Mammalian Gene Regulation through the 3′ UTR

by

Cydney Brooke Nielsen

B.Sc., University of British Columbia (2001)

Submitted to the Department of Biology
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY IN BIOLOGY

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

Author_____

Department of Biology
May 23, 2008

Certified by_____

Christopher B. Burge
Associate Professor
Departments of Biology and Biological Engineering
Thesis Supervisor

Accepted by_____

Stephen P. Bell
Professor of Biology
Chair, Committee for Graduate Students

Mammalian Gene Regulation through the 3′ UTR

Cydney Brooke Nielsen

Submitted to the Department of Biology
on May 23, 2008 in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy in Biology

Abstract

The untranslated region (UTR) at the 3′ end of a mammalian mRNA is typically rich with regulatory motifs that influence the stability, localization, translation and other properties of the message. We explored two classes of motifs, microRNA (miRNA) complementary sites and cleavage and polyadenylation (poly(A)) signals, and provide evidence that specific sequence contextual features are important for their recognition.

MiRNAs are ∼22 nt, non-coding RNAs that function as post-transcriptional gene regulators in animals and plants. They typically interact with target mRNAs through base-pairing predominantly between bases 2-8 (the 'seed' region) at the 5′ end of the miRNA and complementary sites in the target 3′ UTR ('seed matches'). These interactions result in target mRNA translational repression or deadenylation, or both. Through analysis of mRNA expression data following miRNA or siRNA overexpression or inhibition, we uncovered novel targeting determinants that influence mRNA levels. These include the presence of distinct seed match types and sequence context, in particular that increased AU content and conservation were independently associated with greater target down-regulation. Our results demonstrate that mRNA fold change increases multiplicatively (i.e., log-additively) with seed match count. We integrated these features into a target scoring scheme, TargetRank, and demonstrated the effectiveness of our rankings in predicting *in vivo* target responses.

Mammalian genes frequently have multiple, competing poly(A) sites, and the features influencing site selection remain poorly understood. Poly(A) site recognition occurs co-transcriptionally and given that transcription is highly influenced by the tight packaging of genomic DNA into chromatin, we investigated the potential impact of nucleosome positioning on poly(A) site usage. Using recent, public, Illumina sequencing data from human nucleosome boundaries, we found evidence that greater nucleosome density in regions flanking but not overlapping poly(A) sites is associated with more frequent usage.

Thesis Supervisor: Christopher B. Burge
Title: Associate Professor of Biology and Biological Engineering

Acknowledgements

I would like to thank Chris Burge for being an encouraging and dedicated advisor, whose attention to detail and scientific rigor have helped to shape my own research approach. I would also like to thank Rick Young and Phil Sharp who have served on my committee since the start, and have provided excellent advice and guidance over the years. I am grateful to my fellow Burge lab members and department colleagues for their support and willingness to discuss ideas. My family and close friends have been a constant source of encouragement, and I would like to give special thanks to my husband, Peter, whose belief in my abilities and remarkable perspective continue to inspire me.

Contents

Chapter 1

Introduction

Chapter 1

Introduction

Overview

This thesis focuses on the formation and regulation of mammalian 3′ UTRs. The processes studied here are diverse, specifically mammalian post-transcriptional regulation by microRNAs (miRNAs) and regulation of alternative cleavage and polyadenylation (poly(A)) at the 3′ ends of genes. However, in both cases, the approaches begin by examining known regulatory signals and expand on that knowledge by presenting evidence that specific sequence context features influence regulation. Chapter 2 describes an in-depth analysis of several novel mammalian miRNA and siRNA targeting determinants and demonstrates their effectiveness in predicting target mRNA responses. Chapter 3 presents our progress on a very recent project exploring connections between chromatin structure and poly(A) site usage, a direction inspired by my long-term interest in the regulation of alternative polyadenylation and the recent availability of Illumina sequencing data from human nucleosome boundaries. The current chapter introduces the broad topics of (i) chromatin structure and its role in transcriptional regulation, and (ii) roles of miRNAs in post-transcriptional gene regulation, providing context for the research described in Chapters 2 and 3. An earlier publication investigating the dynamics of intron gain and loss across fungal species is provided in Appendix 1. Appendices 2 and 3 contain supplemental material for

Chapters 2 and 3, respectively. A short concluding chapter discusses the significance of the results described in previous chapters and possible directions for future work.

Transcription in a chromatin context

Chromatin structure

Eukaryotic genomic DNA is packaged into chromatin and its core repeating unit, the nucleosome, is composed of 146 bp of DNA wrapped around a histone protein octamer (Kornberg and Lorch, 1999). The nucleosome core contains a central H3/H4 tetramer, flanked on both sides by H2A/H2B dimers, which are among the most conserved proteins known. Histones are mostly globular except for their N-terminal tails, which are less structured and subject to extensive modifications. While the structure of the nucleosome core is now well resolved (Richmond and Davey, 2003), the nature of higher order chromatin structure, such as the 30 nm fiber, remains under study (reviewed by Tremethick (2007)).

Nucleosome affinity for DNA

In order to wrap around the histone core, DNA must bend sharply, yet DNA sequences differ in their intrinsic abilities to accommodate such strain. Shortly after the discovery of the nucleosome, the tendency for some dinucleotides to occur with a periodicity of $\sim$10 bp was observed (close to the helical periodicity of DNA) and proposed to facilitate smooth DNA folding within chromatin (Trifonov and Sussman, 1980). With the availability of early nucleosome crystal structures (Richmond et al., 1984; Bentley et al., 1984), it was suggested that rotational positioning facilitates DNA bending, such that AT-rich dinucleotides are favorable at minor grooves facing

towards the histone core and GC-rich dinucleotides at minor grooves facing away from the core (Satchwell et al., 1986; Travers and Klug, 1987). In addition, long runs of dA:dT were found to occur preferentially in linker sequences presumably due to their relative rigidity (Drew and Travers, 1985). These same sequence properties have been observed in diverse eukaryotes from mammals to yeast (Widlund et al., 1997; Segal et al., 2006).

Several groups have devised prediction methods to locate nucleosomes based on these sequence properties. Early approaches made predictions using rotational preference matrices for dinucleotides (reviewed by Turnell and Travers (1992)) derived from a set of 177 aligned chicken core sequences (Satchwell et al., 1986). More recent genome-wide studies have employed techniques resembling position-specific scoring matrices to capture dinucleotide preferences from ∼200 nucleosome-associated sequences in yeast, *Saccharomyces cerevisiae* (Segal et al., 2006; Ioshikhes et al., 2006). Modest improvements were obtained by using discriminative models that attempt to differentiate between high and low affinity sequences (Peckham et al., 2007; Lee et al., 2007; Yuan and Liu, 2008). Through comparisons with independent data sets, such as those from tiling arrays (Yuan et al., 2005), Segal and coworkers (2006) demonstrated that 54% of *in vivo* nucleosome positions could be predicted by dinucleotide sequence features alone compared to 39% expected by chance. It is likely that further improvements will be made with higher coverage data sets for feature training. Nucleosome occupancy *in vivo* is influenced by a host of chromatin remodeling complexes, as discussed below, and these protein machines likely function to mobilize nucleosomes, allowing intrinsic nucleotide affinities and competition with other DNA-binding factors to guide positioning.

Chromatin as a transcriptional regulator

Early *in vitro* experiments revealed that nucleosomes can impede transcription (Knezetic and Luse, 1986), and it is now appreciated that regulation of chromatin structure has important consequences for transcriptional activity. Eukaryotic cells employ three distinct, but complementary, mechanisms to overcome the nucleosome barrier: histone modification, chromatin remodeling, and incorporation of histone variants (reviewed by Li et al. (2007); Saunders et al. (2006)). Chromatin plays a role in essentially all DNA-related metabolic processes, such as replication, recombination, and repair. Here, we will focus on its regulatory roles in the three phases of transcription: initiation, elongation, and termination.

Regulation of transcription initiation

The transcriptional cycle of a protein coding gene begins with the recruitment of the pre-initiation complex (PIC) to the core promoter region containing the transcriptional start site. This complex, composed of RNA polymerase II (Pol II) and general transcription factors (GTFs) TFIID, TFIIA, and TFIIB, is recruited in part by activator proteins bound upstream of the core promoter. TFIIH helicase activity is then required to open 12-15 bp of promoter DNA which serves as the single-stranded template for Pol II (reviewed by Lee and Young (2000)). There are several outstanding questions regarding how initial activator binding is impacted by the presence of nucleosomes. *In vitro* evidence suggests that some transcription factors can recognize their sequence targets on a nucleosome template (Taylor et al., 1991), however this ability is not universal and it has been proposed that spontaneous unwrapping of nucleosomes may provide initial access (Bucceri et al., 2006). Early high-throughput studies using chromatin immunoprecipitation followed by microarray analysis (ChIP-

chip) revealed reduced nucleosome occupancy in promoter regions (Bernstein et al., 2004; Lee et al., 2004a), and the prevalence of nucleosome-free regions (NFR) was corroborated by higher resolution methods (Sekinger et al., 2005; Pokholok et al., 2005; Yuan et al., 2005; Lee et al., 2007). Nucleosome affinity prediction points to inherent sequence properties of promoters that may explain their reduced occupancy (Segal et al., 2006; Ioshikhes et al., 2006). However, activators recruit extensive machinery which appears to help stabilize their binding, while further exposing the promoter DNA to create a state conducive to active transcription.

