, SE Kearsey, S Watt, and J Bähler (Jan. 2008). "Global role for polyadenylation-assisted nuclear RNA degradation in posttranscriptional gene silencing." In: *Mol Cell Biol* 28.2, pp. 656–65. DOI: 10.1128/MCB.01531-07 (cit. on p. 77).

Watanabe, T, A Takeda, T Tsukiyama, K Mise, T Okuno, H Sasaki, N Minami, and H Imai (July 2006). "Identification and characterization of two novel classes of small RNAs in the mouse germline: retrotransposon-derived siRNAs in oocytes and germline small RNAs in testes." In: *Genes Dev* 20.13, pp. 1732–43. DOI: 10.1101/gad.1425706 (cit. on pp. 71, 78).

Win, TZ, S Draper, RL Read, J Pearce, CJ Norbury, and SW Wang (Mar. 2006). "Requirement of fission yeast Cid14 in polyadenylation of rRNAs." In: *Mol Cell Biol* 26.5, pp. 1710–21. DOI: 10.1128/MCB.26.5.1710-1721.2006 (cit. on pp. 69, 70, 73).

Wyers, F, M Rougemaille, G Badis, JC Rousselle, ME Dufour, J Boulay, B Régnault, F Devaux, A Namane, B Séraphin, D Libri, and A Jacquier (June 2005). "Cryptic pol II transcripts are degraded by a nuclear quality control pathway involving a new poly(A) polymerase." In: *Cell* 121.5, pp. 725–37. DOI: 10.1016/j.cell.2005.04.030 (cit. on pp. 70, 73).

Zamore, PD, T Tuschl, PA Sharp, and DP Bartel (Mar. 2000). "RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals." In: *Cell* 101.1, pp. 25–33. DOI: 10.1016/S0092-8674(00)80620-0 (cit. on p. 69).

Zaratiegui, M, DV Irvine, and RA Martienssen (Feb. 2007). "Noncoding RNAs and gene silencing." In: *Cell* 128.4, pp. 763–76. DOI: 10.1016/j.cell.2007.02.016 (cit. on p. 69).

Chapter 5

Conclusion

5.1 Summary and Progress

In the second chapter, we demonstrated a marked enrichment for nucleosomes on internal exonic DNA, rivaling the nucleosome enrichment at the +1 position downstream of the transcription start site of expressed genes. This signal appears to be mostly encoded at the sequence level by the high GC-content of exonic regions compared to neighboring intronic regions. Nucleosome positioning on exons is potentially a novel readout of the differential nucleotide content of exonic regions that could be used by the cellular machinery to improve recognition of bona fide splice sites. In support of a role for nucleosomes in marking true exons, we found a small but highly significant improvement in the accuracy of a splicing simulation, which takes into account only known cis regulatory motifs. This significant improvement persists even after taking into account known exonic splicing enhancers and silencers. Additionally, the nucleosome enrichment is highest for exons furthest away from neighboring exons, as well as for those exons with the weakest consensus splice site sequences, in favor of a role for nucleosome enrichment in compensating for weak cis regulatory signals in exon recognition. Finally, we identified a number of histone methyl marks which were considerably enriched beyond the average nucleosome level on exons. These results build a case for widespread interplay between chromatin state and the splicing machinery, and emphasize the co-transcriptional nature of mRNA processing events including not only splicng but also $3'$ end formation.

In the third chapter, we explored the functional consequences of tandem $3'$ UTR regulation. We showed that 3P-Seq is not only a highly specific method for annotating poly(A) sites but also functions as a reasonably quantitative measure of $3'$ UTR isoform usage. Although 3P-Seq quantitation appears to be somewhat less exact compared to RNA-Seq, we suggest this may be a general problem for $3'$ UTR isoform quantitation. Despite these issues, we were able to demonstrate a widespread shift in mRNA stability between tandem $3'$ UTR isoforms, with the shorter isoforms globally more stable than the longer isoforms. We were able to attribute these differences in part to the presence of microRNA target sites and PUF-binding motifs in the $3'$ UTR extension region that differs between tandem UTR isoforms. We also showed an inverse correlation between $3'$ UTR sequence conservation and the stability of the mRNA, suggesting that cis regulatory motifs play a more integral role in destabilizing mRNAs rather than stabilizing them.

In the fourth chapter, we characterized the centromere-derived siRNAs that are integral to formation of heterochromatin at the centromeres. We noted that the vast majority of siRNAs ($> 98\%$) begin with a $5'$ U, a marked bias that proved useful in distinguishing bona fide siRNAs from degradation products abnormally associated with the overexpressed and tagged Argonaute protein. Our results demonstrate that tight regulation of nuclear RNA species is important for the functionality of processes including centromeric heterochromatin maintenance and ribosomal and transfer RNA processing. Mutants in the exosome, which is involved in tRNA/rRNA processing, show an overabundance of aberrant tRNAs and rRNAs which can be processed by the RNA-dependent RNA-polymerase and Dicer to produce ribosomal siRNAs and severely diminishing the number of centromeric siRNAs. We were unable to determine whether these ribosomal siRNAs were able to modify the chromatin state of the tandem ribosomal DNA repeats.

5.2 Themes and Perspectives

High-throughput sequencing As I began my first project in graduate school, described in chapter 4, the use of high-throughput sequencing was just becoming widespread. These new sequencing technologies, including Illumina and 454 sequencing, enabled a new level of genome-wide discovery of novel and known functional classes of genes, isoforms, microRNAs and regulatory elements, etc. The sequencing of 250,000 Argonaute-associated small RNA reads allowed us to fully characterize the extent of repeat silencing in fission yeast.

My subsequent project, was based around early ChIP-Seq data of over 20 histone modifications in human T cells. The ChIP-Seq data published by the Zhao lab at NIH (Barski et al. 2007; Schones et al. 2008) was designed as a genome-wide survey of histone placement in the genome with the goal of increasing our understanding of chromati-level regulatory elements. Using these data, we were able to connect the chromatin state to the exonic splicing machinery. Importantly, we would have been unlikely to draw this connection were it not for the increased depth of the data, since nucleosome positioning is quite noisy for any given genomic locus. Only with ChIP-Seq data for 180,000 exons were we able to confidently determine the bias for nucleosomes on exonic DNA.

Over the last few years, high-throughput sequencing has become a fairly routine tool for genome-wide quantitation, in large part replacing the micro-array for this purpose. This has enabled work such as that presented in chapter 3, involving genome-wide $3'$ UTR isoform-specific quantitation of mRNA half-lives. As high-throughput sequencing becomes quotidian, we should continue to take advantage of its dual strengths for both quantitation and discovery of new phenomena. It is important to look for familiar things in new places (h/t R. Friedman), as we were able to do with the histone data when we shifted our view to the interior of genes, rather than the traditional transcription start site- and enhancer-focused view.

Interplay and cross-talk between regulatory step A major theme of the work presented in this thesis is the amount of overlap and interaction between gene regulatory processes. This may involve directly coupled mechanisms, as between RNA interference and chromatin modifications in fission yeast or as might be hypothesized for histone modifications and splicing. Or, this could involve indirectly coupled processes, such as the cytoplasmic regulation of mRNA stability that depends on nuclear alternative cleavage and polyadneylation events.

These results remind us of the importance of looking at the collected inputs and outputs of regulatory systems we study. This means considering not only regulatory events likely to be in close physical proximity to one another within the cell (and hence possibly directly linked) but also the upstream and downstream events which may be indirectly but dramatically affected.

5.3 Future directions

Interplay of chromatin modification and splicing Since the publication of the work presented in chapter 2, as well as a number of concurrent papers with similar results, considerable interest has been shown in elucidating the relationship between chromatin modification and the splicing process. Given the additional enrichment for H3K36me3 on exons above the nucleosome level, an obvious follow-on experiment was to perturb the methyltransferase involved in laying down this mark. Luco et al. (2010) found differential recruitment of splicing factors following knock-down and overexpression of the Set2 methyltransferase. It will be interesting to explore the other differentially enriched and depleted exonic chromatin marks, as well as to uncover biologically relevant uses of chromatin to modify splicing patterns or vice versa.

Quantitative modeling of gene regulation The advent of micro-arrays and the rise of high-throughput sequencing have enabled researchers to predict the genome-wide effects of various regulatory mechanisms, including transcription factors and microRNAs. It is just now becoming feasible to build models that integrate these data into a quantitative prediction of tissue-specific regulation of a specific regulatory step such as splicing (Barash et al. 2010). Other models are working to tease apart the quantitative contributions of mRNA and protein production and degradation (Schwanhäusser et al. 2011). The work I present in chapter 3 is a step towards such a mechanistic understanding of the contribution of destabilizing (mostly) and stabilizing (a little) cis-regulatory motifs found in $3'$ UTRs.

These models serve two imortant purposes. First, they allow researchers to simulate perturbations and predict their outcomes prior to performing lengthy and costly experiments. Second, and perhaps more importantly, they allow us to quantitate the importance of known regulatory mechanisms, and estimate the contributions of unknown mechanisms. With this information in hand, we can guide the direction of the gene regulation field in directions most likely to yield fruitful and meaningful results.

5.4 References

Barash, Y, JA Calarco, W Gao, Q Pan, X Wang, O Shai, BJ Blencowe, and BJ Frey (May 2010). “Deciphering the splicing code.” In: *Nature* 465.7294, pp. 53–9. DOI: 10.1038/nature09000 (cit. on p. 90).

Barski, A, S Cuddapah, K Cui, TY Roh, DE Schones, Z Wang, G Wei, I Chepelev, and K Zhao (May 2007). “High-resolution profiling of histone methylations in the human genome.” In: *Cell* 129.4, pp. 823–37. DOI: 10.1016/j.cell.2007.05.009 (cit. on p. 89).

Luco, RF, Q Pan, K Tominaga, BJ Blencowe, OM Pereira-Smith, and T Misteli (Feb. 2010). “Regulation of alternative splicing by histone modifications.” In: *Science* 327.5968, pp. 996–1000. DOI: 10.1126/science.1184208 (cit. on p. 89).

Schones, DE, K Cui, S Cuddapah, TY Roh, A Barski, Z Wang, G Wei, and K Zhao (Mar. 2008). “Dynamic regulation of nucleosome positioning in the human genome.” In: *Cell* 132.5, pp. 887–98. DOI: 10.1016/j.cell.2008.02.022 (cit. on p. 89).

Schwanhäusser, B, D Busse, N Li, G Dittmar, J Schuchhardt, J Wolf, W Chen, and M Selbach (May 2011). “Global quantification of mammalian gene expression control.” In: *Nature* 473.7347, pp. 337–42. DOI: 10.1038/nature10098 (cit. on p. 90).

Appendix A

Supplementary information for Chapter 2: Biased Chromatin Signatures around Polyadenylation Sites and Exons

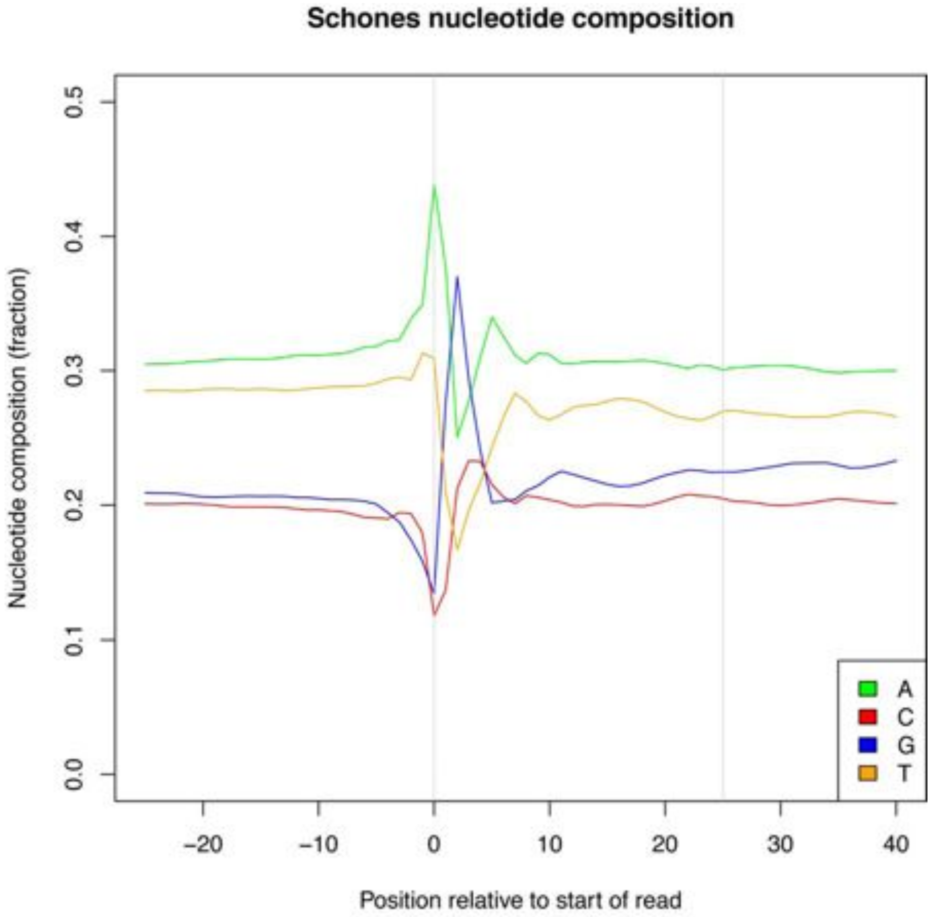


Figure S1. ChIP-seq reads show biased 5'-end nucleotide composition.

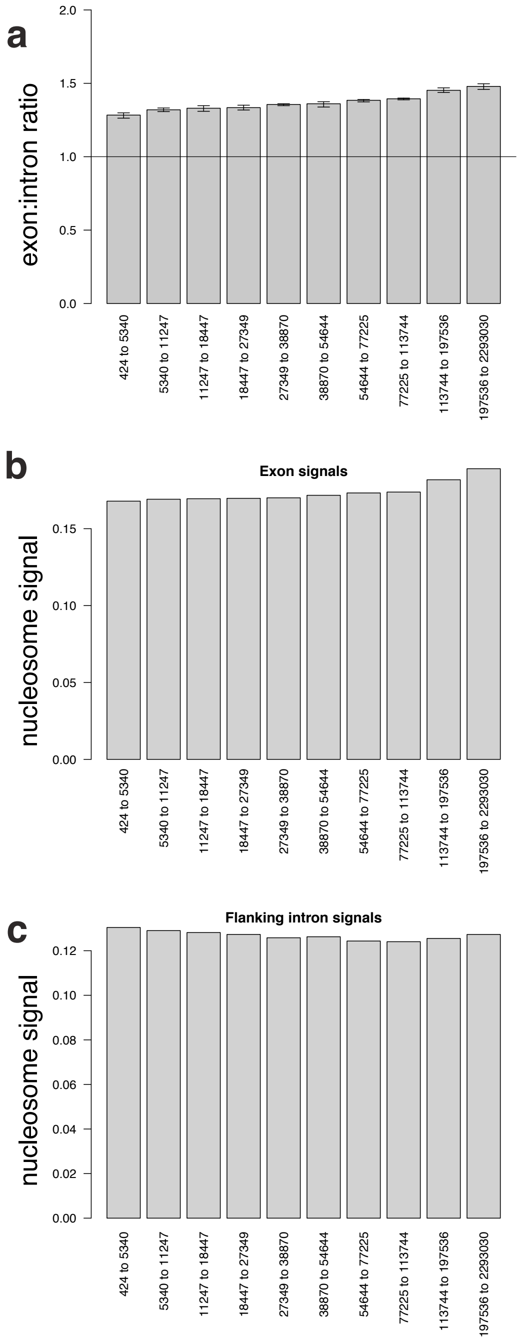


Figure S2. Nucleosome enrichment increases at greater distances from the TSS (a), although nucleosome density in exons (b) and their flanking intronic regions (c) do not consistently increase in direct correlation with distance to TSS. Error bars are 95% confidence intervals (resampling).

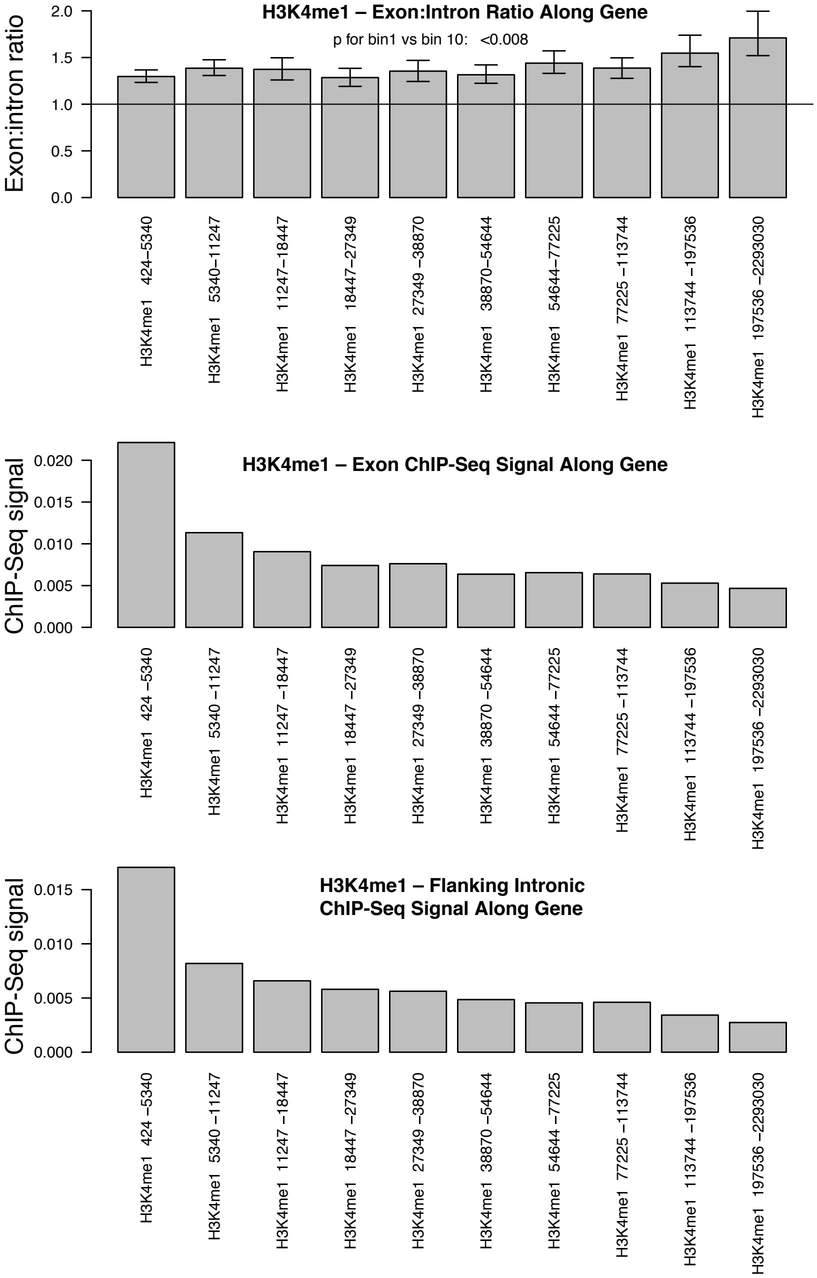


Figure S3. (A) H3K4me1 levels are highest near the TSS, but exon enrichment increases at greater distances to the TSS. Exons were placed into 10 equally-sized bins based on their 5'ss distance from the TSS. Top panel shows exon:intron enrichment values as in Fig. 2A, middle panel shows average ChIP-Seq signal across exons and bottom panel shows average ChIP-Seq signal in intronic regions flanking exons in given bin. Error bars are 95% confidence intervals. p-value comparing first and last bins, and confidence intervals, by resampling.