Histone modifying enzymes represent one such class of machinery recruited to promoters. For example, lysine residues are acetylated by a host of histone acetyltransferase complexes (HATs), including Gcn5 (component of the Spt, Ada, Gcn5 Acetyltransferase complex, SAGA) and Esa1 (component of NuA4 complex). These HATs are recruited to promoters through bound activators (Robert et al., 2004). Consistent with this, Pokholok et al. (2005) demonstrated peaks in H3 and H4 acetylation at active promoters and went on to show that the magnitude of these peaks correlated with transcription rate. One view is that by neutralizing positive charge on lysine residues, acetylation could result in the loosening of inter- and intra-nucleosome DNA-histone interactions. A second method used by cells to handle the nucleosome barrier is the recruitment of ATP-dependent nucleosome-remodeling complexes. Like HATs, chromatin-remodeling complexes are recruited to the promoter via their interactions with bound activators. Histone marks also play a role in recruitment. For example, acteylated nucleosomes at the promoter are recognized by the SWI/SNF chromatin-remodeling complex through its bromodomains (Hassan et al., 2002). Consistent with this idea, nucleosomes are observed to be hyperacetylated prior to being lost at active promoters (Reinke and Hörz, 2003). A third strategy is the use of

histone variants. Histone variants are distinct from the canonical core histones in that they are expressed outside of S phase and thus their incorporation into chromatin is replication independent. While the diverse roles of the H2A variant, H2A.Z, are still being uncovered, this variant appears well positioned on either side of the nucleosome-free region in promoters (Raisner et al., 2005; Barski et al., 2007). H2A.Z has been observed to flank the 5′ ends of both transcriptionally active and inactive genes (Raisner et al., 2005). Zhang et al. (2005) proposed that H2A.Z variants may serve to repress promoters while facilitating activation through their susceptibility to loss, which subsequently increases promoter DNA accessibility.

Negotiating the nucleosome during transcription elongation

Transcriptional elongation refers to the stages from promoter clearance through to assembly of a fully processive Pol II, resulting in the synthesis of a complete RNA transcript. Pol II frequently pauses at the promoter suggesting that the transition into productive elongation is a rate-limiting step (reviewed by Core and Lis (2008)). This promoter-proximal pausing was first observed at *Drosophila melanogaster* heat-shock genes and has since been demonstrated to be a widespread phenomenon (Gilmour and Lis, 1986; Rougvie and Lis, 1988; Rasmussen and Lis, 1993; Kim et al., 2005; Schones et al., 2008). Phosphorylation of the C-terminal domain (CTD) of the largest subunit of Pol II is critical for mediating the transition to elongation. Composed of tandem hepta-peptide repeats with the consensus YSPTSPS (52 copies in human, 26 in *S. cerevisiae*), the CTD is initially hypophosphorylated. Early in the transcriptional cycle, serine 5 (Ser5) is phosphorylated by the Cdk7 (Cyclin-dependent kinase-7) of THIIH. TFIIH-mediated Ser5 phosphorylation occurs on the PIC and recruits capping enzymes that stabilize the transcript 5′ end, through addition of a 7-methylguanosine.

It has been suggested that pausing may allow correct capping to occur, and that subsequent capping may facilitate escape from the pause (reviewed by Saunders et al. (2006)). Serine 2 (Ser2) phosphorylation appears later during elongation and is catalyzed by Cdk9 of P-TEFb (Positive Transcription Elongation Factor b). P-TEFb also phosphorylates DSIF (DRB Sensitivity-Inducing Factor) and NELF (Negative ELongation Factor) relieving their negative effects on elongation. The CTD serves therefore as a critical signaling platform during transcription.

As in transcription initiation, Pol II faces chromatin barriers throughout elongation, and while the themes for managing the DNA-histone interactions remain the same, the machinery employed is different. Transcriptionally active genes display characteristic patterns of histone modifications, including: (i) H3K4 methylation throughout the gene, with tri-, di-, and mono-methyl modifications dominating at the beginning, middle, and end of genes, respectively; (ii) $3'$ bias in H3K36 methylation; and (iii) H2B monoubiquitination throughout promoters and ORFs (reviewed by Li et al. (2007)). Unlike at promoters, recruitment of the modification machinery often occurs through the Pol II CTD. For example, the H3K4 methyltransferase, Set1 (COMPASS complex), is recruited to the Ser5-phosphorylated CTD with the help of elongation factor PAF (Krogan et al., 2003; Ng et al., 2003). It has been proposed that through its association with the CTD, Set1 gradually adds methyl groups to the $5'$ ends of genes, creating the gradient of tri- and di- methylations. *In vitro* studies suggest that H3K4 trimethylation has no direct effect on transcription (Pavri et al., 2006). However, it may play a signaling role as evidenced by its interactions with chromatin-remodeling factors, such as Chd1 (Pray-Grant et al., 2005; Flanagan et al., 2005). In addition to remodelers, histone chaperone proteins are important for nucleosome displacement and deposition enabling Pol II passage. For example, the FACT

heterodimer (FAcilitates Chromatin Transcription) mediates removal and reassembly of H2A-H2B histone dimers during elongation (Belotserkovskaya et al., 2003). Histone acetylation in front of the elongation machinery, later reversed by deacetylases upon reassembly, may play a role in histone removal (Workman, 2006). Pol II transcription is also associated with replacement of H3 histones with the variant H3.3 (Schwartz and Ahmad, 2005) and the chromatin remodeler Chd1 has been implicated in their incorporation (Konev et al., 2007). H3.3 histones, which have a shorter protein half-life in the cell, may help to destabilize active chromatin, promoting passage of Pol II.

Transcriptional termination and RNA $3'$ end formation

The final stage of the transcriptional cycle involves release of Pol II from the DNA template. Experiments conducted in the late '80s demonstrated that transcriptional termination is tightly coupled to transcript $3'$ end processing (Whitelaw and Proudfoot, 1986; Logan et al., 1987; Connelly and Manley, 1988). Specifically, the signals involved in cleavage and subsequent polyadenylation (addition of $\sim$200 adenosines) at the $3'$ end of nascent RNA transcripts were found to be essential for proper termination. While Pol II pause sites have been shown to promote termination (Yonaha and Proudfoot, 1999; Gromak et al., 2006), there appears to be no termination consensus sequence and Pol II release occurs stochastically downstream of the poly(A) site up to distances greater than 1 kb (Tran et al., 2001; Orozco et al., 2002).

Mammalian cleavage and polyadenylation involves the recognition of two core RNA motifs, the poly(A) signal (PAS) characterized by an AAUAAA hexamer or close variant, and a degenerate U-rich downstream element (DSE) (reviewed by

Zhao et al. (1999)). The PAS, located 10-30 nucleotides upstream of the poly(A) site, is recognized by the largest (160 kDa) of four subunits of CPSF (Cleavage and Polyadenylation Specificity Factor). The 64 kDa subunit of the trimeric CstF (Cleavage Stimulation Factor) binds to the U-rich DSE approximately 30 or fewer nucleotides downstream of the poly(A) site. Direct protein-protein interactions between CstF-77 and CPSF-160 result in mutual stabilization of the CPSF-CstF-RNA complex, and recent structural and biochemical evidence suggests that CstF may function as a dimer (Bai et al., 2007). The nuclease activity catalyzing the cleavage reaction has been attributed to CPSF-73 (Ryan et al., 2004; Mandel et al., 2006). Cleavage factors (CF) I_m and II_m, and poly(A) polymerase (PAP), are also required to form a cleavage-competent complex on the transcript. Following cleavage, CstF, CFI_m and CFII_m dissociate, leaving CPSF and PAP to complete the polyadenylation step, together with newly recruited poly(A)-binding protein II (PAB II), which is required for PAP to achieve its full processive activity.

The notion that 3′ end processing is coupled to transcription *in vivo* was initially supported by CTD deletion experiments using α-amanitin resistant forms of murine Pol II which lead to inhibition of polyadenylation and other RNA processing steps, such as capping and splicing (McCracken et al., 1997). Subsequent *in vitro* experiments provided evidence that Pol II, or a recombinant CTD, is required for cleavage at the poly(A) site in a transcription independent manor (Hirose and Manley, 1998). Interactions between the poly(A) machinery and Pol II happen early in the transcriptional process, as revealed by the unexpected purification of CPSF components in a transcription factor TFIID immunopurification and the demonstration that CPSF is transferred to Pol II upon transcription initiation at the promoter (Dantonel et al., 1997). These observations have been corroborated by ChIP experiments mapping

poly(A) machinery to locations across the length of genes (Calvo and Manley, 2005). Direct interactions have been detected between the CTD and *in vitro* translated CstF-50 (McCracken et al., 1997). In contrast, CPSF shows poor affinity for the CTD and its 30 kDa subunit appears to bind to the body of polymerase in a fashion that is mutually exclusive with CstF binding (Nag et al., 2007). While the mechanisms of poly(A) signal recognition are still being investigated, Nag et al. (2007) propose a model whereby CstF transiently associates with the CTD during elongation, but only becomes stably associated upon simultaneous binding of the U-rich DSE on the nascent transcript, a step they suggest is facilitated by Pol II pausing induced by the poly(A) signal itself.

Two predominant models of how cleavage and polyadenylation are coupled to termination have been proposed (Buratowski, 2005). One, coined the anti-terminator model, proposes that the switch from elongation to termination involves a change in Pol II associated factors that is induced upon encountering the poly(A) signals in the nascent RNA. A second, called the torpedo model, suggests that exonucleases target the unprotected RNA 5′ end resulting from cleavage, and process along the RNA until they eventually displace the polymerase. In support of the first model, Tran et al. (2001) used cis-antisense inhibition to demonstrate that recognition of the poly(A) signal at the RNA level is required for termination, and that transcriptional termination can occur in the absence of cleavage. Other studies used ChIP to reveal localization of certain elongation factors throughout transcribed regions, but not beyond the poly(A) site, consistent with factor rearrangement on the CTD (Ahn et al., 2004). The torpedo model is supported by observations of termination defects in yeast lacking the nuclease Rat1 (Kim et al., 2004) or in HeLa cells following RNAi knock-down of the Rat1 ortholog Xrn2 (West et al., 2004). As with many transcrip-

tional processes, both methods may be employed *in vivo* and their relative influences likely vary for different genes (Kim et al., 2006).

In contrast to the wealth of studies documenting a role for chromatin in transcriptional initiation and elongation, only a handful of papers address the potential impact of chromatin on termination. Alén et al. (2002) used transcription run-on analysis in *S. cerevisiae* to demonstrate that deletion of the conserved ATP-dependent chromatin remodeler, Chd1, leads to termination defects in a number of yeast genes. At other loci, Chd1 appeared to act redundantly with other chromatin-remodelers, Isw1 and Isw2, such that faulty termination was only observed in the triple mutant. Micrococcal nuclease (MNase) cleavage patterns across termination regions differed between wild type and *chd1* mutants, and ChIP experiments localized Chd1 to gene $3'$ ends. Intriguingly, both these results were reproducible independent of whether the gene was transcriptionally induced or not, suggesting that Chd1 remodeling is not a consequence of transcriptional termination, but rather configures the nucleosomes into a state conducive to termination. Alén et al. (2002) point to earlier studies showing that deletion of Chd1 reduces the cytotoxic effect of 6-azauracil, a drug that promotes Pol II pausing through depleting the nucleotide pool (Woodage et al., 1997). The authors point out that this is consistent with a role for Chd1 in enhancing Pol II transcriptional pausing, a process known to facilitate the switch from elongation to termination. In a subsequent study, Morillon et al. (2003) demonstrated that deletion of Isw1 partially overcomes sensitivity to 6-azauracil resulting from loss of certain positive elongation factors, suggesting that Isw1 functions to block elongation. As with Chd1, deletion of Isw1 alone leads to defects in termination at certain loci. The authors rule out a reduction in poly(A) site recognition as Northern analysis indicated similar transcript levels in both wild type and mutant yeast. While there is

extensive evidence that chromatin plays a role in all phases of transcription, whether nucleosomes also play important roles in RNA processing events such as cleavage and polyadenylation remains unclear.

Post-transcriptional regulation by microRNAs

The small RNA revolution

The first microRNA (miRNA) was discovered when the isolated locus of *lin-4*, a known regulator of cell linage in *Caenorhabditis elegans* larval development, was demonstrated to encode not a protein-coding mRNA but rather produce a 22 nt non-coding RNA (Lee et al., 1993; Wightman et al., 1993). Suggestions that this was a curiosity of *C. elegans* biology were countered with evidence of another regulatory 21 nt RNA, *let-7* (Reinhart et al., 2000), which proved to be conserved in diverse eukaryotes including human and the fruit fly *D. melanogaster* (Pasquinelli et al., 2000). At about the same time, injection of exogenous double-stranded RNA (dsRNA) into *C. elegans* was reported to induce specific gene silencing of the corresponding endogenous locus, a process termed RNA-interference (RNAi) (Fire et al., 1998). Short 25 nt RNAs were implicated in a similar phenomenon in plants called posttranscriptional gene silencing (PTGS) (Hamilton and Baulcombe, 1999), and demonstrated to function as intermediates in the RNAi pathway in a *Drosophila in vitro* system (Zamore et al., 2000). These discoveries sparked multiple small RNA cloning efforts in both invertebrate and vertebrate model systems, revealing the diversity and cross-species conservation of small RNAs (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Characterization of small RNAs remains ongoing. They were recently found in a unicellular green alga *Chlamydomonas reinhardtii* (Molnár

et al., 2007; Zhao et al., 2007a) and in several eukaryote-infecting viruses (reviewed by Scaria et al. (2007)), further emphasizing their key roles in eukaryotic gene regulation.

Biogenesis

MiRNAs are processed from several kilobase-long transcripts (pri-miRNAs) (Lee et al., 2002), which are capped and polyadenylated pol II products (Lee et al., 2004b; Cai et al., 2004). An exception to this is a cluster of human miRNAs interspersed among repetitive Alu elements which are Pol III transcribed (Borchert et al., 2006). While mammalian pri-miRNAs can be transcribed from distinct loci using their own promoters, over half are found to overlap spliced transcriptional units, often occurring within introns of protein coding genes (Rodriguez et al., 2004). Clusters of miRNA genes resulting in poly-cistronic transcripts are common (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee et al., 2002). The nuclear RNase III Drosha (Lee et al., 2003), together with its interacting partner DGCR8 (Gregory et al., 2004; Han et al., 2004; Landthaler et al., 2004), processes pri-miRNAs into ∼60-70 nt pre-miRNAs. The pre-miRNA sequences form stem-loop structures within the pri-miRNA, and recognition of their ∼33 bp stem and flanking single RNA strands by DGCR8 facilitates Drosha positioning and cleavage ∼11 bp into the stem (Han et al., 2006). The resulting hairpin pre-miRNAs are subsequently exported into the cytoplasm by nuclear transport factor, Exportin-5/Ran-GTP (Yi et al., 2003). Recently, Ruby et al. (2007b) uncovered an alternative processing pathway in which certain debranched introns in *C. elegans* and *D. melanogaster* mimic the structures of pre-miRNAs, and these mirtrons can enter the silencing pathway without Drosha-mediated cleavage.

Unlike miRNAs, small interfering RNAs (siRNAs) involved in RNAi are processed

from long dsRNA molecules from either exogenous or endogenous sources. The RNase III protein Dicer was found to cut these long dsRNA substrates into ∼22 siRNAs (Bernstein et al., 2001), and shortly thereafter, was demonstrated to also function in cleavage of pre-miRNA hairpins, producing mature ∼22 nt miRNAs (Grishok et al., 2001; Hutvágner et al., 2001; Ketting et al., 2001). While the single mammalian Dicer possesses both these activities, recent evidence suggests that two different Dicer genes in *Drosophila*, Dicer-1 and Dicer-2, are responsible for pre-miRNA cleavage and siRNA generation, respectively, with only rare exceptions (Lee et al., 2004c). Cleavage by Dicer (and Drosha in the case of miRNAs) results in duplexes with 5′ phosphates and 3′ 2 nt overhangs characteristic of RNase III endonucleases.