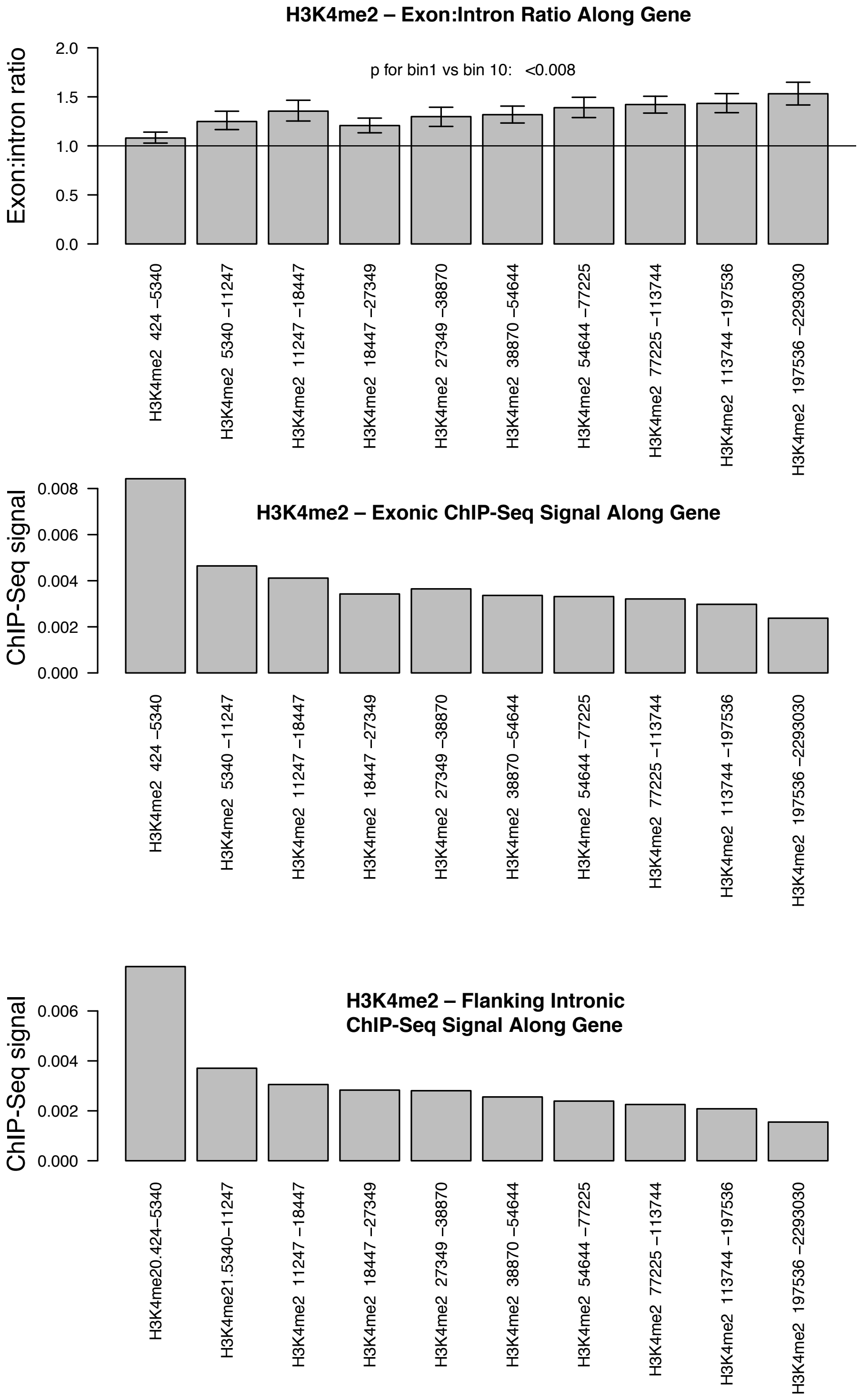


Figure S3. (B) H3K4me2 levels are highest near the TSS, but exon enrichment increases at greater distances to the TSS.

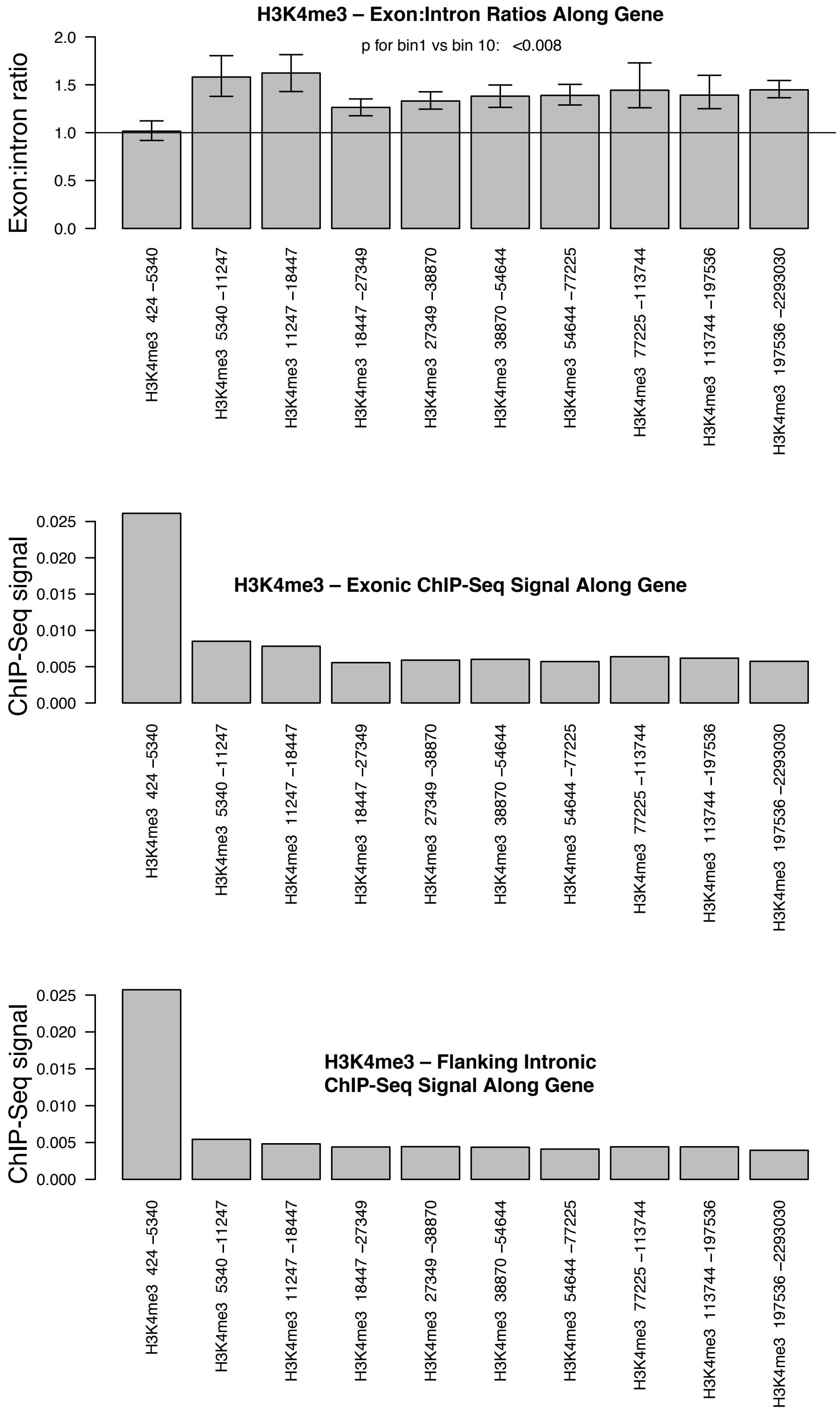


Figure S3. (C) H3K4me3 levels are highest near the TSS, but exon enrichment increases at greater distances to the TSS.

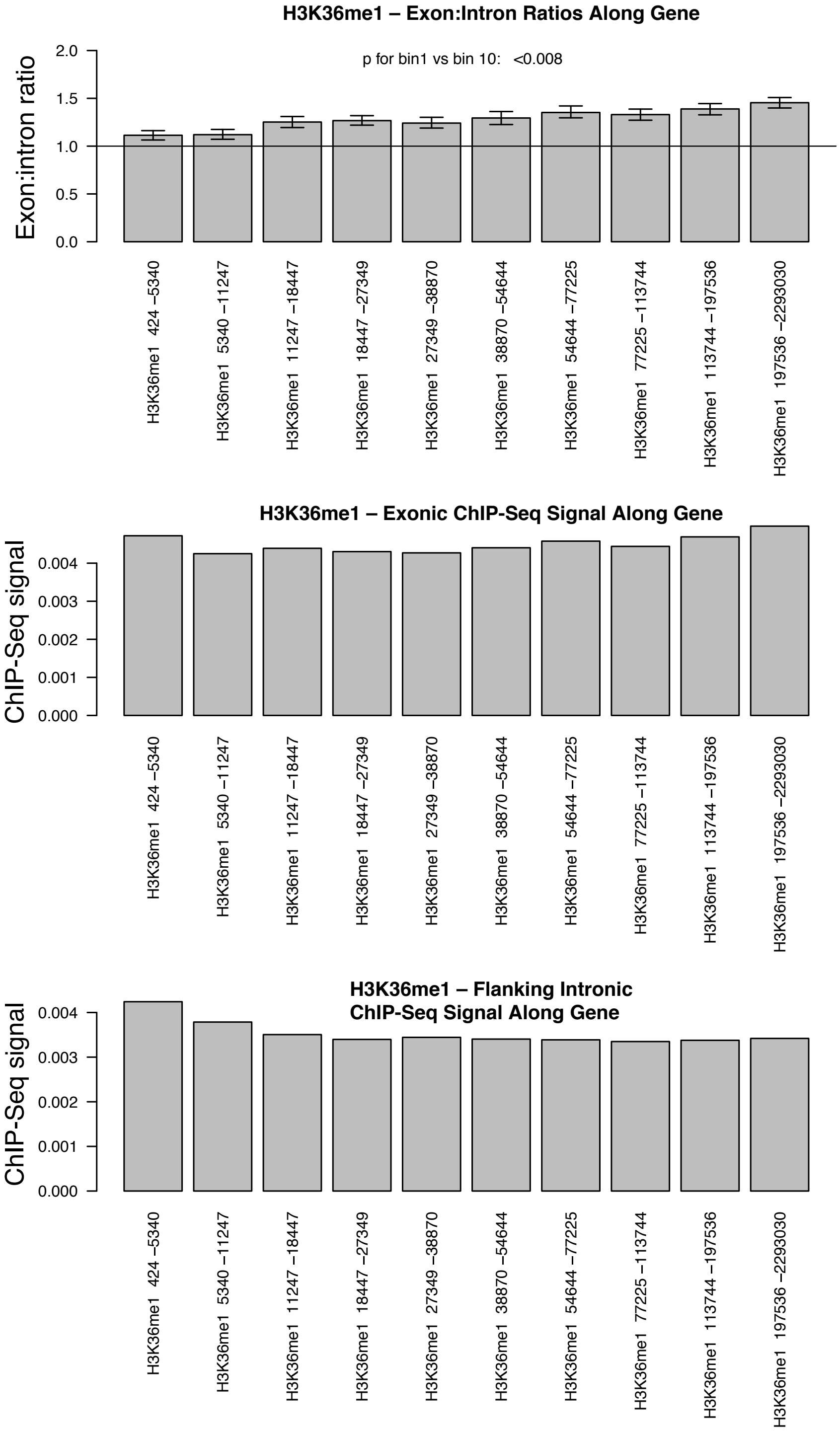
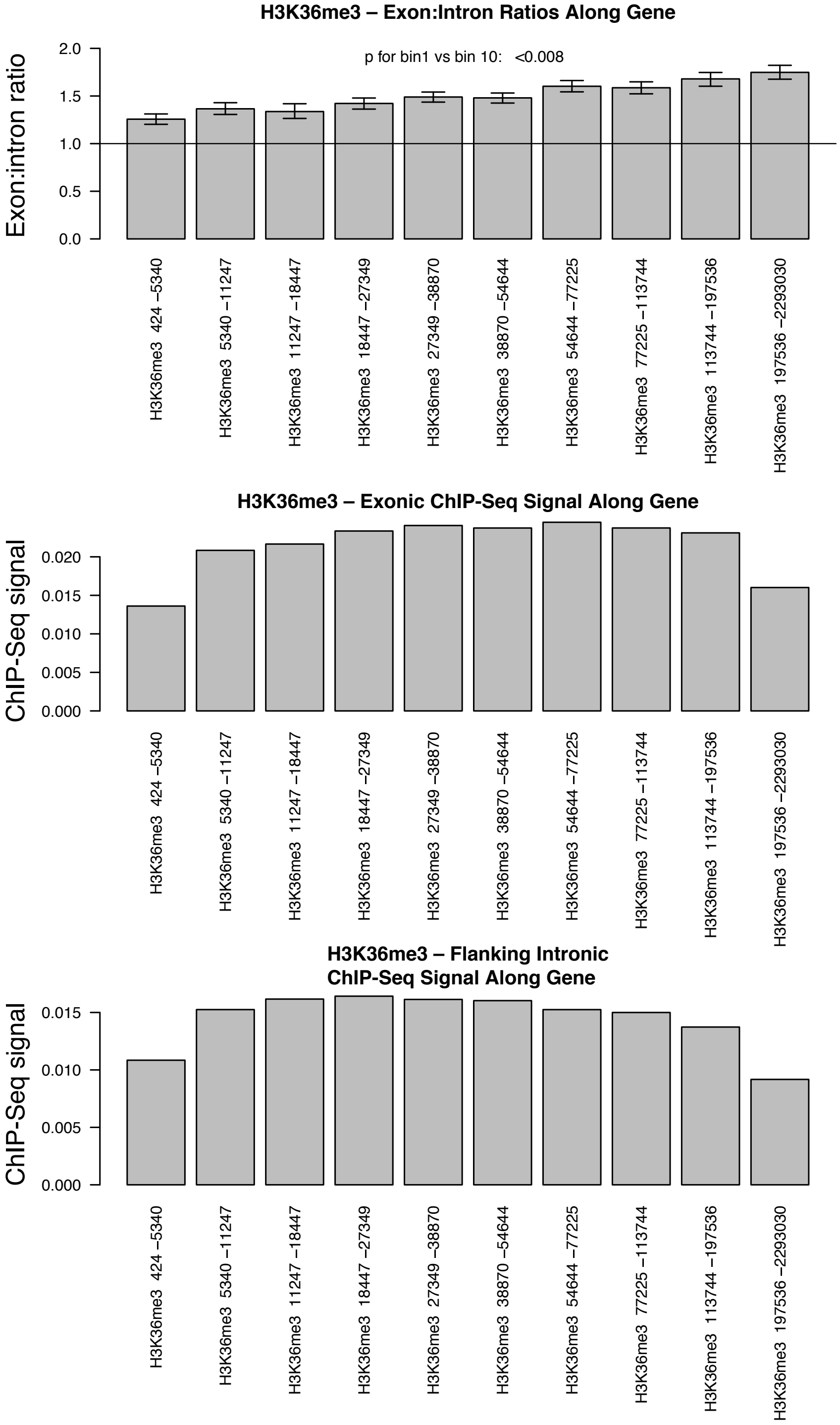


Figure S3. (D) H3K36me1 levels increase at greater distances to the TSS.



H3K36me3 – Exon:Intron Ratios Along Gene

H3K36me3 – Exonic ChIP-Seq Signal Along Gene

H3K36me3 – Flanking Intronic ChIP-Seq Signal Along Gene

Figure S3. (E) H3K36me3 levels are highest far from the TSS.

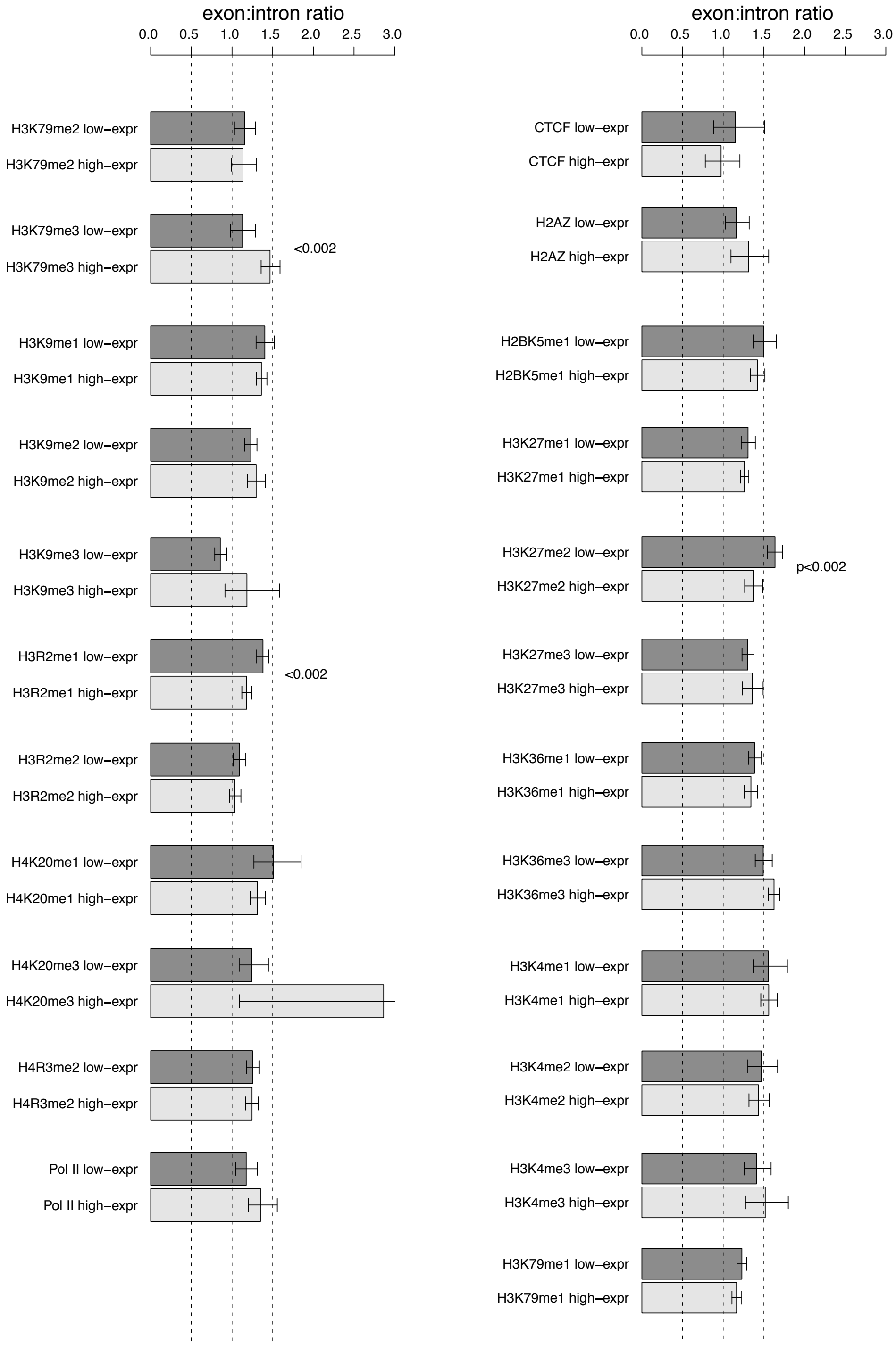


Figure S4. Most marks show consistent exon:intron ratios in lowly and highly expressed genes. See also figure 2b. 95% confidence error bars and p-values (indicated where significant) by resampling.

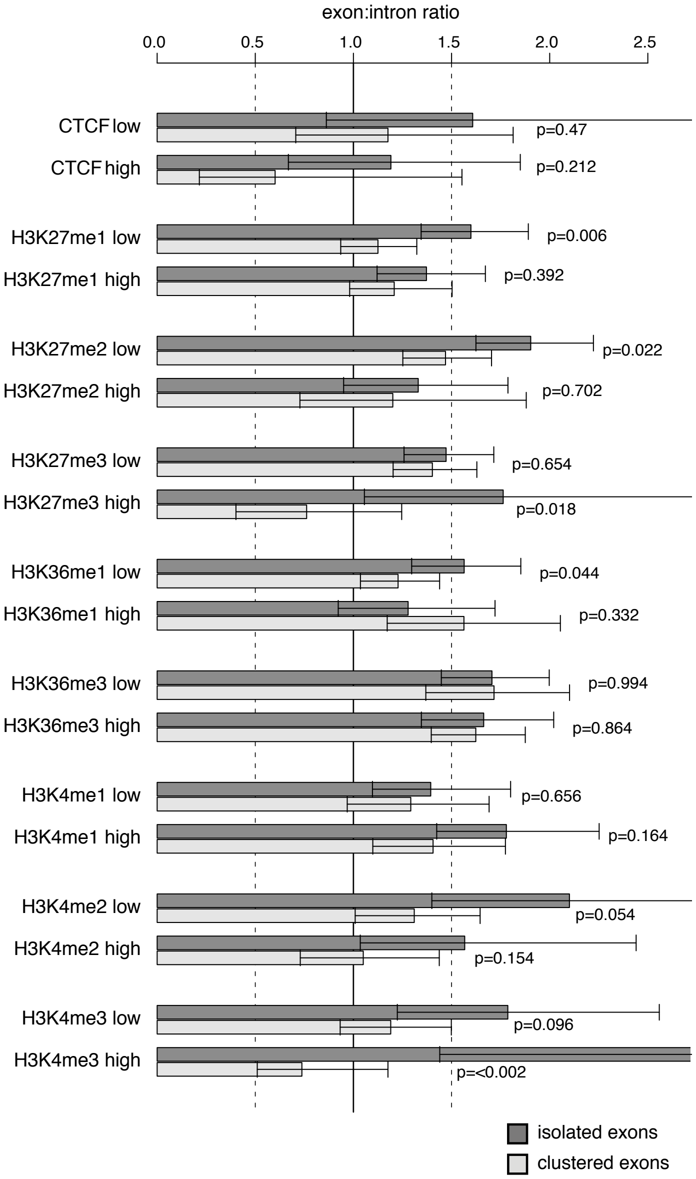


Figure S5. H3K4me3 shows enrichment in isolated exons particularly in highly expressed genes (highly and lowly expressed genes and isolated and clustered exons are defined in the methods). Error bars and 95% confidence intervals by resampling.

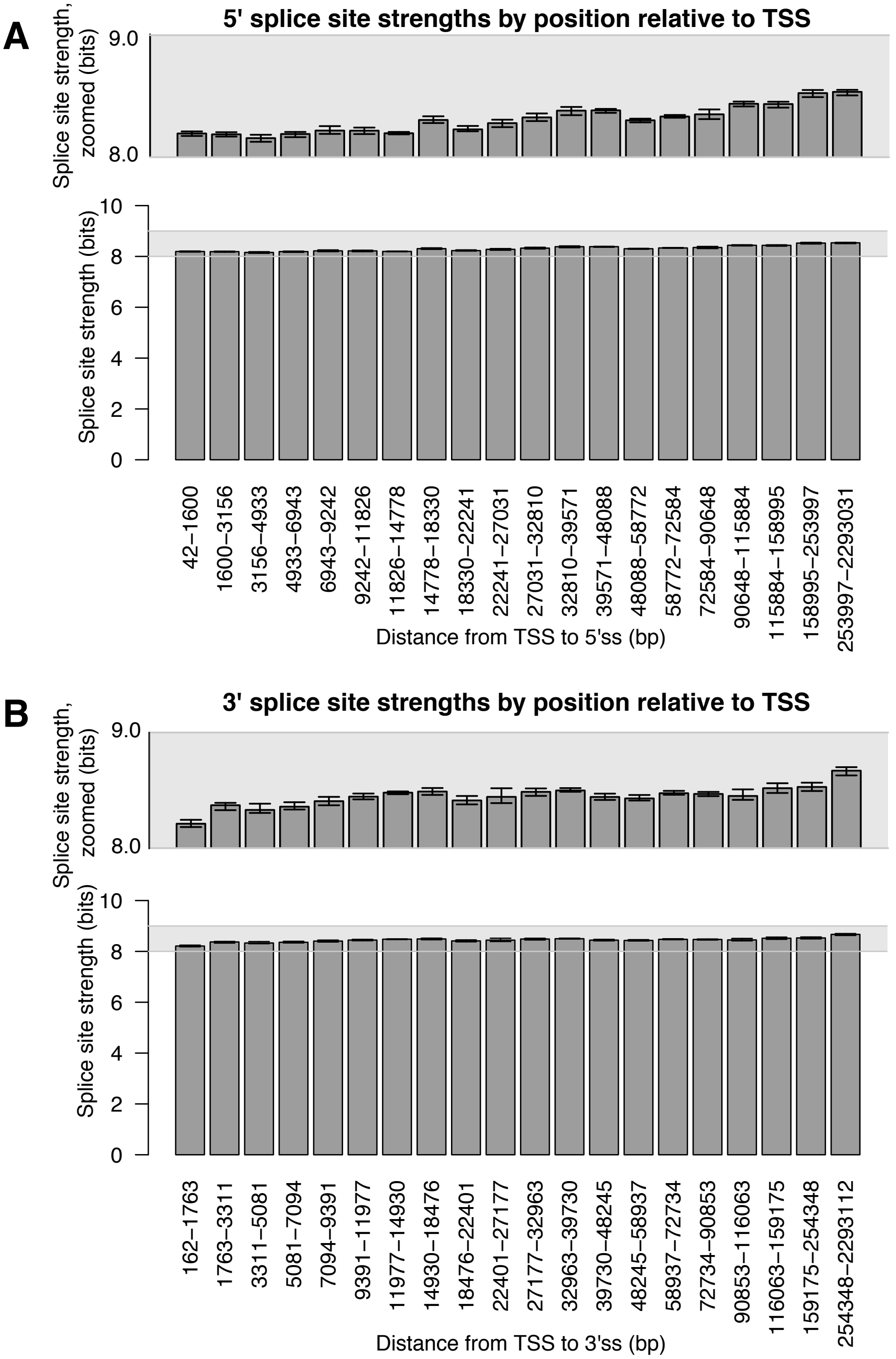


Figure S6. 5' splice site strength increases with distance to TSS (A), although this increase is slight, from ~8.25 to ~8.5 (top panel). 3' splice site strengths increase similarly (B).

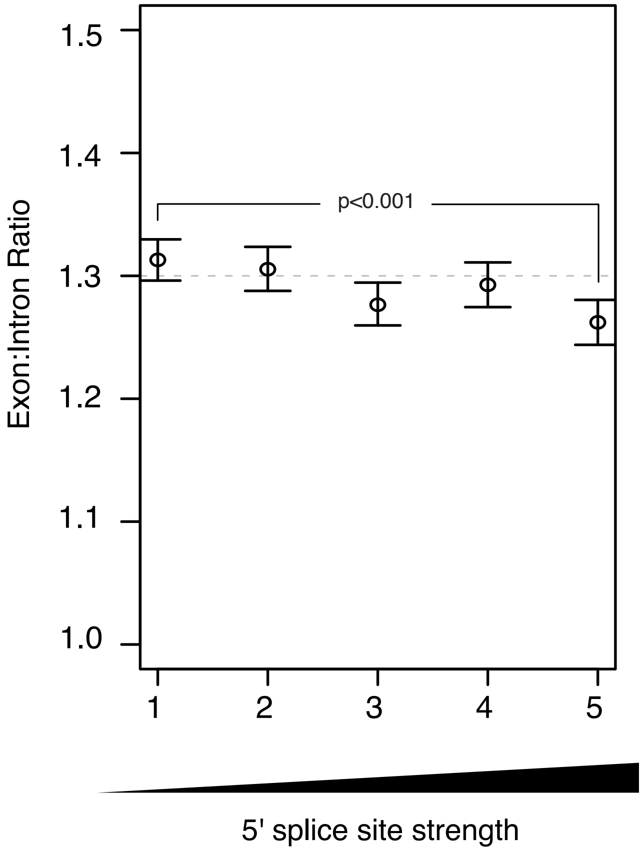


Figure S7. Nucleosome enrichment on exons is inversely correlated with 5' splice site strength. As in Fig. 3c, exons with non-negative splice site strength scores were binned into five equally sized bins (x-axis) and average nucleosome exon:intron ratios were calculated. Error bars indicate 95% confidence intervals. CIs and p-value comparing weakest ss strength bin (bin 1) to strongest ss strength bin (bin 5) were calculated by resampling.

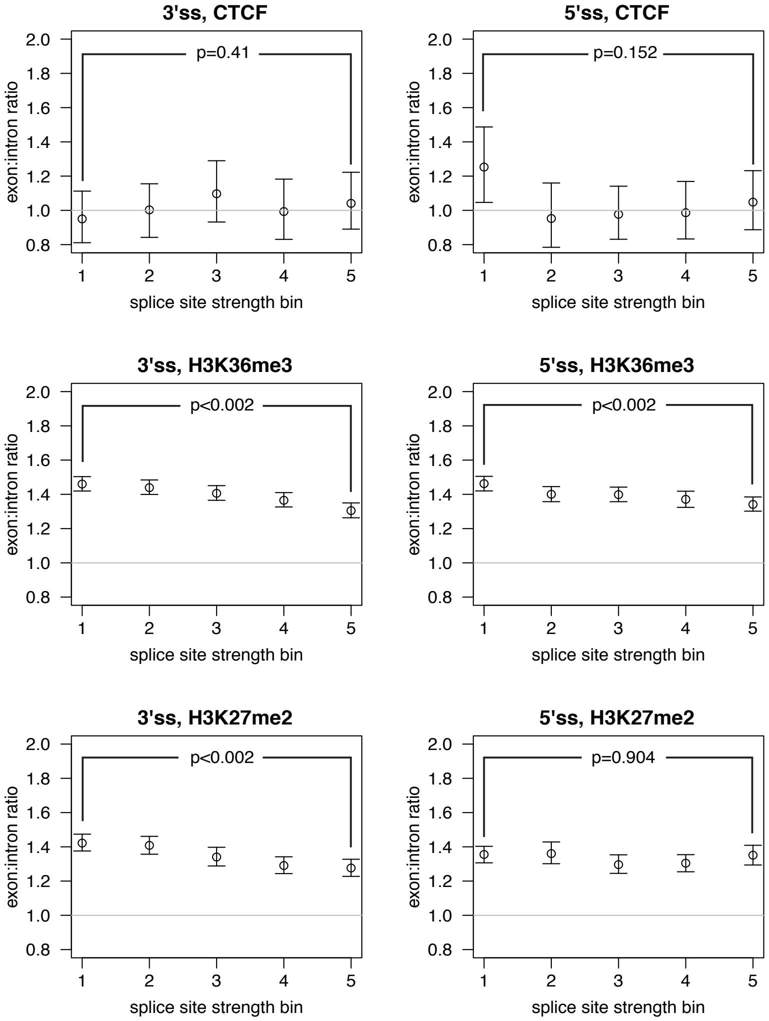


Figure S8. H3K36me3 enrichment dependence on splice site strength mirrors that of nucleosomes. Axes and values as in Fig. 3c and S7 but for datasets as indicated.

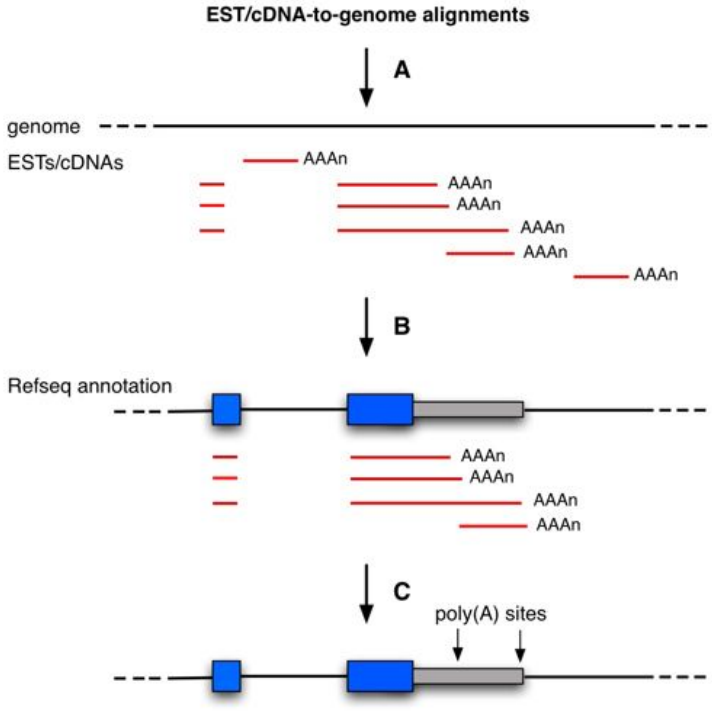


Figure S9. Outline of the approach used to build a database of poly(A) sites.
(A) First EST/cDNA-to-genome alignments from UCSC were filtered to keep only uniquely mapping ESTs/cDNAs with non-genomic poly(A) tails (minimum of 8 terminal A or T characters).
(B) Second, ESTs/cDNAs overlapping Refseq annotations were kept (blue boxes: exons; grey boxes: 3' UTRs). ESTs/cDNAs completely contained within introns or intergenic regions were removed.
(C) Third, genomic coordinates of poly(A) sites were mapped from alignments and poly(A) sites within 24 bp of each other were clustered. The –1 to –40 region upstream of each poly(A) site was searched for a poly(A) signal or variant. If a signal was found, the cluster was recorded as a poly(A) site (black arrow).

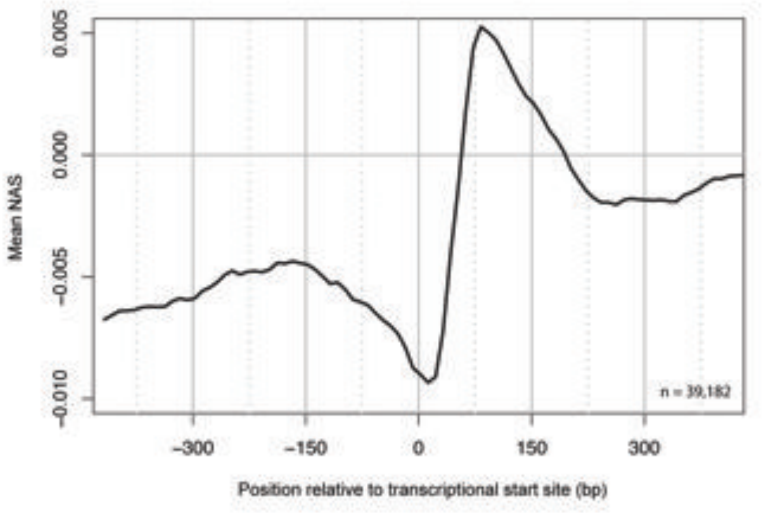
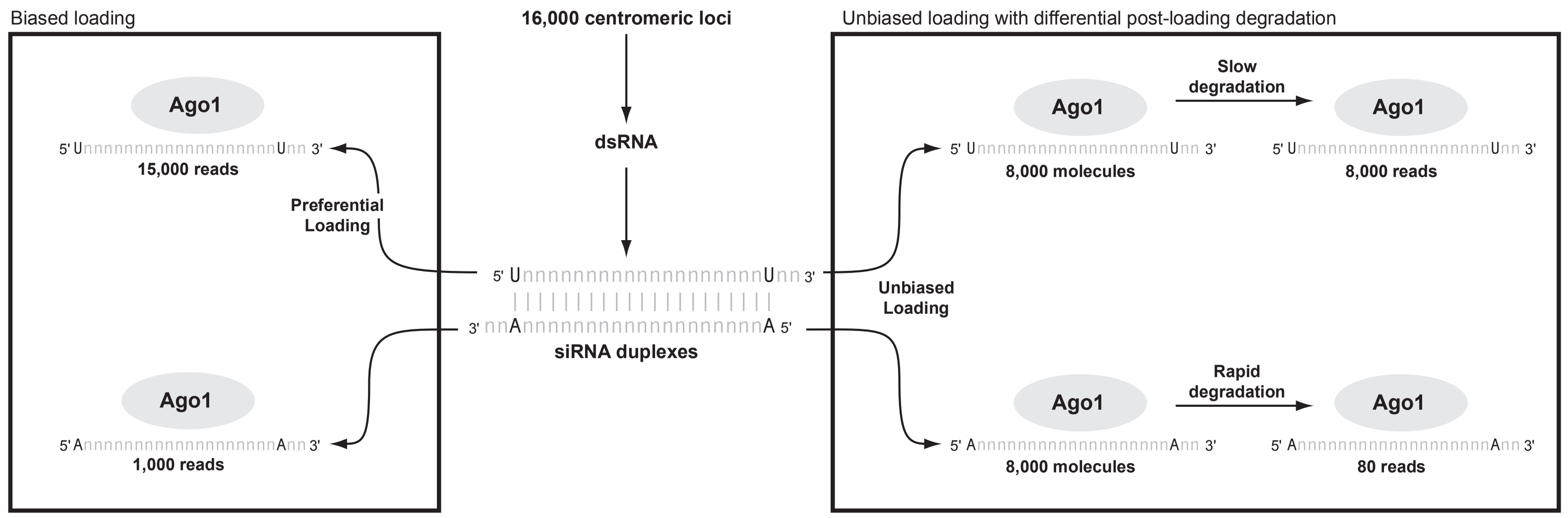


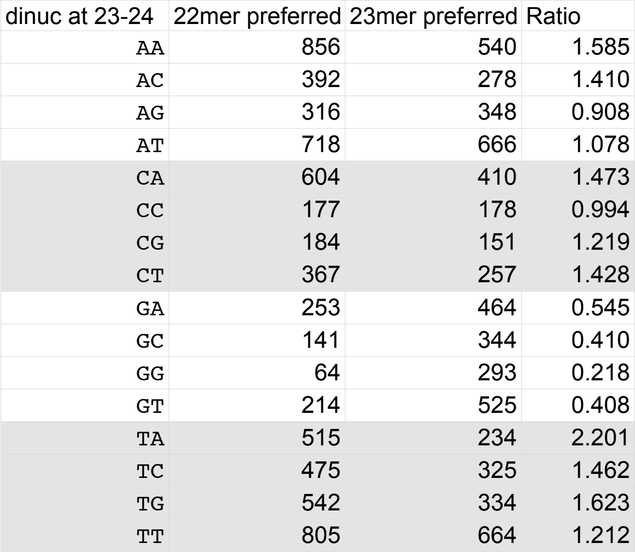
Figure S10. Mean nucleosome affinity scores (NAS) around transcriptional start sites, smoothed using a 50 nt sliding window positioned every 10 nt.

Appendix B

Supplementary information for Chapter 4: TRAMP-mediated RNA surveillance prevents spurious entry of RNAs into the Schizosaccharomyces pombe siRNA pathway

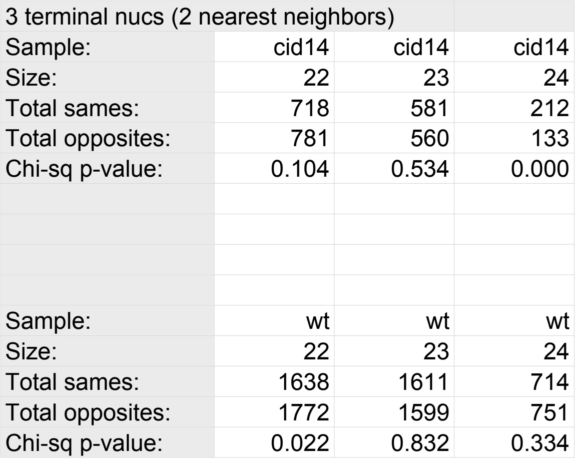

Biased loading
16,000 centromeric loci
Unbiased loading with differential post-loading degradation
Ago1
5' UnnnnnnnnnnnnnnnnnnnUnn 3'
15,000 reads
Preferential Loading
Ago1
5' AnnnnnnnnnnnnnnnnnnnAnn 3'
1,000 reads
dsRNA
5' UnnnnnnnnnnnnnnnnnnnUnn 3'
3' nnAnnnnnnnnnnnnnnnnnnnA 5'
siRNA duplexes
Unbiased Loading
Ago1
5' UnnnnnnnnnnnnnnnnnnnUnn 3'
8,000 molecules
Slow degradation
Ago1
5' UnnnnnnnnnnnnnnnnnnnUnn 3'
8,000 reads
Ago1
5' AnnnnnnnnnnnnnnnnnnnAnn 3'
8,000 molecules
Rapid degradation
Ago1
5' AnnnnnnnnnnnnnnnnnnnAnn 3'
80 reads

Supplementary Figure 1.



dinuc at 23-24	22mer preferred	23mer preferred	Ratio
AA	856	540	1.585
AC	392	278	1.410
AG	316	348	0.908
AT	718	666	1.078
CA	604	410	1.473
CC	177	178	0.994
CG	184	151	1.219
CT	367	257	1.428
GA	253	464	0.545
GC	141	344	0.410
GG	64	293	0.218
GT	214	525	0.408
TA	515	234	2.201
TC	475	325	1.462
TG	542	334	1.623
TT	805	664	1.212

Supplementary Table 1. There is an up to 2-fold preference for 22nt species over 23nt species when the base at position 23 is a U. For each locus where both the 22nt and 23nt sequences were present with identical 5' ends, the species with the greater number of normalized reads was indicated as preferred, and these counts were recorded according to the dinucleotide at position 22–23 from the common 5' end.



3 terminal nucs (2 nearest neighbors)			
Sample:	cid14	cid14	cid14
Size:	22	23	24
Total sames:	718	581	212
Total opposites:	781	560	133
Chi-sq p-value:	0.104	0.534	0.000
Sample:	wt	wt	wt
Size:	22	23	24
Total sames:	1638	1611	714
Total opposites:	1772	1599	751
Chi-sq p-value:	0.022	0.832	0.334

Supplementary Table 2. No evidence for strand preference based on terminal stability. Terminal three base pairs (2 nearest neighbors) were analyzed using a nearest neighbors algorithm and the 5' end with the highest stability was counted as same if it had more reads than the inferred duplex, or opposite if the opposite strand had more reads. Duplexes with identical stability at both ends were ignored. Because of the strong 5' U bias, only those duplexes with the same 5' nucleotide on both strands were counted.

Supplementary Table 3. Genes with highest number of antisense reads (sorted by number of reads antisense to gene in wild-type). The number of reads and sequences is counting only those reads that map uniquely to the genome, with the exception of those mapping to tlh1 and tlh2. Most gene-matching reads either map to several genomic loci (not shown here) or map to the sense strand of the gene and are presumably degradation products.

Wild-type				cid14					
Antisense		Sense		Antisense		Sense			
Reads	Seqs	Reads	Seqs	Reads	Seqs	Reads	Seqs	Systematic Name	Gene product
3,124*	2,172	1,196*	1,133	413*	597	122*	213	SPBCPT2R1.08c	RecQ type DNA helicase Tlh1
3,123*	2,171	1,228*	1,138	412*	596	121*	212	SPAC212.11	RecQ type DNA helicase
54	46	53	36	16	13	67	60	SPCC13B11.01	alcohol dehydrogenase Adh1
43	34	206	108	4	4	34	28	SPCC330.05c	orotidine 5'-phosphate decarboxylase Ura4
33	23	0	0	38	29	0	0	SPAC27E2.13	dubious
8	6	0	0	0	0	0	0	SPBC3D6.16	sequence orphan
7	7	0	0	4	4	5	5	SPBC317.01	MADS-box transcription factor Pvg4
7	5	0	0	5	2	0	0	SPBC725.06c	serine/threonine protein kinase Ppk31
6	4	0	0	2	1	0	0	SPAC212.06c	pseudogene/pseudo
6	6	0	0	2	2	0	0	SPAC22F3.04	AMP binding enzyme
4	4	26	22	0	0	25	24	SPAPB8E5.03	malic acid transport protein Mae1
4	4	0	0	7	5	0	0	SPCC285.14	TRAPP complex subunit Trs130
4	4	0	0	4	3	0	0	SPBC1861.06c	S. pombe specific UPF0300 family protein 4
4	4	0	0	1	1	0	0	SPAC25H1.09	alpha-amylase homolog Mde5
4	3	1	1	6	4	2	2	SPBC776.10c	Golgi transport complex peripheral subunit
4	3	2	1	1	1	1	1	SPBC1215.02c	NatB N-acetyltransferase complex non catalyticsubunit Arm1
4	1	0	0	0	0	1	1	SPAC25H1.05	sequence orphan
4	1	0	0	0	0	1	1	SPCC16C4.02c	DUF1941 family protein
3	3	0	0	5	5	0	0	SPBC19G7.01c	MutS protein homolog 2
3	3	1	1	3	3	4	3	SPAC22F8.11	phosphoinositide phospholipase C Plc1
3	3	1	1	1	1	6	5	SPBC947.02	AP-1 adaptor complex subunit Apl2
3	3	5	4	1	1	9	7	SPBC418.01c	imidazoleglycerol-phosphate synthase
3	3	3	3	0	0	2	2	SPAC10F6.02c	ATP-dependent RNA helicase Prp22
3	3	0	0	0	0	1	1	SPAC11H11.04	pheromone p-factor receptor
3	3	0	0	0	0	2	2	SPAC167.03c	U4/U6 x U5 tri-snRNP complex subunit Snu66
3	3	0	0	0	0	0	0	SPAC3H8.09c	poly(A) binding protein Nab3
3	3	0	0	0	0	0	0	SPAC6C3.07	sequence orphan
3	3	0	0	0	0	0	0	SPAC4F8.08	sequence orphan
3	3	0	0	0	0	0	0	SPBC3D6.10	AP-endonuclease Apn2
3	2	5	5	6	5	2	2	SPAC57A7.05	conserved protein (fungal and plant)
3	2	0	0	1	1	0	0	SPCC126.01c	conserved fungal protein
3	2	0	0	1	1	0	0	SPBC36B7.07	SNARE Tgl1
3	2	3	3	0	0	2	1	SPBC1539.07c	glutathione-dependent formaldehyde dehydrogenase
3	1	7	6	6	3	4	4	SPAC23C11.06c	hydrolase
3	1	0	0	1	1	0	0	SPBC32H8.06	TPR repeat protein, meiotically spliced
3	1	2	2	1	1	0	0	SPCC126.02c	Ku domain protein Pku70
3	1	1	1	0	0	2	2	SPBC21.05c	Ras guanyl-nucleotide exchange factor Ral2
3	1	2	2	0	0	0	0	SPBC776.06c	spindle pole body interacting protein
3	1	0	0	0	0	2	1	SPAC5H10.09c	3-methyl-2-oxobutanoatehydroxymethyltransferase
3	1	0	0	0	0	1	1	SPAC3C7.03c	RecA family ATPase Rhp55
3	1	1	1	0	0	0	0	SPBP8B7.28c	sequence orphan

* Nearly all of these reads map to the genome exactly twice: once to tlh1 and once to tlh2.