Both miRNAs and siRNAs are loaded into a cytoplasmic multi-component complex, RISC (RNA-Induced Silencing Complex) which contains a member of the Argonaute (Ago) protein family as a core component (Hammond et al., 2001; Hutvágner and Zamore, 2002; Martinez et al., 2002). The well conserved Ago/Piwi proteins, which can be clustered into the Ago subfamily and the Piwi subfamily, contain PAZ and PIWI domains and as a result, are sometimes called PPD proteins (reviewed by Farazi et al. (2008)). The PAZ domain recognizes the 3′ end of small RNAs and is also found in most Dicer RNase III family members, where recognition of one end of the duplex occurs at a fixed distance from the endonuclease domain (Zhang et al., 2004; MacRae et al., 2006, 2007). Structural studies of PIWI domains revealed its role in recognition of the RNA 5′ phosphate, which appeared unpaired from the rest of the RNA duplex (Ma et al., 2005; Parker et al., 2005). The PIWI domain forms an RNase H fold which can function as an endonuclease. PIWI domains show sequence variation in the active site, and of the four mammalian Argonaute proteins, only Ago2 possesses this RNA cleavage, or 'slicer', activity (Liu et al., 2004). While purified hu-

man Ago2, combined with an siRNA, can form a minimal RISC (Rivas et al., 2005), Dicer and the dsRNA binding protein TRBP associate with Ago2 and are likely important for proper miRNA loading (Chendrimada et al., 2005). Martinez et al. (2002) observed RISC to contain single-strand RNA, and the choice of strand appears to be guided in part by the relative stability of the terminal base pairs, with the surviving guide strand having weaker pairing to its complement at its 5′ end (Schwarz et al., 2003; Khvorova et al., 2003). Ago2 'slicer' activity was reported to play a role in siRNA biogenesis, as Ago2 was shown to cleave the passenger (non-guide) strand of the bound siRNA duplex, leading to its degradation (Matranga et al., 2005; Rand et al., 2005; Miyoshi et al., 2005). Recently, Ago2 has been implicated in miRNA processing as Diederichs and Haber (2007) uncovered an Ago2-cleaved precursor (ac-pre-miRNA) in human cells consisting of a pre-miRNA hairpin with a cut ∼11 nt from the terminal end of the 3′-arm. They demonstrated that this nick is dependent on Ago2 RNase activity and that the cleaved pre-miRNA is a substrate for Dicer, suggesting that Ago2 plays a role in passenger strand removal.

miRNA genes

Initial miRNA cloning experiments in *C. elegans* exploited characteristic miRNA features, such as the 5′ phosphate and 3′ hydroxyl groups, and uncovered ∼50 miRNAs genes (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Computational methods were devised for miRNA gene prediction that searched for conserved stem loop precursor RNAs mimicking real pre-miRNAs (Lim et al., 2003; Grad et al., 2003). These predictions suggested a miRNA gene count in the hundreds, with ∼1% of genes encoding miRNAs (Bartel, 2004). Today, miRBase contains thousands of miRNA sequences from over 50 organisms (Griffiths-Jones et al., 2008), and current

deep-sequencing approaches continue to reveal additional examples, many of which are expressed at low levels and appear to be species specific (Berezikov et al., 2006; Ruby et al., 2006; Rajagopalan et al., 2006; Morin et al., 2008). These techniques have also been instrumental in uncovering siRNAs, and piRNAs (Piwi-Interacting RNAs) thought to be important for transposon silencing, which are reviewed elsewhere (Chu and Rana, 2007).

Modes of action

Partial complementarity between the *lin-4* miRNA and the 3′ UTR of a negatively regulated downstream mRNA, *lin-14* (Lee et al., 1993; Wightman et al., 1993) pointed to an antisense targeting mechanism at the 3′ end of the transcript. The importance of the miRNA 5′ end, in particular bases 2-8 termed the 'seed' region and the corresponding W-C complementary target sequence called the 'seed match', was initially revealed through comparative genomics. Lewis and coworkers (2003) used a sliding heptamer window along human miRNA sequences and examined the conservation of the corresponding complementary sequences in 3′ UTRs. Compared to control heptamers matching shuffled miRNA sequences, seed matches to miRNA bases 2-8 showed the greatest above-background conservation. This work was complemented by reports implicating the 5′ miRNA end in mediating target regulation (Lai, 2002) and later careful reporter studies (Doench and Sharp, 2004; Brennecke et al., 2005). Decreases in *lin-14* protein levels but stable mRNA levels indicated that *lin-4* regulates target translation. In contrast, siRNAs can direct Ago2-dependent cleavage of their targets through perfect base-pairing along their ∼22 nt length, with cleavage localized to the center of the siRNA spanning region (Elbashir et al., 2001; Liu et al., 2004). This targeting mechanism is reminiscent of that used by plant miRNAs (Jones-

Rhoades et al., 2006). It is now clear that these different modes of action depend on the degree of complementarity between the small RNA and its target, and that the miRNAs and siRNAs are themselves indistinguishable, defined purely by their source and biogenesis, endogenously transcribed hairpin structures or long dsRNA molecules respectively (Hutvágner and Zamore, 2002; Doench et al., 2003; Zeng et al., 2003). The initial paradigm of imperfect base-pairing leading to translational repression without alteration of target mRNA abundance has since been challenged by reports of target mRNA reductions in response to miRNAs, including in the original *lin-4/lin-14* pairing (Lim et al., 2005; Bagga et al., 2005; Wu and Belasco, 2005). While the mechanisms of translational repression and mRNA destabilization are still being uncovered, it is clear that both play important roles in regulation by small RNAs.

How miRNAs repress translation is not completely understood. Evidence that miRNAs inhibit translation initiation came from studies in mammalian cells transfected with *in vitro* transcribed reporter mRNAs. These studies demonstrated that m^7G-capped mRNAs, but not mRNAs containing an internal ribosome entry site (IRES), were repressed by miRNAs (Pillai et al., 2005; Humphreys et al., 2005). Kiriakidou and coworkers (2007) reported protein sequence similarity between human Ago proteins and the cap-binding region of the eukaryotic translation initiation factor, eIF4E, and provided data from Ago2 mutations to support a model whereby Ago proteins compete with eIF4E for cap binding. However, the recent work of Eulalio and coworkers (2008) demonstrated that the equivalent mutations in *D. melanogaster* Ago1 abolished silencing activity without effecting m^7GTP-Sepharose binding, suggesting that the role of these residues is unrelated to cap binding. They further showed that these mutations disrupt interactions between Ago1 and both GW182 and miRNAs, presenting evidence that Ago1-GW182 binding is important for miRNA

silencing. Chendrimada and coworkers (2007) have recently provided evidence that eIF6, a protein known to prevent productive assembly of the 80S ribosome, immunoprecipitates with the Ago2-Dicer-TRBP complex and may play a role.

There is also evidence that inhibition occurs post-initiation. Using sedimentation profiles, several groups have observed target mRNAs to be associated with polysomes (Olsen and Ambros, 1999; Seggerson et al., 2002) which appear to be engaged in translation elongation due to their sensitivity to puromycin (Petersen et al., 2006; Maroney et al., 2006). There are also data demonstrating that miRNAs can indeed repress translation directed by IRES elements, challenging the importance of cap recognition in miRNA regulation (Petersen et al., 2006). There is growing agreement that these two mechanisms are not mutually exclusive, but rather the observed discrepancies may result from differences in experimental systems and procedures that favor detection of one mechanism over another (reviewed by Wu and Belasco (2008); Filipowicz et al. (2008)).

In contrast to translation initiation, there is greater consensus regarding how miRNA-dependent mRNA degradation occurs. miRNAs can direct cleavage of targets with full-length complementarity, however with the exception of miR-196 regulation of HOXB8 (Yekta et al., 2004), this does not appear to be the norm in animals. The vastly more common imperfect base-pairing between a miRNA and its target was observed by Wu et al. (2006) and Giraldez et al. (2006) to result in removal of the $3'$ poly(A) tail and acceleration of target mRNA decay. In these studies, reporter deadenylation was detected in the absence of translation, blocked either through introduction of a $5'$ UTR stem loop structure preventing 80S ribosome assembly or by use of an antisense morpholino to mask the translational start site. Transcripts destined

for decay are known to localize to cytoplasmic processing bodies (P or GW bodies) which contain an abundance of decapping enzymes and exonucleases (Sheth and Parker, 2003). Deadenylation in response to miRNAs is dependent on the deadenylase CCR4:NOT, and decapping complexes DCP1:DCP2 in *Drosophila* (Behm-Ansmant et al., 2006). Ago proteins localize to P-bodies, as do reporter miRNA targets in a miRNA dependent manner (Sen and Blau, 2005; Liu et al., 2005b). A P-body marker protein, GW128, co-purifies with Ago proteins in mammalian cells and was found to interact with the PIWI domain of Ago1 in *Drosophila* (Behm-Ansmant et al., 2006). Depletion of GW182 leads to P-body disruption and also impairs silencing by miRNAs, and to a lesser extent by siRNAs (Liu et al., 2005a; Jakymiw et al., 2005). However, observations that miRNA silencing can occur when P-bodies are disrupted by other means points to a more direct role for GW128 in the miRNA pathway (Chu and Rana, 2006; Eulalio et al., 2008). In addition to being sites of mRNA decay, P-bodies are devoid of ribosomes (Teixeira et al., 2005). Given that mRNA entry into P-bodies seems to require inhibition of translation, a plausible model is that non-translating miRNA targets are directed to P-bodies, where their sequestration and/or decay reinforces the initial silencing (reviewed by Eulalio et al. (2007)).

Functions *in vivo*

In recent years, miRNAs have been implicated in diverse biological processes including apoptosis, cell division, and metabolism (reviewed by Bushati and Cohen (2007)). Deletion of miRNA processing enzymes in model organisms has demonstrated their importance in development. In particular, loss of Dicer (of both maternal and zygotic origin), leads to embryonic lethality in *C. elegans*, zebrafish, and mouse although the stage of impairment varies (Grishok et al., 2001; Bernstein et al., 2003; Giraldez et al.,

2005). Deletion of individual miRNA genes led to severe phenotypes in mice, such as immunodeficiency in the case of miR-155 (Rodriguez et al., 2007) or defects in cardiogenesis resulting from loss of a single genomic copy of miR-1 (Zhao et al., 2007b). Systematic deletion of miRNA genes in *C. elegans* revealed that in most cases disruption of a miRNA gene does not result in grossly abnormal phenotypes, suggesting functional redundancy among miRNAs in nematodes (Miska et al., 2007). Studies that have examined the consequences of both miRNA overexpression and depletion in specific cell types have offered insights into the functions of individual miRNAs. An example of this approach comes from Chen et al. (2006) where mis-expression of miR-1 leads to accelerated C2C12 myoblast differentiation, and depletion of miR-1 by 2′-O-methyl antisense oligonucleotides impedes differentiation, demonstrating its regulatory role in muscle development. These findings are corroborated by loss-of-function studies in mice lacking a copy of miR-1 that displayed defects in cardiogenesis (Zhao et al., 2007b).

The first discovered miRNA, *lin-4*, appears to function as a developmental switch in *C. elegans*, regulating the decision to transition from the first larval stage to the second (Lee et al., 1993; Wightman et al., 1993). This miRNA exerts its effect through a key mRNA target, *lin-14*, which contains multiple *lin-4* seed matches in its 3′ UTR. Given that this pair was identified through forward genetics, it is not surprising that this first example has a dramatic regulatory effect. More recent genome-wide analyses, such as those conducted by Lim et al. (2005) involving transfection of miR-1 or miR-124 RNA duplexes into HeLa cells and subsequent microarray analysis, revealed that miRNAs can regulate hundreds of genes frequently containing only single seed matches to the corresponding miRNA. These results suggest that miRNAs may serve more general functions in altering global expression patterns. Strikingly, they

observed that transcripts down-regulated by miRNA overexpression were biased for mRNAs normally expressed at relatively low levels in the tissue where the endogenous miRNA is expressed. For example, mRNAs responsive to the brain-specific miR-124 also tended to be poorly expressed in brain. Such inverse relations between tissue-specific miRNAs and their targets have been observed in other systems and lead to the idea that miRNAs may reinforce regulation at the transcriptional level, providing robustness to a cell's expression profile (Stark et al., 2005; Farh et al., 2005). Experiments by Giraldez and coworkers (2006) offered a related example whereby miR-430 functions to clear maternal mRNAs during zebrafish embryogenesis, essentially accelerating the shift in expression profile to create a more precise developmental transition.

While miRNAs are clearly regulated at the expression level, displaying restricted expression patterns across cell types or developmental stages, a number of recent studies demonstrated that RNA-binding proteins can modulate miRNA effects on individual target mRNAs. In the zebrafish system mentioned above, Mishima et al. (2006) observed that some miR-430 targets, including *nanos1* which is essential for proper germ cell formation, are expressed in primordial germ cells despite the accumulation of miR-430. They also identified a region of the *nanos1* 3′ UTR that conferred this apparent protection from miRNA regulation. Kedde et al. (2007) recently provided evidence that Dnd1 (dead end 1), an RNA-binding protein required for germ cell survival in zebrafish, binds to U-rich regions near the miR-430 seed matches in the *nanos1* 3′ UTR prohibiting miR-430 binding. They also observed this type of Dnd1 regulation in other miRNA-mRNA pairs in human cells, and traced disruption of the miRNA-mRNA interaction using a 3′-biotin labeled miRNA and streptavidin bead pull-down experiments in the presence and absence of Dnd1. This is reminiscent

of work by Bhattacharyya et al. (2006) showing that HuR, an AU-rich-element binding protein, can relieve miR-122 directed repression of the CAT-1 mRNA in human cells under stress conditions. Ago proteins and miRNAs have been shown to localize to stress granules (SGs) upon exposure to stress stimuli where their activity may be modulated by other SG-associated RNA-binding proteins such as TIA-1 (Leung et al., 2006). Understanding how miRNAs are modulated by other regulators will assist in deciphering their functions *in vivo*.

Target prediction

A critical step in interpreting cellular responses to miRNAs is identifying target genes. The early examples of miRNA-mRNA pairs identified through forward genetics demonstrated preferential base-pairing between the 5′ end of the miRNA and conserved complementary sequences in the target mRNA 3′ UTR (Lee et al., 1993; Wightman et al., 1993; Reinhart et al., 2000). Initial target prediction algorithms built on these observations by searching sets of 3′ UTRs for conserved 'seed matches' (W-C complement of miRNA 'seed' consisting of bases 2-8) (Lewis et al., 2003) or a 'binding nucleus' (6 to 8 bp W-C complement of the miRNA not necessarily at its 5′ end) (Rajewsky and Socci, 2004), with some methods tolerating G:U pairs in the seed region (Stark et al., 2003; Enright et al., 2003). These early methods incorporated estimates of the thermodynamic stability of the miRNA-mRNA interaction in their predictions. False positive rates based on shuffled miRNA controls were estimated as 31% for human targets conserved to mouse and rat (Lewis et al., 2003) and 35% for *D. melanogaster* targets conserved to *D. pseudoobscura* (Enright et al., 2003), leading to the prediction of hundreds of conserved miRNA targets in total (i.e. a few targets per miRNA gene).

More detailed characterizations of the miRNA-mRNA interaction led to improved prediction methods. Lewis and coworkers (2005) demonstrated preferential conservation of adenosine at the target position opposite the first miRNA base for miRNAs beginning with non-U as well as U bases. This led to the hypothesis of direct recognition of this base by a protein component of the silencing complex rather than through base-pairing. Careful reporter analyses in fly revealed the regulatory effects of seed matches in the absence of 3′ pairing, suggesting that target rankings based on overall base-pair maximization across the miRNA may not be biologically meaningful (Brennecke et al., 2005). In addition, the presence of G:U base-pairs within the seed region were found to be more detrimental to effective repression than expected based on standard thermodynamic models (Doench and Sharp, 2004; Brennecke et al., 2005). These considerations, together with the availability of additional whole genome sequences, led to significant increases in predictions with hundreds of conserved target mRNAs reported per miRNA, representing ∼30% of human genes. In addition, it became clear that many 3′ UTRs may be targeted by multiple miRNAs, and some target prediction algorithms scored targets in terms of a set of miRNAs (Krek et al., 2005). However, the requirement for perfect conservation of seed matches in aligned genomic regions can lead to reduced sensitivity due to variable assembly quality or coverage. Ruby et al. (2007a) have recently assessed seed match conservation by a branch length score, a measure of evolutionary distance across which a motif is conserved. In addition, the importance of additional targeting determinants beyond the seed match have been reported (Grimson et al., 2007; Nielsen et al., 2007), and will be discussed in detail in Chapter 2.

Current target predictions are based on mRNA 3′ UTRs. There are examples of miRNAs targeting the coding region, as is the case for DNMT3b1 regulation by

miR-148 in human cells (Duursma et al., 2008), and evidence that seed matches placed in the 5′ UTR of reporter constructs carrying internal ribosome entry sites can drive repression (Lytle et al., 2007). Computational methods have found evidence for conserved coding-region seed matches above background expectation (Lewis et al., 2005; Grimson et al., 2007; Stark et al., 2007). While exonic sites may play a role in individual genes, these studies illustrated that the vast majority of targeting interactions occur through the 3′ UTR. Unfortunately, the quality of 3′ UTR annotation remains a limitation for target prediction, particularly in organisms such as *C. elegans* where fixed length windows downstream of stop codons are typically used (Lall et al., 2006). Legendre and coworkers (2006) provided initial evidence that miRNA targeted isoforms are present at reduced levels in tissues expressing the corresponding miRNA. It is likely that the switch to a non-target, alternative isoform and expression of the corresponding miRNA are coordinated events, as appears to be the case with *Tropomyosin 1* in flies where the non-muscle isoform contains target sites to the muscle-specific miRNA, miR-1 (Stark et al., 2007). However, the impact of alternative 3′ UTRs on miRNA targeting genome-wide remains to be characterized.

References

S. H. Ahn, M. Kim, and S. Buratowski. Phosphorylation of serine 2 within the RNA polymerase II C-terminal domain couples transcription and 3′ end processing. *Mol Cell*, 13(1):67–76, Jan 2004.

C. Alén, N. A. Kent, H. S. Jones, J. O'Sullivan, A. Aranda, and N. J. Proudfoot. A role for chromatin remodeling in transcriptional termination by RNA polymerase II. *Mol Cell*, 10(6):1441–52, Dec 2002.

S. Bagga, J. Bracht, S. Hunter, K. Massirer, J. Holtz, R. Eachus, and A. E. Pasquinelli. Regulation by let-7 and lin-4 mirnas results in target mRNA degradation. *Cell*, 122(4):553–63, Aug 2005.

Y. Bai, T. C. Auperin, C.-Y. Chou, G.-G. Chang, J. L. Manley, and L. Tong. Crystal structure of murine Cstf-77: dimeric association and implications for polyadenylation of mRNA precursors. *Mol Cell*, 25(6):863–75, Mar 2007.