Table S4. List of strains used in this study.

Strain	Genotype
SPY137	SPY137 *h*$^{+}$ *leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4*$^{+}$ *oriA*
SPY1220	*h*$^{+}$ *leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4*$^{+}$ *oriA cid14Δ::nat*^{R}
SPY815	*h*$^{+}$ *leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4*$^{+}$ *oriA clr4Δ::kan*^{R}
SPY1220	*h*$^{+}$ *leu1-32 ade6-M210 ura4DS/E otr1*R*(SphI)::ura4*$^{+}$ *oriA cid14Δ::nat*^{R}
SPY28	*h*$^{+}$ *leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4*$^{+}$ *oriI*
SPY86	*h*$^{+}$ *leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4*$^{+}$ *oriI dcr1Δ::TAP-kan*^{R}
SPY87	*h*$^{+}$ *leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4*$^{+}$ *oriI rdp1Δ::TAP-kan*^{R}
SPY399	*h*$^{+}$ *leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4*$^{+}$ *oriI clr4Δ::nat*^{R}
SPY787	*h*$^{+}$ *leu1-32 ade6-M216 ura4-D18 imr1*R*(NcoI)::ura4*$^{+}$ *oriI cid14Δ::nat*^{R}
SPY139	*h90 leu1-32 ade6-M210 ura4DS/E mat3M::ura4*$^{+}$
SPY1313	*h90 leu1-32 ade6-M210 ura4DS/E mat3M::ura4*$^{+}$ *rrp6Δ:: nat*^{R}
SPY1408	*h90 leu1-32 ade6-M210 ura4DS/E mat3M::ura4*$^{+}$ *mtr4-1(mtr4*$^{+}$*::TAP-nat*^{R}*)*
SPY1284	*h*$^{-}$ *leu1Δ ura4Δ dis3-54*
SPB45	*h? cid14Δ:: nat*^{R} *rdp1Δ:: kan*^{R} *ura4+::5BoxB/HPH leu1-32*
SPB46	*h? cid14Δ:: nat*^{R} *dcr1Δ:: kan*^{R} *ura4+::5BoxB/HPH leu1-32*

Table S5. List of oligonucleotides used in this study.

Name	Sequence
mb86	5’-AACCCTCAGCTTTGGGTCTT-3’
mb87	5’-TTTGCATACGATCGGCAATA-3’
mb90	5’-CAACCCTCAGCTTTGGGTCTTG-3’
mb91	5’-TCCTTTTGCATACGATCGGCAATAC-3’
mb510	5’-AAAATGTTTCTATGCTACTTTAACAATTCGCACAAAG-3’
mb511	5’-AAAGTGCACGCTCTAATTTTAATTTTAACAGTCTATAAAGTTTAG-3’
mb512	5’-CCTAGCCACTGGACCATGACGGA-3’
mb521	5’-AGGCCAGCTACGCTACTC-3’
mb522	5’-CGACTTACTATTAAGCATTGATTGCAAATTACATTTTG-3’
mb523	5’-AAATAGTGTCTGAACAATAATCATAAAACTTTCTATGCTAAC-3’
mb524	5’-CATAGTATCTTAGAAAAATGTGAAAAGTGTTAGTTTACTATTCTC-3’
mb525	5’-TTAAGCATAATAAAAAGATTCTTTGAAAGTGGAAGAAATCATG-3’
mb526	5’-CACTAAAAATTTGAGAAAATAATAAAACGTGTCAAGCTCTTTC-3’
mb527	5’-TTAAACGTAACCGATACATAATTTAGGCAAAAATTGTTG-3’
mb528	5’-GTTCATCTAAAAGCTTCAAAAAATATTAATATTGAGTCTAAAATCAAGT-3’
rsi1	5’-CAAGTTTGTCCAACTTCTCGGCA-3’
rsi2	5’-AGCCAATCCAGAGGCCTCACTAA-3’
rsi3	5’-TAATGATCCTTCCGCAGGTTCACC-3’
rsi4	5’-AGGTAGTGGTATTTCACCGGCGTA-3’
rsi5	5’-AAGCCAATCCAGAGGCCTCACTAA-3’
rsi6	5’-GGCGAGAAAAGACATCGGTCCAC-3’
rsi7	5’-ATTTTTTGCCTACCAACAAGA-3’
rsi8	5’-GACCAGTAAACACGCCTTGCG-3’
rsi9	5’-CCAAGTTTGTCCAACTTCTCGGCA-3’
rsi10	5’-ACCAGTAAACACGCCTTGCG-3’
rsi11	5’-GGTATTGTAAGCAGTAGAGTA-3’
rsi12	5’-CAATGGTAATTCAACTTAGTA-3’
rsi13	5’-CAGAATTCGGTAAGCGTTGGA-3’
rsi14	5’-GCAATGGTAATTCAACTTAGTA-3’
rsi15	5’-TTGGACAAACTTGGTCATTTA-3’
rsi16	5’-GTATTGTAAGCAGTAGAGTA-3’
rsi17	5’-ACTTGTTCCTACTCTCCTGTA-3’
rsi18	5’-TTCCTACTCTCCTGTATCGTA-3’
DM566	5’-TTATTGATGGCGAAGCTAGATCCG-3’
DM567	5’-AACTCCATAACCACCACCATGCTC-3’

Supplementary Text

Strand selection

Duplexes of miRNAs and synthetic siRNAs tend to bind the loading machinery asymmetrically, such that the strand least stably paired at its 5' end is preferentially loaded as the guide strand within the silencing complex[34]. We examined whether the same was true for heterochromatic siRNAs, focusing on the inferred duplexes that match sequenced 23mers deriving from the centromeric *dg*/*dh* repeats. To ensure that the identity of the 5' nucleotides did not influence the result, we considered only the 3210 duplexes in which the two 5' nucleotides were identical (**Supplementary Table 2**).

Ago1 preferentially loads siRNAs with 5'-uracil

Since preferential siRNA processing and pairing asymmetry contribute little to the enrichment for species with a 5'-U, this strong bias must arise either from preferential loading of siRNAs with a 5'-U into Ago1, or preferential stability of species with 5'-U already in Ago1. To discern between these two possibilities, we looked at centromeric reads with 5'-U whose inferred duplex partner also has a 5'-U (U...A.. species, **Table 1**). There are approximately 8,000 centromeric loci that could form such duplexes. We compared reads from these duplexes to 5'-U reads whose inferred duplex partner has a 5' A (U...U.., **Table 1**). Because U...U.. can occur independently on either genomic strand, there are approximately twice as many such loci, and indeed, we counted approximately 16,000 centromeric occurrences of the 23mer sequence U...U.. .

For each U...A.. duplex, either the + or – strand can be loaded into Ago1, presumably with identical affinity, given the lack of pairing asymmetry and that the opposite strand also has a 5'-U. The number of reads from these duplexes would be proportional to the depth of sequencing and the number of genomic loci, assuming each locus produces dsRNA at approximately the same rate. We observed an average of slightly less than one read per centromeric locus (7,654 reads).

Suppose Ago1 loaded siRNAs with equal efficiency, regardless of 5' nucleotide, but that loaded siRNAs beginning with G, C and A were degraded much more rapidly than those with a 5' U (**Supplementary Figure 1**, right box). In this case, we would expect half of all U...U.. duplexes to load the U...U.. strand, and half to load the paired A...A.. strand instead. Once loaded, the A...A.. species would degrade quickly, resulting in the substantial 5' nucleotide bias in our sequencing data. Given the approximately 1 read to 1 genomic locus ratio from above, and 16,000 loci, we would expect approximately 8,000 U...U.. reads. However, we observed 9,245 reads, significantly more than expected by this model.

We turn instead to a model whereby Ago1 loads the strand with 5' U preferentially compared to duplex partners with 5' G, C or A (**Supplementary Figure 1**, left box). Under this model, we expect some majority of reads from U...U../A...A.. duplexes to be loaded from the U...U.. strand. At a rate of 1 read to 1 genomic locus, this would mean observing more than 8,000 reads, which is consistent with the 9,245 reads observed. The number of U...U.. reads plus the number of A...A.. reads is not twice the number of U...A.. reads, despite the fact that there are twice as many centromeric loci. We attribute this observation to a limited number of encounters of the U...U../A...A.. duplex with the Ago1-loading machinery because some duplex molecules that are released after nonproductive encounters in the suboptimal orientation are presumably degraded before they have another opportunity to encounter the loading machinery.

Because the centromeric regions are particularly AT-rich, the number of potential duplexes with 5'-U and an inferred duplex species with a 5'-G or 5'-C is much lower than that with an inferred duplex species with a 5'-A. There are approximately 8,000 U...C.. loci, and a similar number of U...G.. loci, in the centromeric regions. The model in which siRNAs are loaded with equal efficiency and then differential post-loading degradation explains the sequencing bias predicted approximately 4,000 U...C.. reads and approximately 4,000 U...G..

reads. However, the observed number of reads from U...C.. species (7,423 reads) and U...G.. species (7,835 reads) were nearly 1 per locus, consistent with the second model, and further supporting preferential loading as a major factor in the 5'-U bias.

Appendix C

Mammalian microRNAs: experimental evaluation of novel and previously annotated genes

H. Rosaria Chiang, Lori W. Schoenfeld, J. Graham Ruby, Vincent C. Auyeung, Noah Spies, Daehyun Baek, Wendy K. Johnston, Carsten Russ, Shujun Luo, Joshua E. Babiarz, Robert Blelloch, Gary P. Schroth, Chad Nusbaum, and David P. Bartel

Author contributions I contributed the analysis of untemplated U addition to the following work previously published as:

Chiang, HR, LW Schoenfeld, JG Ruby, VC Auyeung, N Spies, D Baek, WK Johnston, C Russ, S Luo, JE Babiarz, R Blelloch, GP Schroth, C Nusbaum, and DP Bartel (May 2010). “Mammalian microRNAs: experimental evaluation of novel and previously annotated genes.” In: *Genes Dev* 24.10, pp. 992–1009. DOI: `10.1101/gad.1884710`.

Mammalian microRNAs: experimental evaluation of novel and previously annotated genes

H. Rosaria Chiang,[1,2] Lori W. Schoenfeld,[1,2] J. Graham Ruby,[1,2,7] Vincent C. Auyeung,[1,2,3] Noah Spies,[1,2] Daehyun Baek,[1,2] Wendy K. Johnston,[1,2] Carsten Russ,[4] Shujun Luo,[5] Joshua E. Babiarz,[6] Robert Blelloch,[6] Gary P. Schroth,[5] Chad Nusbaum,[4] and David P. Bartel[1,2,8]

[1]Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA; [2]Howard Hughes Medical Institute and Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; [3]Harvard-Massachusetts Institute of Technology Division of Health Sciences and Technology, Cambridge, Massachustts 02139, USA; [4]Broad Institute of Massachusetts Institute of Technology and Harvard, Cambridge, Massachusetts 02141, USA; [5]Illumina, Inc., Hayward, California 94545, USA; [6]Institute for Regeneration Medicine, Center for Reproductive Sciences, and Department of Urology, University of California at San Francisco, San Francisco, California 94143, USA

MicroRNAs (miRNAs) are small regulatory RNAs that derive from distinctive hairpin transcripts. To learn more about the miRNAs of mammals, we sequenced 60 million small RNAs from mouse brain, ovary, testes, embryonic stem cells, three embryonic stages, and whole newborns. Analysis of these sequences confirmed 398 annotated miRNA genes and identified 108 novel miRNA genes. More than 150 previously annotated miRNAs and hundreds of candidates failed to yield sequenced RNAs with miRNA-like features. Ectopically expressing these previously proposed miRNA hairpins also did not yield small RNAs, whereas ectopically expressing the confirmed and newly identified hairpins usually did yield small RNAs with the classical miRNA features, including dependence on the Drosha endonuclease for processing. These experiments, which suggest that previous estimates of conserved mammalian miRNAs were inflated, provide a substantially revised list of confidently identified murine miRNAs from which to infer the general features of mammalian miRNAs. Our analyses also revealed new aspects of miRNA biogenesis and modification, including tissue-specific strand preferences, sequential Dicer cleavage of a metazoan precursor miRNA (pre-miRNA), consequential 5′ heterogeneity, newly identified instances of miRNA editing, and evidence for widespread pre-miRNA uridylation reminiscent of miRNA regulation by Lin28.

[*Keywords*: MicroRNA; miRNA biogenesis; noncoding RNA genes; high-throughput sequencing]

Supplemental material is available at http://www.genesdev.org.

Received November 11, 2009; revised version accepted March 19, 2010.

MicroRNAs (miRNAs) are endogenous ~22-nucleotide (nt) RNAs that post-transcriptionally regulate gene expression (Bartel 2004). miRNAs mature through three intermediates: a primary miRNA transcript (pri-miRNA), a precursor miRNA (pre-miRNA), and a miRNA:miRNA* duplex. RNA Polymerase II transcribes the pri-miRNA, which contains one or more segments that each fold into an imperfect hairpin. For canonical metazoan miRNAs, the RNase III enzyme Drosha together with its partner, the RNA-binding protein DGCR8, recognize the hairpin, and Drosha cleaves both strands ~11 base pairs (bp) from the base of the stem (Han et al. 2006). The cut leaves a 5′ phosphate and 2-nt 3′ overhang (Lee et al. 2003). The liberated pre-miRNA hairpin is then exported to the cytoplasm by Exportin-5 (Yi et al. 2003; Lund et al. 2004). There, the RNase III enzyme Dicer cleaves off the loop of the pre-miRNA, ~22 nt from the Drosha cut (Lee et al. 2003), again leaving a 5′ monophosphate and 2-nt 3′ overhang. The resulting miRNA:miRNA* duplex, comprised of ~22-nt strands from each arm of the original hairpin, then associates with an Argonaute protein such that the miRNA strand is usually the one that becomes stably incorporated, while the miRNA* strand dissociates and is degraded.

In addition to canonical miRNAs, some miRNAs mature through pathways that bypass Drosha/DGCR8 recognition and cleavage. Members of the mirtron subclass of pre-miRNAs are excised as intron lariats from the pri-miRNA by the spliceosome and, following debranching, fold into Dicer substrates (Okamura et al. 2007; Ruby et al.

[7]**Present address: Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA 94158, USA.**
[8]**Corresponding author.**
E-MAIL dbartel@wi.mit.edu; FAX (617) 258-6768.

Article published online ahead of print. Article and publication date are online at http://www.genesdev.org/cgi/doi/10.1101/gad.1884710.

2007a). For some mirtrons, known as tailed mirtrons, a longer intron is excised such that only one end of the pre-miRNA is generated by the spliceosome, whereas the other end of the pre-miRNA matures through the Drosha-independent trimming of a 5′ or 3′ tail (Ruby et al. 2007a; Babiarz et al. 2008). Members of another subclass of pre-miRNAs, called endogenous shRNAs, are suitable Dicer substrates without preprocessing by either Drosha or the spliceosome (Babiarz et al. 2008). Other small silencing RNAs are generated from the sequential processing of long hairpins or long bimolecular duplexes. These small RNAs are classified as endogenous siRNAs rather than miRNAs because they derive from extended duplexes that produce many different small RNA species, whereas miRNAs derive from distinctive hairpins that produce one or two dominant species (Bartel 2004).

The first indication of the abundance of miRNA genes came from sequencing small RNAs from mammals, flies, and worms (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001). Hundreds of mammalian miRNAs have been identified by Sanger sequencing of cloned small RNA-derived cDNAs (Lagos-Quintana et al. 2001, 2002, 2003; Houbaviy et al. 2003; Berezikov et al. 2006b; Landgraf et al. 2007). Some miRNAs, however, are expressed only in a limited number of cells or through a limited portion of development, and their rarity makes them difficult to detect. Computational methods have been used to identify mammalian miRNAs initially missed by sequencing, and some of these predicted miRNAs have been evaluated experimentally—e.g., by rapid amplification of cDNA ends (RACE) (Lim et al. 2003; Xie et al. 2005), hybridization to RNA blots (Berezikov et al. 2005), microarrays (Bentwich et al. 2005), and RNA-primed array-based Klenow extension (RAKE) (Berezikov et al. 2006b). Each of these experimental methods, however, can yield false positives. Indeed, recent work in invertebrates and plants (Rajagopalan et al. 2006; Ruby et al. 2006, 2007b) has shown that the fraction of erroneously annotated miRNAs can be quite high, depending on the quality of the initial computational predictions. Even when miRNA genes are predicted correctly, the resolution of the prediction is often insufficient to confidently determine the precise 5′ end of the mature miRNA. Because miRNAs repress target mRNAs by pairing to the seed sequence, which is defined relative to the position of the miRNA 5′ end, single-nucleotide resolution of 5′-end annotations is required for useful downstream analysis of their physiological consequences (Bartel 2009).

Another approach for finding miRNAs and other small RNAs missed in the early discovery efforts is high-throughput sequencing (Lu et al. 2005). In mammals, high-throughput sequencing methods that have contributed to miRNA discovery efforts have included massively parallel signature sequencing (MPSS) (Mineno et al. 2006), miRNA serial analysis of gene expression (miRAGE) (Cummins et al. 2006), 454 pyrosequencing (Berezikov et al. 2006a, 2007; Calabrese et al. 2007), and Illumina sequencing (Babiarz et al. 2008; Kuchenbauer et al. 2008).

Here we use the Illumina sequencing-by-synthesis platform (Seo et al. 2004) for miRNA discovery in mice. Analyses of these reads, combined with experimental evaluation of newly identified miRNAs as well as previous annotations, led us to substantially revise the set of confidently identified murine miRNAs, thereby providing a more accurate picture of the general features of mammalian miRNAs and their abundance in the genome. In addition, our results revealed new aspects of miRNA biogenesis and modification, including tissue-specific strand preferences, sequential Dicer cleavage of a metazoan pre-miRNA, cases of consequential 5′ heterogeneity, newly identified instances of miRNA editing, and widespread pre-miRNA uridylation reminiscent of Lin28-like miRNA regulation.

Results

We sequenced small-RNA libraries from three mouse tissues—brain, ovary, and testes—as well as embryonic day 7.5 (E7.5), E9.5, E12.5, and newborn. Combining these data with data collected similarly from mouse embryonic stem (ES) cells (Babiarz et al. 2008) yielded 28.7 million reads between 16 nt and 27 nt in length that perfectly matched the mouse genome assembly (Supplemental Table 1). Of these reads, 79.3% mapped to miRNA hairpins, and 7.1% mapped to other annotated noncoding RNA genes (Supplemental Table 2). Because the sequencing protocol was selective for RNAs with 5′ monophosphate and 3′ hydroxyl groups, this dominance of miRNA species was expected (Lau et al. 2001).

miRNA gene discovery

As when analyzing high-throughput data from invertebrates (Ruby et al. 2006, 2007b; Grimson et al. 2008), we identified miRNA genes in mice by applying the following criteria: (1) expression of the candidate miRNA, with a relatively uniform 5′ terminus; (2) pairing characteristics of the predicted hairpin; (3) absence of annotation suggesting non-miRNA biogenesis; (4) absence of proximal reads suggesting that the candidate is a degradation intermediate; and (5) presence of reads corresponding to a miRNA* species with potential to pair to the miRNA candidate with ~2-nt 3′ overhangs. Using a low-stringency genomic search strategy that considered the first four criteria, 736 miRNA candidates were identified from the total data set of mouse reads. Manual inspection of these candidates, focusing on all five criteria, narrowed the list to 465 canonical miRNA genes, 377 of which were already annotated in miRBase version 14.0 (Griffiths-Jones 2004) and 88 of which were novel (Fig. 1A; Supplemental Fig. S1; Supplemental Table 3). We also found 14 mirtrons (including 10 tailed mirtrons), four of which were already annotated, and 16 endogenous shRNAs, six of which were annotated previously (Fig. 1B). When added to the 88 novel canonical miRNA genes, the newly identified mirtons and shRNAs raised the total number of novel genes to 108.

Of these 108 genes, 36 appeared to be close paralogs of previously annotated miRNA genes (most of which were paralogs of *mir-466*, *mir-467*, or *mir-669*), producing

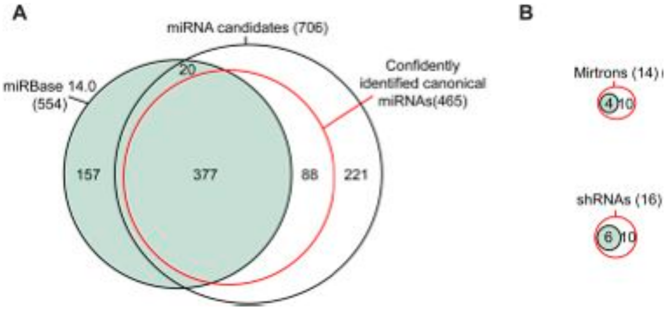


Figure 1. Mouse miRNAs and candidates initially identified by high-throughput sequencing. (*A*) Overlap between previously annotated miRNA hairpins (miRBase version 14.0; green), miRNA candidates identified in the current study, and the subset of these candidates that met our criteria for classification as confidently identified canonical miRNAs (red). Additional considerations increased the number of confidently identified canonical miRNAs to 475. (*B*) Overlap between previously annotated mirtrons and shRNAs and the mirtrons and shRNAs supported by our study, colored as in *A*.

miRNA reads that were identical to the previously annotated miRNAs, creating ambiguity as to which loci contributed to the sequenced reads. Most of these close paralogs (35 of 36), as well as 14 other novel loci, were clustered with annotated miRNAs. The 72 novel genes with reads distinguishable from those of previously identified genes were expressed at a lower level than the previously annotated genes (median read counts 27 and 8206, respectively), and, compared with previously annotated miRNAs, a higher fraction of these novel miRNAs were located within introns of annotated RefSeq (Pruitt et al. 2005) mRNAs (47% and 26%, respectively).

Experimental evaluation of unconfirmed miRNAs

Of 564 miRBase-annotated miRNA genes (including four confirmed mirtons and six confirmed shRNAs) that map to mm8 genome assembly, 157 annotated miRNAs did not pass the filters for miRNA candidates (Fig. 1A,B; Supplemental Fig. S1; Supplemental Table 4). Of these 157, 26 mapped to annotated rRNA and tRNA loci, 52 had no reads mapping to them, and another 72 had some reads but in numbers deemed insufficient for confident annotation. The remaining seven either had reads with very heterogeneous 5′ ends, which suggested nonspecific degradation of a non-pri-miRNA transcript (*mir-464*, *mir-1937a*, and *mir-1937b*); had many reads that mapped well into the loop of the putative hairpin, which were inconsistent with Dicer processing (*mir-451*, *mir-469*, and *mir-805*); or did not give a predicted fold with the requisite pairing involving the candidate and predicted miRNA* (*mir-484*) (Supplemental Fig. S2). For five of these seven, we have no reason to suspect that they might be authentic miRNA genes. Among the remaining two, *mir-484* might be regarded as a miRNA candidate because manual refolding was able to generate a hairpin with the requisite pairing, but, even so, this candidate lacked reads for the predicted miRNA*. miR-451 is a noncanonical miRNA generated from an unusual hairpin without production of a miRNA:miRNA* duplex (S Cheloufi and G Hannon, pers comm.). We do not suspect that any other annotated miRNA genes failed to pass our filters for the same reason as *mir-451*.

An additional 20 annotated miRNA hairpins were in our set of candidates but failed the manual inspection because they lacked predicted miRNA* reads even after allowing for alternate hairpin structures. Hundreds of candidates from other miRNA discovery efforts (Xie et al. 2005; Berezikov et al. 2006b) also failed to pass the filters, usually because no reads mapped to them.

One of the annotated miRNA genes missing from our data sets was *mir-220*, which had been predicted computationally using MiRscan as a miRNA gene candidate conserved in humans, mice, and fish, and was supported experimentally using RACE analysis of zebrafish small RNAs (Lim et al. 2003). In contrast, the other 37 miRNAs newly annotated by Lim et al. (2003) were among our confirmed miRNAs. The absence of *mir-220* in our data sets might have reflected either very low expression in the sequenced samples or inaccuracy of its annotation. Similarly, *mir-207*, annotated in a contemporaneous study that cloned novel miRNAs from mouse tissues, was missing from our data set, but another 27 miRNAs annotated from that study were confirmed (Lagos-Quintana et al. 2003).

To evaluate whether the missing annotated miRNAs and candidates represented authentic miRNAs, we developed a moderate-throughput assay to examine if their respective hairpins could be processed as miRNAs in cultured cells (Fig. 2A). If these putative miRNAs were missing from our data sets because they were not expressed in the sequenced tissues or stages, we reasoned that they would probably be detected in cells ectopically expressing their respective hairpins, because most authentic miRNAs are processed correctly from heterologous transcripts that include the full hairpin flanked by ∼100 nt of genomic sequence on each side of the hairpin (Chen et al. 2004; Voorhoeve et al. 2006). Alternatively, if these putative miRNAs were missing because they were not authentic miRNAs and therefore lacked the features needed for Drosha and Dicer processing, they would not be sequenced from cells ectopically expressing their hairpins. To evaluate many hairpins simultaneously, we transfected pools of hairpin-expressing constructs into HEK293T cells and isolated small RNAs for high-throughput sequencing.

The performance of 26 positive controls, chosen from canonical human/mouse miRNAs confirmed by our sequencing from mice, illustrated the value of the assay. For all but one of these controls, miRNA and miRNA* reads were more abundant in the cells ectopically expressing the hairpin than in the cells without the hairpin constructs (Fig. 2B–D; Supplemental Figs. S3, S4). For example, both hsa-miR-193b and mmu-miR-137 (from humans and mice, respectively) were >10 fold overexpressed (Fig. 2B). The positive controls included genes of tissue-specific miRNAs,

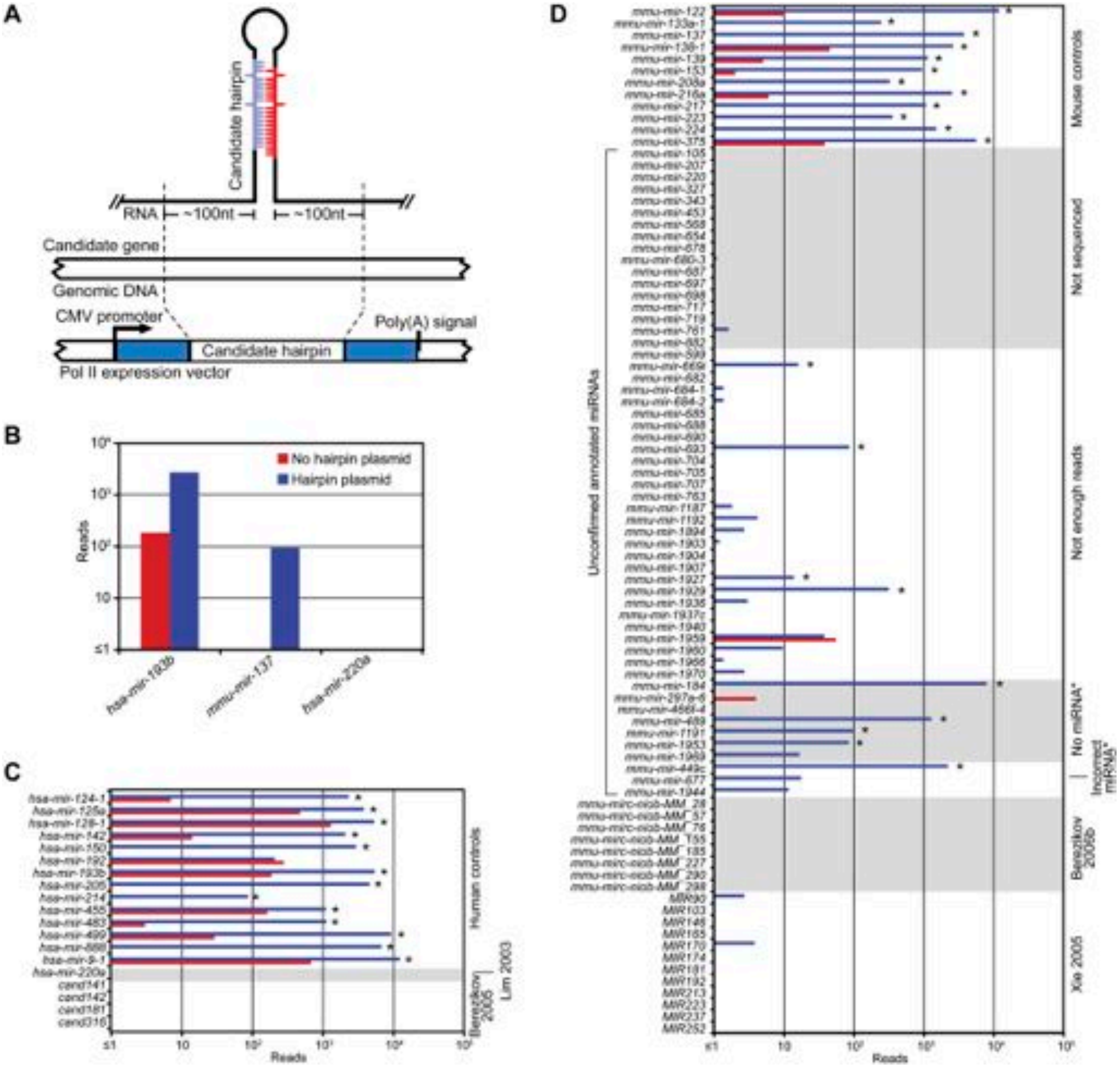


Figure 2. Experimental evaluation of annotated miRNAs and previously proposed candidates. (*A*) Schematic of the expression vector transfected into HEK293T cells. (*B*) Examples of the standard ectopic expression assay, transfecting plasmids indicated in the key. Reads from the control transfection (no hairpin plasmid) were from endogenous expression in HEK293T cells. (*C*) Assay results for annotated human miRNAs and published candidates. Bars are colored as in *B*; asterisks indicate detectable overexpression (≥1 read from both the anticipated miRNA and miRNA*, with miRNA and miRNA* combined expressed more than threefold over endogenous levels). (*D*) Assay results for unconfirmed annotated mouse miRNAs and published candidates. Mouse controls were selected from miRNAs that were sequenced from our mouse samples. Bars are colored as in *B*; detectable overexpression is indicated (asterisks). Shown are the results compiled from two experiments (Supplemental Figs. S3, S4).

including *mir-122* (liver), *mir-133* (muscle), *mir-223* (neutrophil), and several neuron-specific miRNAs, with the idea that hairpins of tissue-specific miRNAs might require tissue-specific factors for their processing, and therefore might be sensitive to the potential absence of such factors in HEK293T cells. Differences were observed, ranging from ~100 to 10,000 reads above the control transfection (Fig. 2C, *hsa-mir-214* and *hsa-mir-9-1*, respectively), consistent with the idea that factors absent in HEK293T cells might play a role in processing of some miRNAs. Alternatively, some miRNA hairpins might be processed less efficiently in all cell types, perhaps because our vectors might not present the hairpins in an optimal context for processing. Perhaps *hsa-mir-192*, the control gene that did not overexpress in our assay, lacked crucial processing determinants needed in all cells. In either scenario, the very high sensitivity of high-throughput sequencing enabled miRNAs to be observed from most of the less efficiently processed hairpins.

From the 52 annotated mouse miRNAs that our study did not sequence, 17 miRNAs, including *mir-220* and *mir-207*, were tested in the ectopic expression assay. One, *mir-698*, generated a single read corresponding to the annotated miRNA, and the rest failed to generate any reads representing the annotated miRNA (Fig. 2D). From the 72 annotated miRNAs that we could not identify due to insufficient number of reads, 28 were tested, and only four of these were found to be overexpressed (Fig. 2D). The difficulty in overexpressing a canonical control miRNA (hsa-miR-192) illustrates that our ectopic expression assay cannot be used to prove conclusively that a particular hairpin does not represent an authentic miRNA gene. However, the inability to overexpress each of the 17 unsequenced miRNAs, as well as most of the 28 insufficiently sequenced miRNAs, strongly indicated that, overall, these annotations have been faulty, and that our failure to detect previously annotated miRNAs in mouse samples was not merely due to inadequate sequencing coverage.

We also tested 10 of the 20 annotated miRNA genes that we identified as candidates but did not confidently classify as miRNA genes because the predicted miRNA* species was not sequenced. Four of seven genes without a miRNA* read and one of three genes with substantially offset miRNA* reads produced the predicted miRNA* species in our ectopic expression assay (Fig. 2D). *mir-184* and *mir-489*, both of which tested positive in this assay, are conserved. *mir-184* is conserved throughout mammals, and *mir-489* is conserved to chicken, although the miRNA seed, which is highly conserved in mammals and chickens, differs in mice and rats. Thus, these two genes, as well as *mir-875*, which is a broadly conserved gene without a miRNA* read, were added to our set of confidently identified miRNA genes. Also added were *mir-290*, *mir-291a*, *mir-291b*, *mir-292*, *mir-293*, *mir-294*, and *mir-295*, which were missing in the genome assembly (mm8) used in our analysis because they fall in the region of the genome that is difficult to assemble. Including these 10 genes, plus *mir-451*, brings the total number of confidently identified miRNA genes to 506, which includes 475 canonical genes.

Our sets of confirmed and novel murine miRNAs also provided the opportunity to evaluate results of more recent computational efforts to find miRNAs conserved among mammals. One set of studies predicted miRNAs based on phylogenetic conservation, and then tested these and additional murine-specific hairpins using RAKE and cloning (Berezikov et al. 2005, 2006b). Among the 322 candidates supported by these experiments, 11 were in our sets of miRNAs (two in our confirmed set, and nine in our novel set), and another nine did not satisfy our annotation criteria but had at least one read consistent with the predictions. Another study started with MiRscan predictions conserved in four mammals, and filtered these predictions for potential seed pairing to conserved motifs in 3′ untranslated regions (UTRs) (Xie et al. 2005). Of their 144 final candidates, 45 were paralogs of miRNAs already published at the time of prediction. Of the remaining 99 candidates, 27 were in our sets of miRNAs (26 in our confirmed set and one in our novel set), and one did not satisfy our annotation criteria but had three reads consistent with the miRNA* of the predicted miRNA. However, only four of the 27 confirmed miRNA genes (4% of the 99 novel predictions) gave rise to the mature miRNA with the predicted seed, suggesting that filtering MiRscan predictions for potential seed pairing provided little, if any, added benefit. This conclusion concurs with a recent analysis of miRNA targeting: miRNAs that are not conserved beyond mammals do not have enough preferentially conserved sites to place these sites as among the most conserved UTR motifs (Friedman et al. 2009). Therefore, it stands to reason that preferentially conserved UTR motifs would provide little value for predicting such miRNAs.

To investigate whether the computational candidates might have been missed because of low expression in tissues and stages from which we sequenced, we included representatives from each study in our ectopic expression assay. We randomly selected 12 Xie et al. (2005) candidates and eight Berezikov et al. (2006b) candidates that our study did not sequence, as well as four human candidates from the Berezikov et al. (2005) set whose mouse orthologs were not sequenced. None generated reads representing the candidate miRNAs (Fig. 2C,D). Taken together, our results raise new questions regarding the authenticity of these candidates, and suggest that previous extrapolation from these candidates, which had suggested that mammals have a surprisingly high number of conserved miRNA genes (as many as 1000) (Berezikov et al. 2005), should be revised accordingly.

Experimental evaluation of novel miRNAs and new candidates

We also used the ectopic expression assay to evaluate novel miRNAs identified from our sequencing. Of the 25 evaluated hairpins, 18 (72%) generated a significant number of miRNA-like reads in HEK293T cells, indicating that most, although perhaps not all, of our 108 novel annotations represented authentic miRNAs (Fig. 3; Supplemental Figs. S5, S6). These 25 hairpins were selected arbitrarily for evaluation, except for a preference for rare miRNAs; i.e., those that had <10 mature miRNA reads. The rare miRNAs and the higher-abundance miRNAs performed similarly (five of seven and 11 of 14 positives, respectively).

To evaluate Drosha and Dicer dependence of the overexpressed hairpins, the experiment was repeated with and without a plasmid encoding a dominant-negative allele of either Drosha or Dicer (Fig. 3A; Han et al. 2009). All but two canonical miRNA controls and most of the novel canonical miRNAs (16 of 17) responded to TNdrosha coexpression (Fig. 3B; Supplemental Fig. S7). Fewer responded to TNdicer, suggesting that this construct was less disruptive of normal miRNA processing (Supplemental Fig. S7).

The tested hairpins included several noncanonical miRNA precursors. The level of mmu-miR-1224, an annotated mirtronic miRNA (Berezikov et al. 2007), increased in the presence of TNdrosha, as expected if this pre-miRNA had more access to Exportin-5 and Dicer when the canonical pre-miRNAs were reduced (Grimm et al. 2006). Although mmu-miR-1839, an annotated shRNA (Babiarz et al. 2008), did not overexpress, mmu-miR-344e and mmu-miR-344f, novel shRNAs, did overexpress from our vector, and, as expected for shRNAs, their biogenesis was Drosha-independent (Fig. 3B; Supplemental Figs. S5–S7). Repeating the ectopic expression assay in Dicer knockout and control cells confirmed that mmu-miR-344e biogenesis was Dicer-dependent (data not shown).

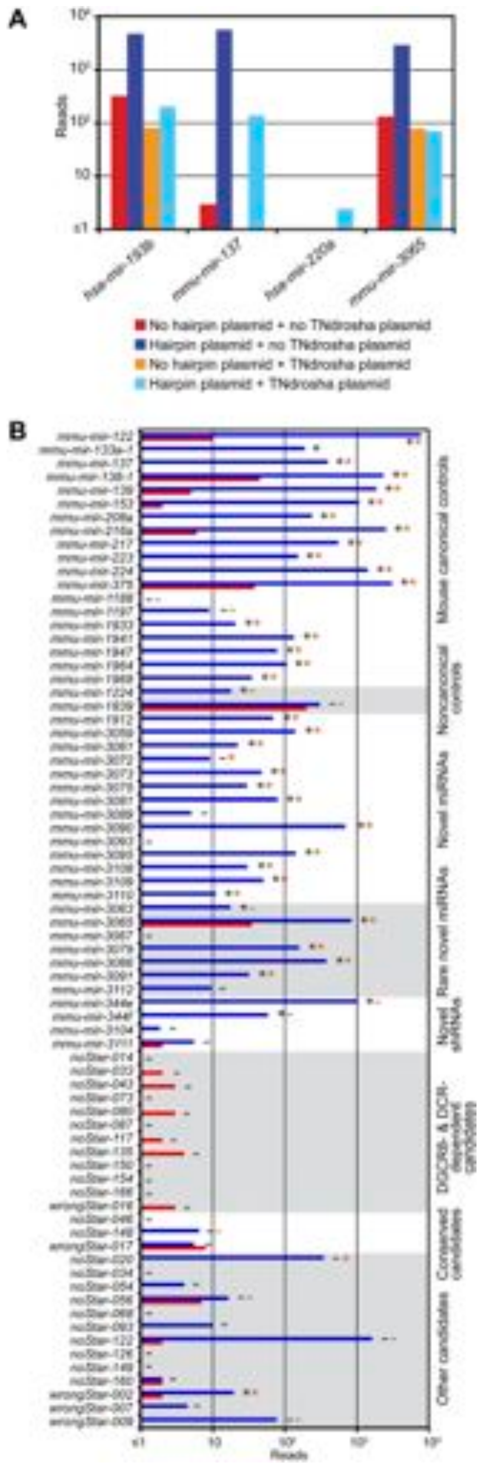


Figure 3. Experimental evaluation of novel miRNAs and candidates. (*A*) Examples of assays evaluating Drosha dependence, transfecting plasmids indicated in the key. (*B*) Assay results for control miRNAs, novel miRNAs, and miRNA candidates. Bars are colored as in *A*; detectable overexpression (black asterisks), overexpression attempted but not detected (black minus sign), detectable Drosha dependence (orange asterisks), and Drosha dependence assayed but not detected (orange minus sign) are all indicated. Shown are the results compiled from three experiments (Supplemental Figs. S5–S7).

We also evaluated our candidates that had not satisfied our criteria for confident annotation as miRNAs, usually because they lacked reads representing the predicted miRNA*. We tested three sets of these candidates. One set represented our candidates that lacked predicted miRNA* reads, yet, based on small RNA sequencing results from wild-type and mutant ES cells (Babiarz et al. 2008), appeared DGCR8- and Dicer-dependent. Another set represented candidates that appeared conserved in syntenic regions of other mammalian genomes, and the third set was selected at random from among the remaining candidates. All but one of the 28 tested candidates failed to generate miRNA-like reads, and the processing of the candidate that did generate miRNA-like reads in HEK293T cells was not dependent on Dicer, based on its presence in Dicer knockout ES cells (Babiarz et al. 2008).