A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007.

D. P. Bartel. Microrna: genomics, biogenesis, mechanism, and function. *Cell*, 116 (2):281–97, Jan 2004.

I. Behm-Ansmant, J. Rehwinkel, T. Doerks, A. Stark, P. Bork, and E. Izaurralde. mRNA degradation by miRNAs and GW182 requires both CCR4:NOT deadenylase and DCP1:DCP2 decapping complexes. *Genes Dev*, 20(14):1885–98, Jul 2006.

R. Belotserkovskaya, S. Oh, V. A. Bondarenko, G. Orphanides, V. M. Studitsky, and D. Reinberg. FACT facilitates transcription-dependent nucleosome alteration. *Science*, 301(5636):1090–3, Aug 2003.

G. A. Bentley, A. Lewit-Bentley, J. T. Finch, A. D. Podjarny, and M. Roth. Crystal structure of the nucleosome core particle at 16°A resolution. *J Mol Biol*, 176(1): 55–75, Jun 1984.

E. Berezikov, F. Thuemmler, L. W. van Laake, I. Kondova, R. Bontrop, E. Cuppen, and R. H. A. Plasterk. Diversity of microRNAs in human and chimpanzee brain. *Nat Genet*, 38(12):1375–7, Dec 2006.

B. E. Bernstein, C. L. Liu, E. L. Humphrey, E. O. Perlstein, and S. L. Schreiber. Global nucleosome occupancy in yeast. *Genome Biol*, 5(9):R62, Jan 2004.

E. Bernstein, A. A. Caudy, S. M. Hammond, and G. J. Hannon. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature*, 409(6818):363–6, Jan 2001.

E. Bernstein, S. Y. Kim, M. A. Carmell, E. P. Murchison, H. Alcorn, M. Z. Li, A. A. Mills, S. J. Elledge, K. V. Anderson, and G. J. Hannon. Dicer is essential for mouse development. *Nat Genet*, 35(3):215–7, Nov 2003.

S. N. Bhattacharyya, R. Habermacher, U. Martine, E. I. Closs, and W. Filipowicz. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–24, Jun 2006.

G. M. Borchert, W. Lanier, and B. L. Davidson. RNA polymerase III transcribes human microRNAs. *Nat Struct Mol Biol*, 13(12):1097–101, Dec 2006.

J. Brennecke, A. Stark, R. B. Russell, and S. M. Cohen. Principles of microRNA-target recognition. *PLoS Biol*, 3(3):e85, Mar 2005.

A. Bucceri, K. Kapitza, and F. Thoma. Rapid accessibility of nucleosomal DNA in yeast on a second time scale. *EMBO J*, 25(13):3123–32, Jul 2006.

S. Buratowski. Connections between mRNA 3′ end processing and transcription termination. *Current Opinion in Cell Biology*, 17:257–261, 2005.

N. Bushati and S. M. Cohen. microRNA functions. *Annu Rev Cell Dev Biol*, 23: 175–205, Jan 2007.

X. Cai, C. H. Hagedorn, and B. R. Cullen. Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *RNA*, 10 (12):1957–66, Dec 2004.

O. Calvo and J. L. Manley. The transcriptional coactivator PC4/Sub1 has multiple functions in RNA polymerase II transcription. *EMBO J*, 24(5):1009–20, Mar 2005.

J.-F. Chen, E. M. Mandel, J. M. Thomson, Q. Wu, T. E. Callis, S. M. Hammond, F. L. Conlon, and D.-Z. Wang. The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat Genet*, 38(2):228–33, Feb 2006.

T. P. Chendrimada, R. I. Gregory, E. Kumaraswamy, J. Norman, N. Cooch, K. Nishikura, and R. Shiekhattar. TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature*, 436(7051):740–4, Aug 2005.

T. P. Chendrimada, K. J. Finn, X. Ji, D. Baillat, R. I. Gregory, S. A. Liebhaber, A. E. Pasquinelli, and R. Shiekhattar. MicroRNA silencing through RISC recruitment of eIF6. *Nature*, 447(7146):823–8, Jun 2007.

C. Y. Chu and T. M. Rana. Translation repression in human cells by microRNA-induced gene silencing requires RCK/p54. *PLoS Biol*, 4(7):e210, Jul 2006.

C. Y. Chu and T. M. Rana. Small RNAs: regulators and guardians of the genome. *J Cell Physiol*, 213(2):412–9, Nov 2007.

S. Connelly and J. L. Manley. A functional mRNA polyadenylation signal is required for transcription termination by RNA polymerase II. *Genes Dev*, 2(4):440–52, Apr 1988.

L. J. Core and J. T. Lis. Transcription regulation through promoter-proximal pausing of RNA polymerase II. *Science*, 319(5871):1791–2, Mar 2008.

J. C. Dantonel, K. G. Murthy, J. L. Manley, and L. Tora. Transcription factor TFIID recruits factor CPSF for formation of 3′ end of mRNA. *Nature*, 389(6649):399–402, Sep 1997.

S. Diederichs and D. A. Haber. Dual role for argonautes in microRNA processing and posttranscriptional regulation of microRNA expression. *Cell*, 131(6):1097–108, Dec 2007.

J. G. Doench and P. A. Sharp. Specificity of microRNA target selection in translational repression. *Genes Dev*, 18(5):504–11, Mar 2004.

J. G. Doench, C. P. Petersen, and P. A. Sharp. siRNAs can function as miRNAs. *Genes Dev*, 17(4):438–42, Feb 2003.

H. R. Drew and A. A. Travers. DNA bending and its relation to nucleosome positioning. *J Mol Biol*, 186(4):773–90, Dec 1985.

A. Duursma, M. Kedde, M. Schrier, C. le Sage, and R. Agami. miR-148 targets human DNMT3b protein coding region. *RNA*, Mar 2008.

S. M. Elbashir, W. Lendeckel, and T. Tuschl. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev*, 15(2):188–200, Jan 2001.

A. J. Enright, B. John, U. Gaul, T. Tuschl, C. Sander, and D. S. Marks. MicroRNA targets in Drosophila. *Genome Biol*, 5(1):R1, Jan 2003.

A. Eulalio, I. Behm-Ansmant, and E. Izaurralde. P bodies: at the crossroads of post-transcriptional pathways. *Nat Rev Mol Cell Biol*, 8(1):9–22, Jan 2007.

A. Eulalio, E. Huntzinger, and E. Izaurralde. GW182 interaction with Argonaute is essential for miRNA-mediated translational repression and mRNA decay. *Nat Struct Mol Biol*, 15(4):346–53, Apr 2008.

T. A. Farazi, S. A. Juranek, and T. Tuschl. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, 135 (7):1201–14, Apr 2008.

K. K.-H. Farh, A. Grimson, C. Jan, B. P. Lewis, W. K. Johnston, L. P. Lim, C. B. Burge, and D. P. Bartel. The widespread impact of mammalian microRNAs on mRNA repression and evolution. *Science*, 310(5755):1817–21, Dec 2005.

W. Filipowicz, S. N. Bhattacharyya, and N. Sonenberg. Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*, 9(2):102–14, Feb 2008.

A. Fire, S. Xu, M. K. Montgomery, S. A. Kostas, S. E. Driver, and C. C. Mello. Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature*, 391(6669):806–11, Feb 1998.

J. F. Flanagan, L.-Z. Mi, M. Chruszcz, M. Cymborowski, K. L. Clines, Y. Kim, W. Minor, F. Rastinejad, and S. Khorasanizadeh. Double chromodomains cooperate to recognize the methylated histone H3 tail. *Nature*, 438(7071):1181–5, Dec 2005.

D. S. Gilmour and J. T. Lis. RNA polymerase II interacts with the promoter region of the noninduced hsp70 gene in Drosophila melanogaster cells. *Mol Cell Biol*, 6 (11):3984–9, Nov 1986.

A. J. Giraldez, R. M. Cinalli, M. E. Glasner, A. J. Enright, J. M. Thomson, S. Baskerville, S. M. Hammond, D. P. Bartel, and A. F. Schier. MicroRNAs regulate brain morphogenesis in zebrafish. *Science*, 308(5723):833–8, May 2005.

A. J. Giraldez, Y. Mishima, J. Rihel, R. J. Grocock, S. van Dongen, K. Inoue, A. J. Enright, and A. F. Schier. Zebrafish miR-430 promotes deadenylation and clearance of maternal mrnas. *Science*, 312(5770):75–9, Apr 2006.

Y. Grad, J. Aach, G. D. Hayes, B. J. Reinhart, G. M. Church, G. Ruvkun, and J. Kim. Computational and experimental identification of C. elegans microRNAs. *Mol Cell*, 11(5):1253–63, May 2003.

R. I. Gregory, K.-P. Yan, G. Amuthan, T. Chendrimada, B. Doratotaj, N. Cooch, and R. Shiekhattar. The microprocessor complex mediates the genesis of microRNAs. *Nature*, 432(7014):235–40, Nov 2004.