The results evaluating the novel miRNAs and candidates illustrated the importance of requiring a convincing miRNA* read as a criterion for confident miRNA annotation. Five previously annotated miRNAs that were initially rejected due to lack of a convincing miRNA* read had tested positive in our overexpression assay (Fig. 2D), which indicated that this criterion was too stringent for some of the previously annotated genes. However, the results for the newly identified miRNAs and candidates showed that the presence of a convincing miRNA* read was the primary criterion that distinguished the novel canonical miRNAs (most of which tested positive) from the remaining candidates (nearly all of which tested negative). By requiring a convincing miRNA* read in addition to the other four annotation criteria, our approach accurately distinguished miRNA reads from the millions of other small RNA reads generated by high-throughput sequencing, with relatively few false positives among the novel annotations and few false negatives among the rejected candidates.

miRNA expression profiles

To compare expression levels of each miRNA in different sequenced samples, we constructed relative miRNA expression profiles (Fig. 4; Supplemental Table 5), and to compare the relative expression of various miRNAs with

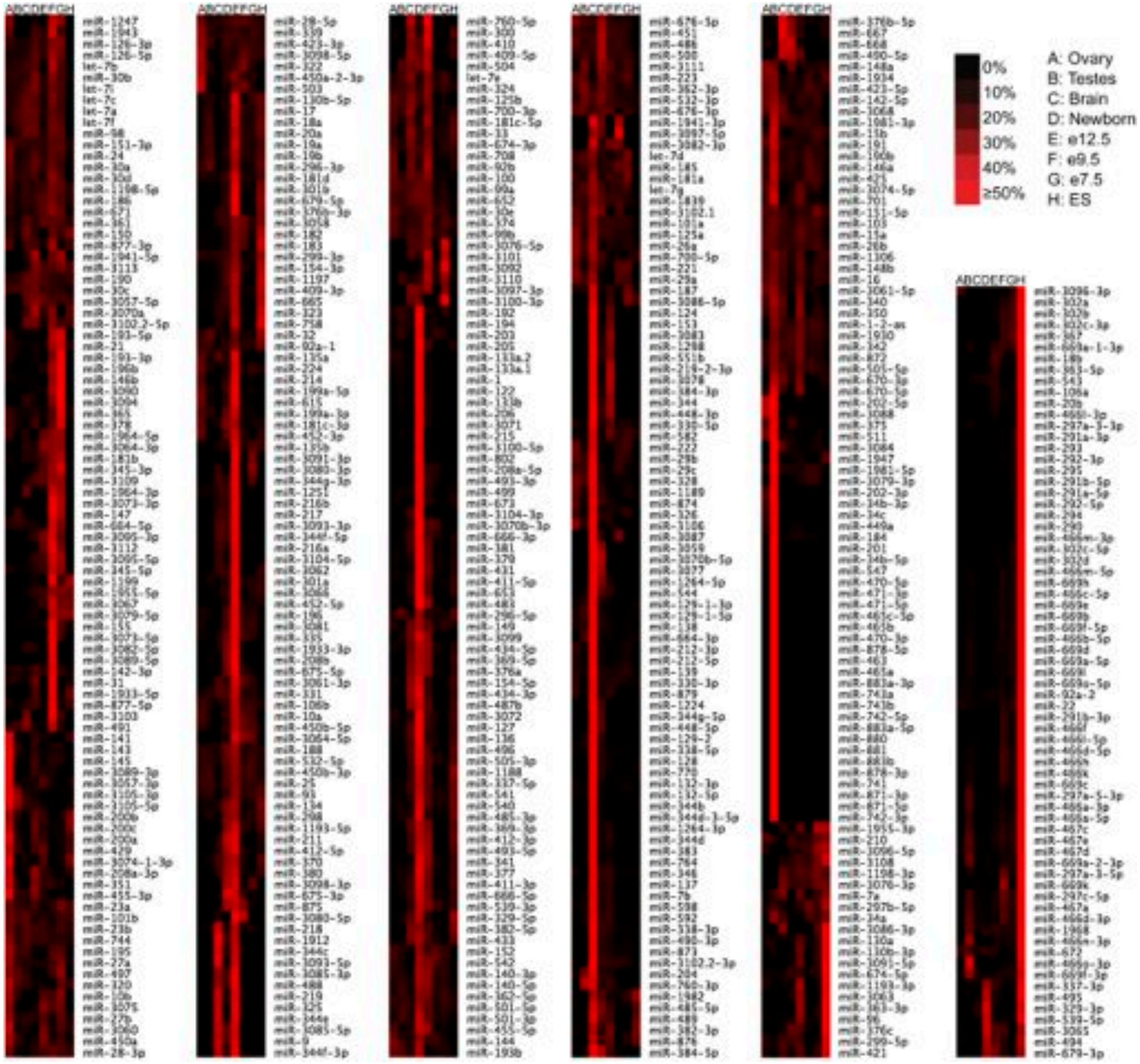


Figure 4. miRNA relative expression profiles. Profiles of mature miRNAs were constructed as described (Ruby et al. 2007b). The relative contribution of each miRNA from each sample and the sum of the normalized reads of all samples are provided (Supplemental Table 5).

each other, we generated a table of overall miRNA abundance (Supplemental Table 5). Most miRNAs had substantially stronger expression in some tissues or stages than in others, in agreement with previous observations (Wienholds et al. 2005). We expect that strong tissue- or stage-specific expression preferences inferred from our limited sample set will be revised as more tissues and stages are surveyed.

General features of mammalian miRNAs

Our analyses of high-throughput sequencing data and subsequent experimental evaluation reshaped the set of known murine miRNAs, setting aside 173 questionable annotations and adding 108 novel miRNA genes to bring the total number of confidently identified murine genes to 506. A majority (60%) of the 506 genes appeared conserved in other mammals (Supplemental Fig. S1; Supplemental Table 6). However, only 15 of the 108 novel miRNA genes were conserved in other mammals, suggesting that the number of nonconserved miRNA genes will soon surpass that of conserved ones as high-throughput sequencing is applied more deeply and more broadly.

Five novel miRNAs (*mir-3065*, *mir-3071*, *mir-3074-1*, *mir-3074-2*, and *mir-3111*) mapped to the antisense strand of previously annotated miRNAs (*mir-338*, *mir-136*, *mir-24-1*, *mir-24-2*, and *mir-374*, respectively), which, when added to the previously identified *mir-1-2*/*mir-1-2-as* pair, brings the total number of sense/antisense miRNA pairs to six. In addition, the *mir-486* hairpin has a palindromic sequence, which resulted in the same reads mapping to both the sense (*mir-486*) and antisense (*mir-3107*) hairpins. Analysis of the antisense loci of all 498 miRNA genes identified six additional loci that gave rise to some antisense reads resembling miRNAs (antisense loci of *mir-21*, *mir-126*, *mir-150*, *mir-337*, *mir-434*, and *mir-3073*). As more high-throughput data is acquired, these as well as other antisense loci are likely to be annotated as miRNA genes. However, <0.00002 of our miRNA reads corresponded to miRNAs from antisense loci (excluding the reads mapping ambiguously to *mir-486*/*mir-3107*), raising the possibility that none of the murine antisense miRNAs have a function comparable with that of miR-iab-as in flies (Bender 2008; Stark et al. 2008; Tyler et al. 2008).

Our substantially revised set of miRNA genes provided the opportunity to speak to the general features of 475 canonical miRNAs in mice, with the properties of the 295 conserved genes applying also to the conserved genes of humans and other mammals (Table 1). Most canonical miRNA genes (61%) were clustered in the genome, falling within 50 kb of another miRNA gene, on the same genomic strand. Even when excluding the four known megaclusters (Calabrese et al. 2007), which are on chromosomes 2, 12 (two clusters), and X (with 69, 35, 16, and 18 genes, respectively), a sizable fraction of the remaining genes (153 of 337) were in clusters of two to seven genes. As observed in humans (Baskerville and Bartel 2005), miRNAs from these loci within 50 kb of each other tended to have correlated expression, consistent with their processing from polycistronic pri-miRNA transcripts (Supplemental Fig. S8). In a scenario of one transcript per cluster, the 475 canonical miRNA genes would derive from 245 transcription units. In addition, many miRNA hairpins mapped to introns. Just over a third (38%) of the hairpins fell within introns of annotated mRNAs. Several lines of evidence—including coexpression correlations, chromatin marks, and directed experiments—indicate that miRNAs can be processed from introns (Baskerville and Bartel 2005; Kim and Kim 2007; Marson et al. 2008). In this scenario, as many as 107 (44%) of the 245 transcription units could double as pre-mRNAs. Other hairpins were found within transcripts that lacked other annotated functions, falling either within introns or exons, or in transcripts without evidence of splicing.

Table 1. *Properties of canonical miRNAs*

	Total	Conserved	Nonconserved
Hairpins	475	295	180
Cluster analysis			
In clusters	291	163	128
In small clusters	153	129	24
In large clusters	138	34	104
Not in clusters	184	132	52
Intron overlap			
In introns (same strand)	180	77	103
Opposite introns	22	18	4
Not in introns	273	200	73
Arm preferences			
With miRNA from 5′ arm	202	137	65
With miRNA from 3′ arm	141	102	39
With miRNAs from both arms	132	56	76

miRNA hairpins are generally thought to each give rise to a single dominant mature guide RNA. This was usually the case for the murine miRNAs, although, as in other species, this result relied on grouping together as a single functional species all the isoforms that share the same 5′ terminus. This grouping is justified based on the current understanding of miRNA target recognition, which stipulates that heterogeneity often observed at miRNA 3′ termini should have no effect on miRNA target recognition (Bartel 2009). Most mature miRNA reads (97%) were 20–24 nt in length, with 20mer, 21mer, 22mer, 23mer, and 24mer comprising 5%, 19%, 47%, 21%, and 4% of the reads, respectively (Supplemental Fig. S9). Although a single dominant mature species appears to be the most frequent outcome of miRNA biogenesis, some miRNA hairpins give rise to two or more species that each could function to target different sets of mRNAs. This expanded targeting potential arises from multiple mechanisms, including utilization of both strands of the miRNA:miRNA* duplex with similar frequency, 5′ heterogeneity, sequential Dicer cleavage, and RNA editing. Addition of untemplated nucleotides to the 3′ termini of the miRNAs can also occur, and although not thought to change targeting specificity, these changes could indicate post-transcriptional regulation of miRNA stability. Occurrence of each of these phenomena is described below.

miRNAs from both arms, with occasional tissue-specific differences in the preferred arm

Most canonical miRNA genes produced one dominant mature miRNA species, from either the 5′ or 3′ arm of the pre-miRNA hairpin, with an overall tendency to derive from the 5′ arm (Table 1), as reported for previously annotated human miRNAs (Hu et al. 2009). Some, however, yielded a similar number of reads from both arms, suggesting that the two species enter the silencing complex with similar frequencies. For these genes, mature species from the 5′ and 3′ arms were annotated using the -5p and -3p suffixes, as is conventional in such cases (Griffiths-Jones 2004). Discrimination favoring one arm over the other was less pronounced for both the nonconserved miRNAs and the less highly expressed miRNAs (Fig. 5A), although for the miRNAs with very few reads this trend was likely enhanced by our requirement for a miRNA* read. Overall, the discrimination was high, with the species from the less dominant arm comprising 4.1% of the reads that map to a miRNA or miRNA*. For the 10 most abundant miRNAs (sampling just the most abundant member in cases of repetitive miRNAs), discrimination was even higher, with the less dominant arm comprising only 1.3% of the reads. Nevertheless, the miRNA* species of these more highly expressed miRNAs were sequenced at a median frequency 13-fold greater than that of the median nonconserved miRNA, suggesting that a search for

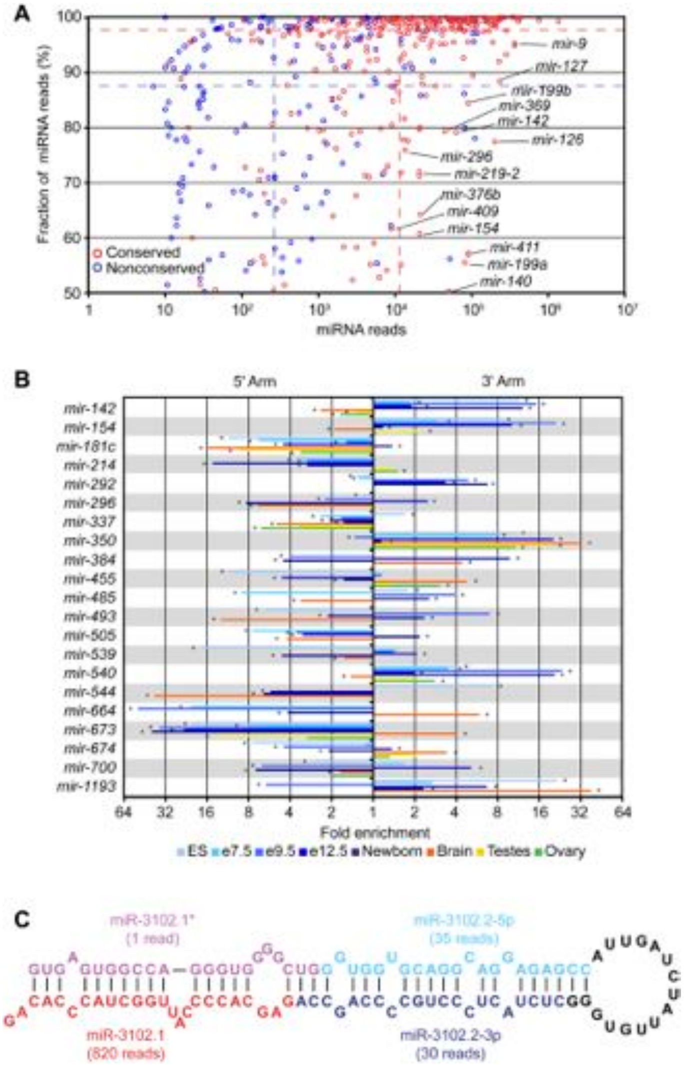


Figure 5. Reads from both arms of a hairpin, and sequential reads from the same arm. (*A*) Fraction and abundance of miRNA reads from each miRNA hairpin. To calculate the fraction, the miRNA reads were divided by the total number of miRNA and miRNA* reads, considering on each arm only the major 5′ terminus. The dashed lines indicate the median fraction of miRNA reads and the median number of miRNA reads for conserved (red) and nonconserved (blue) miRNAs. (*B*) Switching of the dominant arm in different samples. For each sample, the fold enrichment of miRNA reads produced from the 5′ arm over those produced from the 3′ arm and vice versa was calculated. Shown are results for nonrepetitive miRNAs that switch dominant arms, with at least a fivefold differential between two samples. The samples are color-coded (key), and an asterisk indicates samples with statistically significant enrichment of miRNAs produced from one arm over the other ($P < 0.05$, χ^2 test). (*C*) Sequential Dicer cleavage. Predicted secondary structure of *mmu-mir-3102* pre-miRNA (Hofacker et al. 1994).

biological function for these miRNA* species might be at least as fruitful as that for the poorly expressed nonconserved miRNAs.

If the mature miRNA accumulated preferentially from one arm of the pre-miRNA hairpin, the preferred arm generally remained consistent across the various libraries. For a few miRNAs, however, the preferred arms switched between samples (Fig. 5B), as reported previously using PCR-based miRNA quantification (Ro et al. 2007). For example, miR-142-5p was sequenced more frequently in ovary, testes, and brain, and miR-142-3p was sequenced more frequently in embryonic and newborn samples. These results imply a developmental switch in targeting preferences. A similar arm-switching phenomena has been reported for a sponge miRNA (Grimson et al. 2008), and was observed for 20 other nonrepetitive mouse miRNA genes (Fig. 5B).

Sequential Dicer cleavage of a mirtron hairpin

In plants, a few pri-miRNA hairpins with long, continuous RNA duplexes are cleaved sequentially by Dicer to generate two adjacent miRNA:miRNA* duplexes (Kurihara and Watanabe 2004; Rajagopalan et al. 2006). Those precursors bear little resemblance to the shorter, imperfectly base-paired hairpins of metazoan miRNA genes. In mice, similar precursors are found in the form of hairpin siRNA (hp-siRNA) precursors, but their expression appears to be limited to germline tissues and totipotent ES cells, which lack a robust interferon response to intracellular dsRNA (Babiarz et al. 2008; Tam et al. 2008; Watanabe et al. 2008). However, we detected two miRNA:miRNA* duplexes deriving from the *mmu-mir-3102* pre-miRNA hairpin, an apparent mirtron as evidenced by reads mapping to both boundaries of an intron (Fig. 5C; Supplemental Table 3). After splicing and debranching, the excised intron was predicted to fold into a 104-nt pre-miRNA hairpin—substantially longer than the average pre-miRNA length of 61 nt (calculated from the set of confirmed miRNAs). Reads from this locus suggested that Dicer cleaved this pre-miRNA twice, with the first cut generating the outer miRNA:miRNA* duplex and the second cut generating the inner miRNA: miRNA* duplex (Fig. 5C). The inner miRNA (miR-3102.2-3p) was among a set of proposed miRNA candidates (Berezikov et al. 2006b), but the most frequently sequenced species from this hairpin was the outer miRNA (miR-3102.1) (Fig. 5C). Of the 16 genomes examined, the extended *mir-3102* hairpin with both the inner and outer miRNAs appeared conserved only in rats, although the orthologous loci in cows, dogs, and humans also could fold into shorter hairpins, with miR-3102.1 potentially conserved in cows.

We suspect that it is more than a coincidence that the single metazoan example of a sequentially diced miRNA is initially processed by the spliceosome rather than by Drosha. One way to explain this observation is that DGCR8/Drosha interacts directly with the loop of pri-miRNA stem–loops when recognizing its substrates (Zeng et al. 2005), and that the lack of sequentially diced Drosha-dependent miRNA hairpins in animals reflects the limited reach of this complex.

5′ Heterogeneity

Most conserved miRNAs had very precise 5′ processing, with alternative 5′ isoforms comprising only 8% of all miRNA reads (Fig. 6A,B). These results, analogous to those observed in worms and flies (Ruby et al. 2006, 2007b), are consistent with the idea that selective pressure to avoid off-targeting acts to optimize precision of the cleavage event that produces the 5′ terminus of the dominant species so as to prevent a consequential number of molecules with seed sequences in the wrong register. Moreover, 5′ termini of conserved miRNAs were more precise than those of miRNA* reads (4% and 12% offset reads, respectively, excluding those that produce comparable numbers of small RNAs from each arm). For cases in which Dicer produced the 5′ terminus of the miRNA, the Dicer cut appeared somewhat more precise than the Drosha cut (5% offset reads for miRNAs on the 3′ arm, compared with 7% offset reads for miRNA* on the 5′ arm), hinting that features of the pre-miRNA structure may supplement the distance from the Drosha cut as determinants of Dicer cleavage specificity (Ruby et al. 2006, 2007b).

A few miRNAs had less uniform 5′ termini (Fig. 6A,B). For some miRNAs, 5′ heterogeneity has been documented previously (Ruby et al. 2007b; Stark et al. 2007; Azuma-Mukai et al. 2008; Wu et al. 2009), the most prominent example being hsa-miR-124, a conserved neuronal miRNA for which the 5′-shifted isoform was initially annotated as the miRNA and eventually replaced by the more prominent isoform following more extensive sequencing (Lagos-Quintana et al. 2002; Landgraf et al. 2007). Another prominent miRNA with unusually diverse 5′ termini was miR-133a. This conserved miRNA, which is highly expressed in heart and muscle, had a second dominant isoform (miR-133a.2) that was shifted 1 nt downstream from the annotated miRNA (miR-133a.1) (Fig. 6C; Supplemental Table 3). To test whether this heterogeneity might be explained by differential processing of the two *mir-133a* paralogous hairpins, as observed for the two *Drosophila mir-2* hairpins (Ruby et al. 2007b), we tested the two *mir-133a* hairpins in our ectopic expression assay. Although *mir-133a-1* was somewhat more prone to produce the miR-133a.2 isoform, both hairpins produced a substantial amount of both isoforms (Fig. 6C).

To investigate the functional consequences of miRNA 5′ heterogeneity, we examined published array data showing the responses of mRNAs after deleting either *mir-223*, a miRNA with substantial heterogeneity, or *mir-155*, a miRNA with little heterogeneity. miR-223 is highly expressed in neutrophils, and analysis of small RNA sequences from isolated neutrophils (Baek et al. 2008) was consistent with our sequencing results (Supplemental Table 3) in showing 5′ heterogeneity, with 81% of the reads mapping to the 5′ end of the major isoform miRNA and 12% mapping to the 5′ end of a second isoform that was shifted by 1 nt in the 3′ direction (Fig. 6D). As expected, mRNAs with canonical 7–8mer sites (Bartel 2009) matching the seed of the major isoform were significantly derepressed in the *mir-223* deletion mutant ($P < 10^{-12}$, Kolmogorov–Smirnov [K–S] test, compared with no site distribution). mRNAs with canonical sites matching the minor isoform also showed a significant tendency to be derepressed, albeit to a lesser degree ($P = 0.0022 \times 10^{-7}$, 0.013×10^{-7}, and 1.7×10^{-7}, for 8mer, 7mer-m8, and 7–8mers combined, respectively) (Fig. 6D). This result could not be attributed to the overlap between sites matching the major and minor isoforms because all mRNAs with a 6mer seed match to the major isoform (ACUGAC) were excluded, and additional analyses ruled out participation of the "shifted 6mer" match (Friedman et al. 2009) to the major isoform (AACUGA) (Supplemental Fig. S10A). Analogous analysis of miR-155 yielded strong evidence for function of the major isoform (Rodriguez et al. 2007) but no sign of function for the minor isoform, which comprised very few (1%) of our miR-155 reads (Fig. 6E; Supplemental Table 3).

Taken together, our results show that some miRNAs have alternative 5′ miRNA isoforms that are expressed at levels sufficient to direct the repression of a distinct set of endogenous targets and thereby broaden the regulatory impact of the miRNA genes. Therefore, we suggest that, rather than choosing one isoform over the other for annotation as the authentic miRNA, more of these alternative isoforms should be annotated, with the expectation that, for some highly expressed miRNAs, more than one 5′ isoform contributes to miRNA function.

RNA editing

RNA editing in which adenosine is deaminated and thereby converted to inosine (I) has been reported for

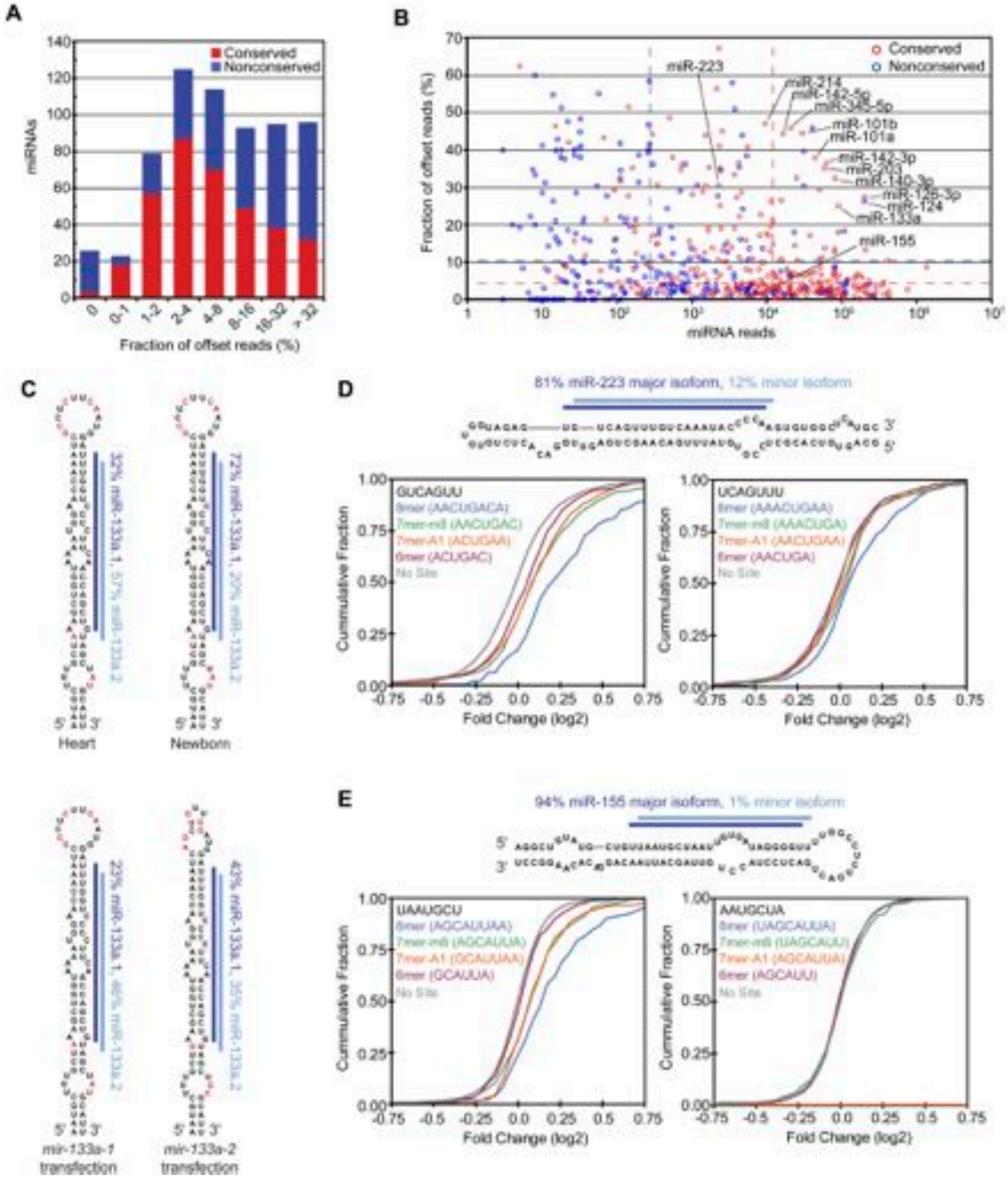


Figure 6. miRNAs with 5′ heterogeneity. (*A*) The distribution of conserved (red) and nonconserved (blue) miRNAs with reads ≤5 nt offset at their 5′ terminus. (*B*) The fraction of offset reads and abundance of reads for each miRNA hairpin, colored as in *A*. The dashed lines indicate the median level of reads for conserved (red) and nonconserved (blue) miRNAs. (*C*) 5′ Heterogeneity of miR-133a. Data from mouse heart (Rao et al. 2009) and newborn are mapped to the *mir-133a-1* hairpin (*top*), and data from the ectopic expression assay are mapped to the indicated transfected hairpin (*bottom*). The lines indicate miR-133a.1 (dark blue) and miR-133a.2 (light blue), and red nucleotides indicate those that differ between *mir-133a-1* and *mir-133a-2*. (*D*) Effect of losing miR-223 on messages with 3′ UTR sites for miR-223 major and minor isoforms. (*Top*) Small RNA sequencing data from mouse neutrophils (Baek et al. 2008) were mapped to the *mir-223* hairpin as in *C*. For each set of messages with the indicated 3′ UTR site for miR-233 (major isoform sites, *bottom left*; minor isoform sites, *bottom right*), the fraction that changed at least to the degree indicated following loss of miR-223 is plotted, using data published for neutrophils differentiated in vivo (Baek et al. 2008). (*E*) Effect of losing miR-155 on messages with 3′ UTR sites for miR-155 major and minor isoforms, plotted as in *D* using published data from T cells (Rodriguez et al. 2007). (*Top*) Sequencing data from our study are mapped to the *mir-155* hairpin as in *C*. The mRNAs with 8mer and 7mer-A1 sites for the minor isoform were excluded from the analysis because these sites overlapped with 7mer-m8 sites for the major isoform.

some miRNA precursors (Blow et al. 2006; Landgraf et al. 2007; Kawahara et al. 2008). Because I pairs with C, such edits could change miRNA target recognition. Reasoning that the mammalian adenosine deaminases (ADARs) responsible for A-to-I editing are expressed primarily in the brain, we searched for sequencing reads from the brain that did not match the genome and had as their closest match a mature miRNA or miRNA*. After filtering for mismatches occurring >2 nt from the 3′ end, a step taken to avoid considering instances of untemplated 3′-terminal addition, only 4% of the reads had single mismatches to the genome (Supplemental Fig. S11A). Moreover, the fraction of sequences with A-to-G changes (indicative of A-to-I editing) was only 0.61%, a fraction resembling that of other mismatches (Supplemental Fig. S11A). This fraction was also similar to that of the A-to-G changes in our synthetic internal standards used for preparing the sequencing libraries. These results indicate that mature edited miRNAs are very rare and difficult to distinguish above the background level of sequencing errors. The low frequency of editing in mature miRNAs was consistent with the findings that edited processed miRNAs are more than fourfold less common in mice relative to humans (Landgraf et al. 2007), and are less common than edited miRNA precursors (Kawahara et al. 2008). The latter observation might be due to rapid degradation or impaired processing, which has been shown for miR-142 (Yang et al. 2006) and miR-151 (Kawahara et al. 2007a).

Although editing did not appear to be a widespread phenomenon among all mature miRNAs, editing at specific sites might still be important for a few individual miRNAs. To investigate this possibility, mismatch fractions were calculated as the fraction of reads bearing a particular mismatch over all reads covering that genomic position. For each library, a change was considered significant if the fraction exceeded 5% and at least 10 reads contained the mismatch. Additional filters designed to remove sequencing errors, alignment artifacts, and instances of untemplated nucleotide addition preferentially retained A-to-G changes while removing nearly all other events (Supplemental Fig. S11B). Sixteen A-to-G events passed the filters and subsequent manual examination, all of which occurred only in the brain library (Table 2). Five of these inferred editing sites were also observed in a low-throughput sequencing effort in human brain samples (Kawahara et al. 2008), indicating that editing of some miRNAs is conserved between mammals. Consistent with that study, eight of 16 editing sites occurred in a UAG motif. A separate examination of read alignments with up to three mismatches showed that the vast majority of edited reads were edited at one position, suggesting that either editing of multiple sites in the same RNA molecule is rare, or multiply edited RNAs are degraded more rapidly.

A-to-I editing of a seed nucleotide would dramatically affect targeting. In addition to editing in the miR-376 cluster described previously (Kawahara et al. 2007b, 2008), we found another eight miRNAs that are edited within the seed of either the miRNA or the miRNA*. A-to-I editing could also affect miRNA loading, and thereby indirectly affect targeting. Indeed, the editing of miR-540 might help explain why the 5′ arm is more abundant in the brain than in other tissues, although editing is too infrequent to fully explain the switch in strand bias. Altering Drosha and Dicer processing could also indirectly affect targeting. Analysis of 5′ ends showed that seven of 16 instances of editing were associated with a statistically significant ($P < 0.05$) shift in the 5′ nucleotide, presumably due to changes in the Drosha and Dicer cleavage site (Supplemental Fig. S11D).

Table 2. *Inferred A-to-I editing sites in miRNAs*

miRNA	Position	Fraction edited
miR-219-2-3p	15	0.064
miR-337-3p	10	0.062
miR-376a*	4	0.297
miR-376b-3p	6	0.501
miR-376c	6	0.311
miR-378	16	0.087
miR-379*	5	0.095
miR-381	4	0.125
miR-411-5p	5	0.239
miR-421	14	0.054
miR-467d	3	0.094
miR-497	2	0.104
miR-497*	20	0.699
miR-540*	3	0.080
miR-1251	6	0.431
miR-3099	7	0.209

Untemplated nucleotide addition

Much more prevalent than editing of internal nucleotides was addition of untemplated nucleotides to miRNA 3′ termini. As reported previously for miRNAs in mammals (Landgraf et al. 2007), and also observed for those of worms and flies (Ruby et al. 2006, 2007b), nucleotides most frequently added to murine miRNAs were U and A (Fig. 7A). Addition of C or G was no higher than background, as estimated by monitoring apparent addition to tRNA fragments (Fig. 7A). Possible sources of the background rate could be sequencing error, transcription error, or a low level of biological nucleotide addition. Some miRNAs were much more frequently extended than others (Supplemental Table 7). One very frequently extended miRNA was miR-143, for which the extended reads outnumbered the nonextended ones (196,565 compared with 114,980 reads, respectively).

For extension by U, RNAs from the pre-miRNA 3′ arm were three times more frequently extended than were those from the 5′ arm (Fig. 7A,B, $P = 2.3 \times 10^{-4}$, K–S test). This preference, not observed for the A extension (Fig. 7A,C), suggests that much of the U extension occurs to the pre-miRNA, prior to Dicer cleavage—a state in which the 3′ arm but not the 5′ arm would be available for extension (Fig. 7D). TUT4-catalyzed poly(U) addition to the *let-7* pre-miRNA, which is specified by Lin28, plays an important role in post-transcriptional repression of *let-7* expression (Heo et al. 2008, 2009; Hagan et al. 2009). Our analyses indicating untemplated U extension to many other pre-miRNAs hint that this type of regulation may not be

A

	5' Arm	3' Arm	tRNA
A	7.3% (343)	4.5% (318)	0.9% (186)
C	0.2% (348)	0.2% (318)	0.3% (186)
G	0.2% (288)	0.2% (336)	0.5% (186)
T	6.5% (287)	19.9% (324)	2.8% (186)

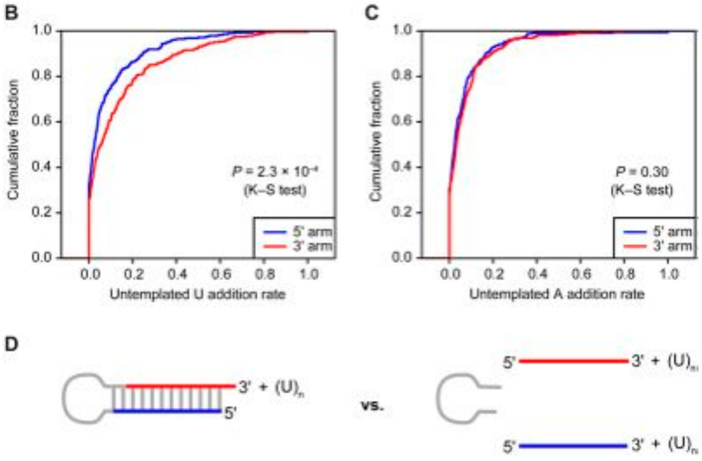


Figure 7. Untemplated nucleotide addition. (*A*) Untemplated nucleotide addition rate for miRNA and miRNA* reads from the indicated arm. Rates for each miRNA are provided (Supplemental Table 6). As a control, tRNA degradation fragments were analyzed similarly. Numbers of genes analyzed are indicated in parentheses. (*B*) Distribution of rates for untemplated U addition to RNAs from the 5′ arm (blue) and from the 3′ arm (red). (*C*) Distribution of rates for untemplated A addition to RNAs from the 5′ arm (blue) and from the 3′ arm (red). (*D*) Schematic of the biogenesis stage in which U could be added to the RNA of only one arm (pre-miRNA, *left*), and the stage in which U could be added to the RNA of either arm (mature miRNA and miRNA*, *right*).

limited to *let-7*, but that analogous pathways, presumably using mediators other than Lin28, act to regulate the expression of other murine miRNAs.

Discussion

The status of miRNA gene discovery in mammals

Our current study sets aside nearly a third (173 of 564) of the miRBase version 14.0 gene annotations for lack of convincing evidence that these produce authentic miRNAs. It also adds another 108 novel miRNA loci, raising the question of how many more authentic loci remain undiscovered. This question is difficult to answer. Ever since the recognition that the poorly conserved miRNAs are also the ones expressed at lower levels in mammals, and thus are the most difficult to detect by both computational and experimental methods, we have known that it is impossible to provide a meaningful estimate of the number of mammalian miRNA genes remaining to be discovered (Bartel 2004). The broadly conserved miRNAs are another matter. Only three of the 88 novel canonical miRNAs had recognizable orthologs sequenced in chickens, lizards, frogs, or fish, and these three were antisense to previously annotated broadly conserved miRNA genes. Therefore, apart from miRNAs expressed at very low levels from the antisense strand of known genes, we suspect that the list of broadly conserved miRNA gene families is nearing completion. The current set of murine miRNA genes includes 192 genes that fall into 89 broadly conserved miRNA gene families (Supplemental Table 6).

Another 107 miRNA gene families appeared conserved in other mammals (Supplemental Table 6). These were represented by 120 murine genes, including 14 novel genes. Of these novel genes, 11 were founding members of novel conserved gene families. Some of these were identified with only 11 reads, indicating that additional pan-mammalian gene families remain to be found, although we have no evidence supporting the idea that the number of conserved gene families will rise to the very high levels suggested by some earlier computational studies (Berezikov et al. 2005, 2006b; Xie et al. 2005). For now, we can say that mammals have at least 196 conserved miRNA gene families represented in mice by at least 312 pre-miRNA hairpins (303 canonical and nine noncanonical hairpins) produced from at least 194 unique transcription units.

Because a single miRNA hairpin can produce multiple functional isoforms, generated by either 5′ processing heterogeneity or utilization of both arms of the miRNA duplex, a single conserved hairpin can produce more than one conserved miRNA isoform. Because the different isoforms have different seed sequences, they fall into different families of mature miRNAs. Thus, the number of conserved families of miRNAs (i.e., mature guide RNAs) will exceed the number of conserved families of genes (i.e., hairpins). Perhaps the best known example of a hairpin with two broadly conserved isoforms is *mir-9*, for which conserved miRNAs from both arms of the hairpin are readily detected by using in situ hybridization in both zebrafish and marine annelids (Wienholds et al. 2005; Christodoulou et al. 2010). Numerous conserved genes produce more than one miRNA isoform (Figs. 5A, 6A), but for most of these we do not yet know whether production of the alternative isoform is conserved in other species. High-throughput sequencing from other species will help identify many additional conserved isoforms. We anticipate that the discovery of multiple conserved isoforms will contribute much more to the future growth in the list of broadly conserved miRNA families than will the discovery of new conserved genes.

As expected, the conserved miRNAs tended to be expressed at much higher levels than were the nonconserved ones, with the median read frequency of conserved miRNAs 44-fold greater than that of the nonconserved miRNAs (Figs. 5A, 6B). Therefore, even if many nonconserved miRNA genes remained to be found, these would add little to the number of annotated miRNA molecules in a given cell or tissue, and presumably even less to the impact of miRNAs on gene expression (Bartel 2009). Indeed, even more pressing than the question of how many poorly conserved miRNAs remain undetected is the question of whether any of the known poorly conserved miRNAs have any consequential function in the animal.

Most of these poorly conserved miRNAs could have derived from transcripts that fortuitously acquired hairpin regions with features needed for some Drosha/Dicer processing. In this scenario, most of these newly emergent miRNAs will be lost during the course of evolution before ever acquiring the expression levels needed to have a targeting function sufficient for their selective retention in the genome. Consistent with the hypothesis that most of these miRNAs play inconsequential regulatory roles, these miRNAs generally accumulated to much lower levels in our ectopic expression assay, (Fig. 3B, median read frequencies of 58 and 844 for nonconserved and conserved miRNAs, respectively), and they displayed weaker specificity for one arm of the hairpin (Fig. 5A), as would be expected if there was no advantage for the cell to efficiently use their respective hairpins. Nonetheless, some were processed efficiently, and at least a few poorly conserved miRNAs probably have acquired consequential species-specific functions. Although none have known functions, such hairpins are worthy of annotation as miRNA loci (just as protein-coding genes can be annotated before the protein is known to be functional), and as a class these newly emergent miRNAs could provide an important evolutionary substrate for the emergence of new regulatory activities.

The major challenge for miRNA gene discovery stems from the difficulty in proving that a nonconserved, poorly expressed candidate is an authentic miRNA, combined with the even greater difficulty in proving that a questionable candidate is not an authentic miRNA. This challenge has become all the more acute as miRNA discovery has reached the point to which nearly all of the novel candidates are both nonconserved and poorly expressed. Our approach of testing pools of candidates in an ectopic expression assay provides useful data for evaluating miRNA authenticity. However, our approach cannot provide conclusive proof for or against the authenticity of a proposed candidate, leaving open the possibility that some of the nonconserved, poorly expressed candidates that we classify as "confidently identified miRNAs" are false positives. When considering the limitations of the current tools for miRNA gene identification, this possibility cannot be avoided. Therefore, if any nonconserved, poorly expressed miRNAs are annotated as miRNAs, the resulting list of miRNAs will have to be somewhat fuzzy, with an expectation that some of the annotated genes will not be authentic miRNAs. This expectation should not be viewed as advocating the indiscriminant annotation of all candidates as miRNAs. Our proposal is that miRNA gene discovery efforts should annotate as miRNAs only those novel candidates that both are found in high-thoughput sequencing libraries and pass a set of criteria that is sufficiently stringent such that a majority of the novel canonical miRNAs are cleanly processed in a Drosha-dependent manner when using the ectopic expression assay. Although implementing this proposal would not prevent all false positives from entering the databases, it would preserve a higher quality set of miRNAs while eliminating few authentic annotations. Those wanting to take additional measures to avoid false positives could focus on only the subset of miRNAs that both meet these criteria and are conserved in other species.

Unknown features required for Drosha/Dicer processing

Before learning the results of our experiments, we wondered whether any ectopically overexpressed hairpin of suitable length would be processed as if it were a miRNA, a result that would have rendered our assay too permissive to be of value. In this scenario, most of the specificity that distinguished authentic miRNA genes from other regions of the genome with the potential to produce transcripts that fold into seemingly miRNA-like hairpins would have been a function of whether or not the regions were transcribed. This scenario was not realized, however, and our assay turned out to be informative, which illustrates how much of Drosha/Dicer substrate recognition still remains unknown. Many of the previously proposed miRNA hairpins that had no reads in our mouse samples were indistinguishable from authentic miRNA hairpins with regard to the known determinants for Drosha/Dicer recognition, yet none of these unconfirmed hairpins produced miRNA and miRNA* molecules in our very sensitive assay (Fig. 2C,D). These results showing that major processing specificity determinants still remain undiscovered point to the importance of finding these determinants—efforts that, if successful, will mark the next substantive advance in accurately predicting and annotating metazoan miRNAs.