S. Griffiths-Jones, H. K. Saini, S. van Dongen, and A. J. Enright. miRBase: tools for microRNA genomics. *Nucleic Acids Res*, 36(Database issue):D154–8, Jan 2008.

A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, Jul 2007.

A. Grishok, A. E. Pasquinelli, D. Conte, N. Li, S. Parrish, I. Ha, D. L. Baillie, A. Fire, G. Ruvkun, and C. C. Mello. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell*, 106(1):23–34, Jul 2001.

N. Gromak, S. West, and N. J. Proudfoot. Pause sites promote transcriptional termination of mammalian RNA polymerase II. *Mol Cell Biol*, 26(10):3986–96, May 2006.

A. J. Hamilton and D. C. Baulcombe. A species of small antisense RNA in posttranscriptional gene silencing in plants. *Science*, 286(5441):950–2, Oct 1999.

S. M. Hammond, S. Boettcher, A. A. Caudy, R. Kobayashi, and G. J. Hannon. Argonaute2, a link between genetic and biochemical analyses of RNAi. *Science*, 293(5532):1146–50, Aug 2001.

J. Han, Y. Lee, K.-H. Yeom, Y.-K. Kim, H. Jin, and V. N. Kim. The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev*, 18(24):3016–27, Dec 2004.

J. Han, Y. Lee, K.-H. Yeom, J.-W. Nam, I. Heo, J.-K. Rhee, S. Y. Sohn, Y. Cho, B.-T. Zhang, and V. N. Kim. Molecular basis for the recognition of primary microRNAs by the Drosha-DGCR8 complex. *Cell*, 125(5):887–901, Jun 2006.

A. H. Hassan, P. Prochasson, K. E. Neely, S. C. Galasinski, M. Chandy, M. J. Carrozza, and J. L. Workman. Function and selectivity of bromodomains in anchoring chromatin-modifying complexes to promoter nucleosomes. *Cell*, 111(3):369–79, Nov 2002.

Y. Hirose and J. L. Manley. RNA polymerase II is an essential mRNA polyadenylation factor. *Nature*, 395(6697):93–6, Sep 1998.

D. T. Humphreys, B. J. Westman, D. I. K. Martin, and T. Preiss. MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. *Proc Natl Acad Sci USA*, 102(47):16961–6, Nov 2005.

G. Hutvágner and P. D. Zamore. A microRNA in a multiple-turnover RNAi enzyme complex. *Science*, 297(5589):2056–60, Sep 2002.

G. Hutvágner, J. McLachlan, A. E. Pasquinelli, E. Bálint, T. Tuschl, and P. D. Zamore. A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science*, 293(5531):834–8, Aug 2001.

I. P. Ioshikhes, I. Albert, S. J. Zanton, and B. F. Pugh. Nucleosome positions predicted through comparative genomics. *Nat Genet*, 38(10):1210–5, Oct 2006.

A. Jakymiw, S. Lian, T. Eystathioy, S. Li, M. Satoh, J. C. Hamel, M. J. Fritzler, and E. K. L. Chan. Disruption of GW bodies impairs mammalian RNA interference. *Nat Cell Biol*, 7(12):1267–74, Dec 2005.

M. W. Jones-Rhoades, D. P. Bartel, and B. Bartel. MicroRNAs and their regulatory roles in plants. *Annual review of plant biology*, 57:19–53, Jan 2006.

M. Kedde, M. J. Strasser, B. Boldajipour, J. A. F. O. Vrielink, K. Slanchev, C. le Sage, R. Nagel, P. M. Voorhoeve, J. van Duijse, U. A. Ørom, A. H. Lund, A. Perrakis, E. Raz, and R. Agami. RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell*, 131(7):1273–86, Dec 2007.

R. F. Ketting, S. E. Fischer, E. Bernstein, T. Sijen, G. J. Hannon, and R. H. Plasterk. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes Dev*, 15(20):2654–9, Oct 2001.

A. Khvorova, A. Reynolds, and S. D. Jayasena. Functional siRNAs and miRNAs exhibit strand bias. *Cell*, 115(2):209–16, Oct 2003.

M. Kim, N. J. Krogan, L. Vasiljeva, O. J. Rando, E. Nedea, J. F. Greenblatt, and S. Buratowski. The yeast Rat1 exonuclease promotes transcription termination by RNA polymerase II. *Nature*, 432(7016):517–22, Nov 2004.

M. Kim, L. Vasiljeva, O. J. Rando, A. Zhelkovsky, C. Moore, and S. Buratowski. Distinct pathways for snoRNA and mRNA termination. *Mol Cell*, 24(5):723–34, Dec 2006.

T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–80, Aug 2005.

M. Kiriakidou, G. S. Tan, S. Lamprinaki, M. D. Planell-Saguer, P. T. Nelson, and Z. Mourelatos. An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell*, 129(6):1141–51, Jun 2007.

J. A. Knezetic and D. S. Luse. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell*, 45(1):95–104, Apr 1986.

A. Y. Konev, M. Tribus, S. Y. Park, V. Podhraski, C. Y. Lim, A. V. Emelyanov, E. Vershilova, V. Pirrotta, J. T. Kadonaga, A. Lusser, and D. V. Fyodorov. CHD1 motor protein is required for deposition of histone variant H3.3 into chromatin in vivo. *Science*, 317(5841):1087–90, Aug 2007.

R. D. Kornberg and Y. Lorch. Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, 98(3):285–94, Aug 1999.

A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, May 2005.

N. J. Krogan, J. Dover, A. Wood, J. Schneider, J. Heidt, M. A. Boateng, K. Dean, O. W. Ryan, A. Golshani, M. Johnston, J. F. Greenblatt, and A. Shilatifard. The Paf1 complex is required for histone H3 methylation by COMPASS and Dot1p:

linking transcriptional elongation to histone methylation. *Mol Cell*, 11(3):721–9, Mar 2003.

M. Lagos-Quintana, R. Rauhut, W. Lendeckel, and T. Tuschl. Identification of novel genes coding for small expressed RNAs. *Science*, 294(5543):853–8, Oct 2001.

E. C. Lai. MicroRNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet*, 30(4):363–4, Apr 2002.

S. Lall, D. Grün, A. Krek, K. Chen, Y.-L. Wang, C. N. Dewey, P. Sood, T. Colombo, N. Bray, P. Macmenamin, H.-L. Kao, K. C. Gunsalus, L. Pachter, F. Piano, and N. Rajewsky. A genome-wide map of conserved microRNA targets in C. elegans. *Curr Biol*, 16(5):460–71, Mar 2006.

M. Landthaler, A. Yalcin, and T. Tuschl. The human DiGeorge syndrome critical region gene 8 and its D. melanogaster homolog are required for miRNA biogenesis. *Curr Biol*, 14(23):2162–7, Dec 2004.

N. C. Lau, L. P. Lim, E. G. Weinstein, and D. P. Bartel. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science*, 294 (5543):858–62, Oct 2001.

C.-K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*, 36(8):900–5, Aug 2004a.

R. C. Lee and V. Ambros. An extensive class of small RNAs in Caenorhabditis elegans. *Science*, 294(5543):862–4, Oct 2001.

R. C. Lee, R. L. Feinbaum, and V. Ambros. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell*, 75(5):843–54, Dec 1993.

T. I. Lee and R. A. Young. Transcription of eukaryotic protein-coding genes. *Annu Rev Genet*, 34:77–137, Jan 2000.

W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet*, 39(10):1235–44, Oct 2007.

Y. Lee, K. Jeon, J.-T. Lee, S. Kim, and V. N. Kim. MicroRNA maturation: stepwise processing and subcellular localization. *EMBO J*, 21(17):4663–70, Sep 2002.

Y. Lee, C. Ahn, J. Han, H. Choi, J. Kim, J. Yim, J. Lee, P. Provost, O. Rådmark, S. Kim, and V. N. Kim. The nuclear RNase III Drosha initiates microRNA processing. *Nature*, 425(6956):415–9, Sep 2003.

Y. Lee, M. Kim, J. Han, K.-H. Yeom, S. Lee, S. H. Baek, and V. N. Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J*, 23(20):4051–60, Oct 2004b.

Y. S. Lee, K. Nakahara, J. W. Pham, K. Kim, Z. He, E. J. Sontheimer, and R. W. Carthew. Distinct roles for Drosophila Dicer-1 and Dicer-2 in the siRNA/miRNA silencing pathways. *Cell*, 117(1):69–81, Apr 2004c.

M. Legendre, W. Ritchie, F. Lopez, and D. Gautheret. Differential repression of alternative transcripts: a screen for miRNA targets. *PLoS Comput Biol*, 2(5):e43, May 2006.

A. K. L. Leung, J. M. Calabrese, and P. A. Sharp. Quantitative analysis of Argonaute protein reveals microRNA-dependent localization to stress granules. *Proc Natl Acad Sci USA*, 103(48):18125–30, Nov 2006.