Materials and methods

Library preparation

Total RNA samples from mouse ovary, testes, and brain were purchased from Ambion, and total RNA from mouse E7.5, E9.5, E12.5, and newborn were obtained from the Chess laboratory. The small RNA cDNA libraries were made as described (Grimson et al. 2008), except for the 3′ adaptor ligation, which was 5′ adenylated pTCGTATGCCGTCTTCTGCTTGidT. For a detailed protocol, see http://web.wi.mit.edu/bartel/pub/protocols.html.

miRNA discovery

The reads with inserts of 16–27 nt were processed as described (Babiarz et al. 2008). The miRNA candidates were identified using reads matching genomic regions that were not very highly repetitive (reads with <500 genomic matches). Reads from all data sets were combined and grouped by their 5′-terminal loci, requiring that each candidate 5′ locus pass five criteria listed in the text. (1) To pass the expression criterion, a candidate required ≥10 normalized reads. (2) To address the hairpin requirement, the secondary structure of the candidate was evaluated by selecting for each 5′-terminal locus the most abundant sequence and extending its 5′ end by 2 nt to define the range of the potential miRNA/miRNA* duplex. Three genomic windows were extracted with the 5′ end extended an additional 10 nt and the 3′ end extended either 50 nt, 100 nt, or 150 nt. Three more windows were extracted extending the 3′ end by 10 nt and the 5′ end another 50 nt, 100 nt, or 150 nt. The secondary structure of each of the six windows was predicted using RNAfold (Hofacker et al. 1994), and the number of hairpin base pairs (denoted using bracket notation) involving the 5′-extended miRNA candidate was calculated as the absolute value of ([number of 5′-facing brackets] – [number of 3′-facing brackets]). A candidate with a minimum of 16 bp using at least one of the six genomic windows satisfied the hairpin criteria. (3) The candidates with non-miRNA biogenesis were found by mapping to annotated noncoding RNA loci (rRNA, tRNA, snRNA, and srpRNA). (4) The candidates likely produced by degradation were defined as those failing the 5′ homogeneity requirement. A candidate satisfied the 5′ homogeneity requirement if at least half of the reads within 30 nt of the candidate 5′ end were present within 2 nt of the candidate 5′ end and if the candidate 5′ end comprised at least half of the reads within 2 nt of the candidate 5′ end, or if there was only one other 5′ end within 30 nt of the candidate 5′ end that had more than half of the reads mapping to the candidate 5′ end. (5) Manual inspection of reads mapped to predicted secondary structures identified candidates accompanied by potential miRNA* reads. For 10 previously annotated miRNAs and seven novel miRNAs, a suitable miRNA* read was found only after considering alternative hairpin folds predicted to be suboptimal using mfold (Mathews et al. 1999; Zuker 2003).

For the analysis of *mir-290*, *mir-291a*, *mir-291b*, *mir-292*, *mir293*, *mir-294*, and *mir-295*, which are not present in mm8 genome assembly, we mapped all reads to mm9 genome assembly corresponding to the region [chr7(+): 3,218,627–3,220,842].

For conservation analysis, a candidate was considered broadly conserved if the hairpin structure and the seed sequence were conserved to chickens, fish, frogs, or lizards (galGal3, danRer5, xenTro2, and anoCar1, respectively) in the University of California at Santa Cruz whole-genome alignments (Kuhn et al. 2009). To identify a candidate conserved in mammals, we looked at 12 additional genomes (bosTau3, canFam2, cavPor2, equCab1, hg18, loxAfr1, monDom4, ornAna1, panTro2, ponAbe2, rheMac2, and rn4) and calculated the branch length score from a phylogenetic tree trained on mouse 3′ UTR data (Friedman et al. 2009), using the cutoff score of 0.7. A gene was considered to be in a conserved miRNA gene family if the hairpin produced a miRNA with a seed matching that of a conserved miRNA (Supplemental Table 6).

Ectopic overexpression assays

To generate expression constructs, pre-miRNA hairpins and the surrounding regions were amplified from human genomic DNA (NCI-BL2126) or from mouse BL6 genomic DNA using Pfu Ultra II polymerase (Stratagene) and primers with Gateway (Invitrogen)-compatible ends designed to anneal ~100 nt upstream of and downstream from the miRNA hairpins. PCR products were inserted into Gateway vector pDONR221 and subsequently into pcDNA3.2/V5-DEST, and the resulting plasmids were transformed into DH5-α cells. Positive clones were selected by colony PCR and were sequenced. Clones that did not have a mutation within pre-miRNA hairpins were selected. Plasmid DNA from the confirmed expression clones was purified for transfection using the Plasmid Mini Kit (Qiagen). For each standard assay, plasmids for up to 10 hairpin expression constructs were mixed in equal amounts to create seven or eight pools of ~1.4 μg of DNA each, with each pool including one to three positive control hairpins.

HEK293T cells were cultured in DMEM supplemented with 10% FBS, and were plated in 12-well plates ~24 h prior to transfection to reach ~80%–90% confluency. Each well of cells was transfected with one pool of DNA using Lipofectamine 2000 (Invitrogen). For the standard assays, 145–200 ng of pMaxGFP (Amaxa) was cotransfected with each pool to enable transfection efficiency to be confirmed by GFP expression. Control wells (no hairpin plasmid) were transfected only with 145 ng of pMaxGFP. For the Drosha/Dicer dependency assays, seven to eight hairpin constructs were combined to create six pools of ~400 ng each. Each pool was mixed with 1.2 μg of the pCK-Drosha-Flag(TN) (TNdrosha), pCK-Flag-Dicer(TN) (TNdicer), or pCK-dsRed.T4 (control vector, constructed by replacing the Drosha-coding sequence of TNdrosha with dsRed-coding sequence) and used to transfect one well of HEK293T cells as above. Control wells were transfected with 1.2 μg of either TNdrosha, TNdicer, or control vector. For the dependency assays, each transfection was performed in duplicate wells. Cells from all assays were harvested 39–48 h after transfection. Cells from each treatment were combined, total RNA was extracted using TriReagent (Ambion), and small RNA libraries were prepared for Illumina sequencing.

The reads were processed as above, and RNA species were matched to the transfected hairpins. In the standard assay, reads were normalized by the median of the 30 most frequently sequenced endogenous miRNAs. For assays testing Drosha/Dicer dependency, reads were normalized based on the number of reads corresponding to an 18-nt internal standard that had been spiked into equivalent amounts of total RNA prior to beginning library preparation. Reads matching the transfected hairpins were grouped by their 5′ termini (5′-terminal locus). The locus with the largest number of reads was considered the 5′-terminal locus of the mature miRNA produced by the hairpin, and similarly, the most dominant 5′ locus on the opposite arm was considered the miRNA*. The normalized miRNA and miRNA* read numbers were summed to calculate the expression level.

If an overexpressed hairpin generated mature miRNA with the dominant 5′-terminal locus corresponding to the expected locus and at least one read corresponding to the miRNA* with an ~2-nt 3′ overhang, it was considered expressed. A hairpin was classified as overexpressed if there were at least threefold more reads in the hairpin transfection than in the control transfection, after adding psuedocounts of five to both. A hairpin was classified as Drosha- or Dicer-dependent if the knockdown was at least threefold.

Identification of arm-switching miRNAs

To determine the read numbers from the 5′ and 3′ arms, reads from each sample were grouped based on their 5′ termini, and the read numbers were tallied for those corresponding to the miRNA or miRNA* 5′ terminus. Only samples with five or more reads on either arm were considered. The fold enrichment was calculated as the ratio of 5′ and 3′ arm reads after adding pseudocounts of one.

RNA editing analysis

Sequencing libraries from individual tissues were combined and mapped to the genome using the Bowtie alignment tool (Langmead et al. 2009). The alignments were filtered for sequences that uniquely aligned to the genome, contained at most one mismatch to the genome, and had 5′ ends that mapped to within 1 nt of an annotated miRNA or miRNA* 5′ end. The 12 possible mismatch types were then quantified at each position covered by the filtered reads. For example, to screen for A-to-G mismatches indicative of A-to-I editing sites, the editing fraction was calculated as the number of reads containing an A-to-G mismatch at a particular position, divided by the number of filtered reads covering that position. Sites were considered editing candidates if the editing fraction was >5%, had at least 10 A-to-G mismatch reads, and did not occur in the last 2 nt of the corresponding miRNA or miRNA*. Candidate editing sites were then manually examined and discarded if an alternative explanation was more parsimonious. For example, the only nonbrain editing candidate mapped to let-7c-1, but was most likely due to a handful of let-7b reads containing untemplated nucleotide additions that fortuitously matched the let-7c-1 locus. Consistent with this explanation, the putatively edited reads were unusually long and at unusually low abundance. Candidate editing sites were also checked in the Perlegen SNP database (Frazer et al. 2007) and dbSNP; no editing candidates corresponded to known SNPs.

Untemplated nucleotide analysis

To examine untemplated nucleotide addition, non-genome-mapping reads were filtered for those that matched miRNA or miRNA* sequences but also included a nongenomic poly(N) at the 3′ end. The untemplated nucleotide addition rate was calculated as the ratio of reads with the untemplated nucleotide to the sum of the reads with and without the untemplated nucleotide. After excluding miRNAs that map to multiple loci, and any miRNAs or miRNA*s with a genomic T at the position immediately 3′ of the annotated sequence, there were 343 miRNA/miRNA* species with untemplated U on the 5′ arm and 318 on the 3′ arm. Similarly, there were 287 5′ arm species with untemplated A on the 5′ arm and 324 on the 3′ arm. The background tRNA untemplated U addition rate was calculated similarly. A two-sided K–S test was used to assess significant differences in distributions.

Accession numbers

All small RNA reads are available at the GEO database with accession number GSE20384.

Acknowledgments

We thank N. Lau and A. Chess for embryonic and newborn total RNA, R. Friedman for calculating branch length scores for the analysis of conservation, A. Marson and N. Hannet for technical advice, and V.N. Kim for TNdrosha and TNdicer plasmids. This work was supported by a grant from the NIH (GM067031) to D.B.

References

Azuma-Mukai A, Oguri H, Mituyama T, Qian ZR, Asai K, Siomi H, Siomi MC. 2008. Characterization of endogenous human Argonautes and their miRNA partners in RNA silencing. *Proc Natl Acad Sci* **105:** 7964–7969.

Babiarz JE, Ruby JG, Wang YM, Bartel DP, Blelloch R. 2008. Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes & Dev* **22:** 2773–2785.

Baek D, Villén J, Shin C, Camargo FD, Gygi SP, Bartel DP. 2008. The impact of microRNAs on protein output. *Nature* **455:** 64–71.

Bartel DP. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* **116:** 281–297.

Bartel DP. 2009. MicroRNAs: Target recognition and regulatory functions. *Cell* **136:** 215–233.

Baskerville S, Bartel DP. 2005. Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. *RNA* **11:** 241–247.

Bender W. 2008. MicroRNAs in the *Drosophila* bithorax complex. *Genes & Dev* **22:** 14–19.

Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat Genet* **37:** 766–770.

Berezikov E, Guryev V, van de Belt J, Wienholds E, Plasterk RHA, Cuppen E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21–24.

Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RHA. 2006a. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet* **38:** 1375–1377.

Berezikov E, van Tetering G, Verheul M, van de Belt J, van Laake L, Vos J, Verloop R, van de Wetering M, Guryev V, Takada S, et al. 2006b. Many novel mammalian microRNA candidates identified by extensive cloning and RAKE analysis. *Genome Res* **16:** 1289–1298.

Berezikov E, Chung WJ, Willis J, Cuppen E, Lai EC. 2007. Mammalian mirtron genes. *Mol Cell* **28:** 328–336.

Blow MJ, Grocock RJ, van Dongen S, Enright AJ, Dicks E, Futreal PA, Wooster R, Stratton MR. 2006. RNA editing of human microRNAs. *Genome Biol* **7:** R27. doi: 10.1186/gb-2006-7-4-r27.

Calabrese JM, Seila AC, Yeo GW, Sharp PA. 2007. RNA sequence analysis defines Dicer's role in mouse embryonic stem cells. *Proc Natl Acad Sci* **104:** 18097–18102.

Chen C-Z, Li L, Lodish HF, Bartel DP. 2004. MicroRNAs modulate hematopoietic lineage differentiation. *Science* **303:** 83–86.

Christodoulou F, Raible F, Tomer R, Simakov O, Trachana K, Klaus S, Snyman H, Hannon GJ, Bork P, Arendt D. 2010. Ancient animal microRNAs and the evolution of tissue identity. *Nature* **463:** 1084–1088.

Cummins JM, He YP, Leary RJ, Pagliarini R, Diaz LA, Sjoblom T, Barad O, Bentwich Z, Szafranska AE, Labourier E, et al. 2006. The colorectal microRNAome. *Proc Natl Acad Sci* **103:** 3687–3692.

Frazer KA, Eskin E, Kang HM, Bogue MA, Hinds DA, Beilharz EJ, Gupta RV, Montgomery J, Morenzoni MM, Nilsen GB, et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448:** 1050–1053.

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

Griffiths-Jones S. 2004. The microRNA registry. *Nucleic Acids Res* **32:** D109–D111. doi: 10.1093/nar/gkh023.

Grimm D, Streetz KL, Jopling CL, Storm TA, Pandey K, Davis CR, Marion P, Salazar F, Kay MA. 2006. Fatality in mice due to oversaturation of cellular microRNA/short hairpin RNA pathways. *Nature* **441:** 537–541.

Grimson A, Srivastava M, Fahey B, Woodcroft BJ, Chiang HR, King N, Degnan BM, Rokhsar DS, Bartel DP. 2008. Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature* **455:** 1193–1197.

Hagan JP, Piskounova E, Gregory RI. 2009. Lin28 recruits the TUTase Zcchc11 to inhibit let-7 maturation in mouse embryonic stem cells. *Nat Struct Mol Biol* **16:** 1021–1025.

Han JJ, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho YJ, Zhang BT, Kim VN. 2006. Molecular basis for the recognition of primary microRNAs by the Drosha–DGCR8 complex. *Cell* **125:** 887–901.

Han J, Pedersen JS, Kwon SC, Belair CD, Kim Y-K, Yeom K-H, Yang W-Y, Haussler D, Blelloch R, Kim VN. 2009. Post-transcriptional crossregulation between Drosha and DGCR8. *Cell* **136:** 75–84.

Heo I, Joo C, Cho J, Ha M, Han JJ, Kim VN. 2008. Lin28 mediates the terminal uridylation of let-7 precursor micro-RNA. *Mol Cell* **32:** 276–284.

Heo I, Joo C, Kim Y-K, Ha M, Yoon M-J, Cho J, Yeom K-H, Han J, Kim VN. 2009. TUT4 in concert with Lin28 suppresses microRNA biogenesis through pre-microRNA uridylation. *Cell* **138:** 696–708.

Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P. 1994. Fast folding and comparison of rna secondary structures. *Monatsh Chem* **125:** 167–188.

Houbaviy HB, Murray MF, Sharp PA. 2003. Embryonic stem cell-specific microRNAs. *Dev Cell* **5:** 351–358.

Hu H, Yan Z, Xu Y, Hu H, Menzel C, Zhou Y, Chen W, Khaitovich P. 2009. Sequence features associated with microRNA strand selection in humans and flies. *BMC Genomics* **10:** 413.

Kawahara Y, Zinshteyn B, Chendrimada TP, Shiekhattar R, Nishikura K. 2007a. RNA editing of the microRNA-151 precursor blocks cleavage by the Dicer–TRBP complex. *EMBO Rep* **8:** 763–769.

Kawahara Y, Zinshteyn B, Sethupathy P, Iizasa H, Hatzigeorgiou AG, Nishikura K. 2007b. Redirection of silencing targets by adenosine-to-inosine editing of miRNAs. *Science* **315:** 1137–1140.

Kawahara Y, Megraw M, Kreider E, Iizasa H, Valente L, Hatzigeorgiou AG, Nishikura K. 2008. Frequency and fate of microRNA editing in human brain. *Nucleic Acids Res* **36:** 5270–5280.

Kim Y-K, Kim VN. 2007. Processing of intronic microRNAs. *EMBO J* **26:** 775–783.

Kuchenbauer F, Morin RD, Argiropoulos B, Petriv OI, Griffith M, Heuser M, Yung E, Piper J, Delaney A, Prabhu AL, et al. 2008. In-depth characterization of the microRNA transcriptome in a leukemia progression model. *Genome Res* **18:** 1787–1797.

Kuhn RM, Karolchik D, Zweig AS, Wang T, Smith KE, Rosenbloom KR, Rhead B, Raney BJ, Pohl A, Pheasant M, et al. 2009. The UCSC Genome Browser Database: Update 2009. *Nucleic Acids Res* **37:** D755–D761. doi: 10.1093/nar/gkn875.

Kurihara Y, Watanabe Y. 2004. Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci* **101:** 12753–12758.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853–858.

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, Tuschl T. 2002. Identification of tissue-specific microRNAs from mouse. *Curr Biol* **12:** 735–739.

Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T. 2003. New microRNAs from mouse and human. *Rna* **9:** 175–179.

Landgraf P, Rusu M, Sheridan R, Sewer A, Iovino N, Aravin A, Pfeffer S, Rice A, Kamphorst AO, Landthaler M, et al. 2007. A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129:** 1401–1414.

Langmead B, Trapnell C, Pop M, Salzberg S. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol* **10:** R25. doi: 10.1186/gb-2009-10-3-r25.

Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* **294:** 858–862.

Lee RC, Ambros V. 2001. An extensive class of small RNAs in *Caenorhabditis elegans*. *Science* **294:** 862–864.

Lee Y, Ahn C, Han JJ, Choi H, Kim J, Yim J, Lee J, Provost P, Radmark O, Kim S, et al. 2003. The nuclear RNase III Drosha initiates microRNA processing. *Nature* **425:** 415–419.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP. 2003. Vertebrate microRNA genes. *Science* **299:** 1540.

Lu C, Tej SS, Luo S, Haudenschild CD, Meyers BC, Green PJ. 2005. Elucidation of the small RNA component of the transcriptome. *Science* **309:** 1567–1569.

Lund E, Guttinger S, Calado A, Dahlberg JE, Kutay U. 2004. Nuclear export of microRNA precursors. *Science* **303:** 95–98.

Marson A, Levine SS, Cole MF, Frampton GM, Brambrink T, Johnstone S, Guenther MG, Johnston WK, Wernig M, Newman J, et al. 2008. Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. *Cell* **134:** 521–533.

Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288:** 911–940.

Mineno J, Okamoto S, Ando T, Sato M, Chono H, Izu H, Takayama M, Asada K, Mirochnitchenko O, Inouye M, et al. 2006. The expression profile of microRNAs in mouse embryos. *Nucleic Acids Res* **34:** 1765–1771.

Okamura K, Hagen JW, Duan H, Tyler DM, Lai EC. 2007. The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila*. *Cell* **130:** 89–100.

Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): A curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **33:** D501–D504. doi: 10.1093/nar/gki025.

Rajagopalan R, Vaucheret H, Trejo J, Bartel DP. 2006. A diverse and evolutionarily fluid set of microRNAs in *Arabidopsis thaliana*. *Genes & Dev* **20:** 3407–3425.

Rao PK, Toyama Y, Chiang HR, Gupta S, Bauer M, Medvid R, Reinhardt F, Liao R, Krieger M, Jaenisch R, et al. 2009. Loss of cardiac microRNA-mediated regulation leads to dilated xardiomyopathy and heart failure. *Circ Res* **105:** 585–594.

Ro S, Park C, Young D, Sanders KM, Yan W. 2007. Tissue-dependent paired expression of miRNAs. *Nucleic Acids Res* **35:** 5944–5953.

Rodriguez A, Vigorito E, Clare S, Warren MV, Couttet P, Soond DR, van Dongen S, Grocock RJ, Das PP, Miska EA, et al. 2007. Requirement of bic/microRNA-155 for normal immune function. *Science* **316:** 608–611.

Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP. 2006. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127:** 1193–1207.

Ruby JG, Jan CH, Bartel DP. 2007a. Intronic microRNA precursors that bypass Drosha processing. *Nature* **448:** 83–86.

Ruby JG, Stark A, Johnston WK, Kellis M, Bartel DP, Lai EC. 2007b. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res* **17:** 1850–1864.

Seo TS, Bai XP, Ruparel H, Li ZM, Turro NJ, Ju JY. 2004. Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci* **101:** 5488–5493.

Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M. 2007. Systematic discovery and characterization of fly microRNAs using 12 *Drosophila* genomes. *Genome Res* **17:** 1865–1879.

Stark A, Bushati N, Jan CH, Kheradpour P, Hodges E, Brennecke J, Bartel DP, Cohen SM, Kellis M. 2008. A single Hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands. *Genes & Dev* **22:** 8–13.

Tam OH, Aravin AA, Stein P, Girard A, Murchison EP, Cheloufi S, Hodges E, Anger M, Sachidanandam R, Schultz RM, et al. 2008. Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature* **453:** 534–538.

Tyler DM, Okamura K, Chung W-J, Hagen JW, Berezikov E, Hannon GJ, Lai EC. 2008. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes & Dev* **22:** 26–36.

Voorhoeve PM, le Sage C, Schrier M, Gillis AJM, Stoop H, Nagel R, Liu Y-P, van Duijse J, Drost J, Griekspoor A, et al. 2006. A genetic screen implicates miRNA-372 and miRNA-373 as oncogenes in testicular germ cell tumors. *Cell* **124:** 1169–1181.

Watanabe T, Totoki Y, Toyoda A, Kaneda M, Kuramochi-Miyagawa S, Obata Y, Chiba H, Kohara Y, Kono T, Nakano T, et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* **453:** 539–543.

Wienholds E, Kloosterman WP, Miska E, Alvarez-Saavedra E, Berezikov E, de Bruijn E, Horvitz HR, Kauppinen S, Plasterk RHA. 2005. MicroRNA expression in zebrafish embryonic development. *Science* **309:** 310–311.

Wu H, Ye C, Ramirez D, Manjunath N. 2009. Alternative processing of primary microRNA transcripts by Drosha generates 5′ end variation of mature microRNA. *PLoS One* **4:** e7566. doi: 10.1371/journal.pone.0007566.

Xie XH, Lu J, Kulbokas EJ, Golub TR, Mootha V, Lindblad-Toh K, Lander ES, Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′ UTRs by comparison of several mammals. *Nature* **434:** 338–345.

Yang W, Chendrimada TP, Wang Q, Higuchi M, Seeburg PH, Shiekhattar R, Nishikura K. 2006. Modulation of microRNA processing and expression through RNA editing by ADAR deaminases. *Nat Struct Mol Biol* **13:** 13–21.

Yi R, Qin Y, Macara IG, Cullen BR. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes & Dev* **17:** 3011–3016.

Zeng Y, Yi R, Cullen BR. 2005. Recognition and cleavage of primary microRNA precursors by the nuclear processing enzyme Drosha. *EMBO J* **24:** 138–148.

Zuker M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res* **31:** 3406–3415.