B. P. Lewis, I. hung Shih, M. W. Jones-Rhoades, D. P. Bartel, and C. B. Burge. Prediction of mammalian microRNA targets. *Cell*, 115(7):787–98, Dec 2003.

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005.

B. Li, M. Carey, and J. L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–19, Feb 2007.

L. P. Lim, M. E. Glasner, S. Yekta, C. B. Burge, and D. P. Bartel. Vertebrate microRNA genes. *Science*, 299(5612):1540, Mar 2003.

L. P. Lim, N. C. Lau, P. Garrett-Engele, A. Grimson, J. M. Schelter, J. Castle, D. P. Bartel, P. S. Linsley, and J. M. Johnson. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature*, 433(7027): 769–73, Feb 2005.

J. Liu, M. A. Carmell, F. V. Rivas, C. G. Marsden, J. M. Thomson, J.-J. Song, S. M. Hammond, L. Joshua-Tor, and G. J. Hannon. Argonaute2 is the catalytic engine of mammalian RNAi. *Science*, 305(5689):1437–41, Sep 2004.

J. Liu, F. V. Rivas, J. Wohlschlegel, J. R. Yates, R. Parker, and G. J. Hannon. A role for the P-body component GW182 in microRNA function. *Nat Cell Biol*, 7 (12):1261–6, Dec 2005a.

J. Liu, M. A. Valencia-Sanchez, G. J. Hannon, and R. Parker. MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. *Nat Cell Biol*, 7(7):719–23, Jul 2005b.

J. Logan, E. Falck-Pedersen, J. E. Darnell, and T. Shenk. A poly(A) addition site and a downstream termination region are required for efficient cessation of transcription by RNA polymerase II in the mouse beta maj-globin gene. *Proc Natl Acad Sci USA*, 84(23):8306–10, Dec 1987.

J. R. Lytle, T. A. Yario, and J. A. Steitz. Target mRNAs are repressed as efficiently by microRNA-binding sites in the 5′ UTR as in the 3′ UTR. *Proc Natl Acad Sci USA*, 104(23):9667–72, Jun 2007.

J.-B. Ma, Y.-R. Yuan, G. Meister, Y. Pei, T. Tuschl, and D. J. Patel. Structural basis for 5′-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. *Nature*, 434(7033):666–70, Mar 2005.

I. J. MacRae, K. Zhou, F. Li, A. Repic, A. N. Brooks, W. Z. Cande, P. D. Adams, and J. A. Doudna. Structural basis for double-stranded RNA processing by Dicer. *Science*, 311(5758):195–8, Jan 2006.

I. J. MacRae, K. Zhou, and J. A. Doudna. Structural determinants of RNA recognition and cleavage by Dicer. *Nat Struct Mol Biol*, 14(10):934–40, Oct 2007.

C. R. Mandel, S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley, and L. Tong. Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature*, 444(7121):953–6, Dec 2006.

P. A. Maroney, Y. Yu, J. Fisher, and T. W. Nilsen. Evidence that microRNAs are associated with translating messenger RNAs in human cells. *Nat Struct Mol Biol*, 13(12):1102–7, Dec 2006.

J. Martinez, A. Patkaniowska, H. Urlaub, R. Lührmann, and T. Tuschl. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell*, 110(5): 563–74, Sep 2002.

C. Matranga, Y. Tomari, C. Shin, D. P. Bartel, and P. D. Zamore. Passenger-strand cleavage facilitates assembly of siRNA into Ago2-containing RNAi enzyme complexes. *Cell*, 123(4):607–20, Nov 2005.

S. McCracken, N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens, and D. L. Bentley. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, 385(6614):357–61, Jan 1997.

Y. Mishima, A. J. Giraldez, Y. Takeda, T. Fujiwara, H. Sakamoto, A. F. Schier, and K. Inoue. Differential regulation of germline mRNAs in soma and germ cells by zebrafish miR-430. *Curr Biol*, 16(21):2135–42, Nov 2006.

E. Miska, E. Alvarez-Saavedra, A. Abbott, N. Lau, A. Hellman, S. McGonagle, D. Bartel, V. Ambros, and H. Horvitz. Most Caenorhabditis elegans microRNAs are individually not essential for development or viability. *PLoS Genet*, 3(12):e215, Dec 2007.

K. Miyoshi, H. Tsukumo, T. Nagami, H. Siomi, and M. C. Siomi. Slicer function of Drosophila Argonautes and its involvement in RISC formation. *Genes Dev*, 19(23): 2837–48, Dec 2005.

A. Molnár, F. Schwach, D. J. Studholme, E. C. Thuenemann, and D. C. Baulcombe. miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii. *Nature*, 447(7148):1126–9, Jun 2007.

A. Morillon, N. Karabetsou, J. O'Sullivan, N. Kent, N. Proudfoot, and J. Mellor. Isw1 chromatin remodeling ATPase coordinates transcription elongation and termination by RNA polymerase II. *Cell*, 115(4):425–35, Nov 2003.

R. D. Morin, M. D. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, A.-L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C. J. Eaves, and M. A. Marra. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res*, 18(4):610–21, Apr 2008.

A. Nag, K. Narsinh, and H. G. Martinson. The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nat Struct Mol Biol*, 14(7):662–9, Jul 2007.

H. H. Ng, F. Robert, R. A. Young, and K. Struhl. Targeted recruitment of Set1 histone methylase by elongating Pol II provides a localized mark and memory of recent transcriptional activity. *Mol Cell*, 11(3):709–19, Mar 2003.

C. B. Nielsen, N. Shomron, R. Sandberg, E. Hornstein, J. Kitzman, and C. B. Burge. Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA*, 13(11):1894–910, Nov 2007.

P. H. Olsen and V. Ambros. The lin-4 regulatory RNA controls developmental timing in Caenorhabditis elegans by blocking LIN-14 protein synthesis after the initiation of translation. *Dev Biol*, 216(2):671–80, Dec 1999.

I. J. Orozco, S. J. Kim, and H. G. Martinson. The poly(A) signal, without the assistance of any downstream element, directs RNA polymerase II to pause in vivo and then to release stochastically from the template. *J Biol Chem*, 277(45):42899–911, Nov 2002.

J. S. Parker, S. M. Roe, and D. Barford. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature*, 434(7033):663–6, Mar 2005.

A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, J. Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature*, 408(6808):86–9, Nov 2000.

R. Pavri, B. Zhu, G. Li, P. Trojer, S. Mandal, A. Shilatifard, and D. Reinberg. Histone H2B monoubiquitination functions cooperatively with FACT to regulate elongation by RNA polymerase II. *Cell*, 125(4):703–17, May 2006.

H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng. Nucleosome positioning signals in genomic DNA. *Genome Res*, 17(8):1170–7, Aug 2007.

C. P. Petersen, M.-E. Bordeleau, J. Pelletier, and P. A. Sharp. Short RNAs repress translation after initiation in mammalian cells. *Mol Cell*, 21(4):533–42, Feb 2006.

R. S. Pillai, S. N. Bhattacharyya, C. G. Artus, T. Zoller, N. Cougot, E. Basyuk, E. Bertrand, and W. Filipowicz. Inhibition of translational initiation by let-7 microRNA in human cells. *Science*, 309(5740):1573–6, Sep 2005.

D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young. Genome-wide map of nucleosome acetylation and methylation in yeast. *Cell*, 122(4):517–27, Aug 2005.

M. G. Pray-Grant, J. A. Daniel, D. Schieltz, J. R. Yates, and P. A. Grant. Chd1 chromodomain links histone H3 methylation with SAGA- and SLIK-dependent acetylation. *Nature*, 433(7024):434–8, Jan 2005.

R. M. Raisner, P. D. Hartley, M. D. Meneghini, M. Z. Bao, C. L. Liu, S. L. Schreiber, O. J. Rando, and H. D. Madhani. Histone variant H2A.Z marks the 5′ ends of both active and inactive genes in euchromatin. *Cell*, 123(2):233–48, Oct 2005.

R. Rajagopalan, H. Vaucheret, J. Trejo, and D. P. Bartel. A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev*, 20(24):3407–25, Dec 2006.

N. Rajewsky and N. D. Socci. Computational identification of microRNA targets. *Dev Biol*, 267(2):529–35, Mar 2004.

T. A. Rand, S. Petersen, F. Du, and X. Wang. Argonaute2 cleaves the anti-guide strand of siRNA during RISC activation. *Cell*, 123(4):621–9, Nov 2005.

E. B. Rasmussen and J. T. Lis. In vivo transcriptional pausing and cap formation on three Drosophila heat shock genes. *Proc Natl Acad Sci USA*, 90(17):7923–7, Sep 1993.

B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature*, 403(6772):901–6, Feb 2000.

H. Reinke and W. Hörz. Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. *Mol Cell*, 11(6):1599–607, Jun 2003.

T. J. Richmond and C. A. Davey. The structure of DNA in the nucleosome core. *Nature*, 423(6936):145–50, May 2003.

T. J. Richmond, J. T. Finch, B. Rushton, D. Rhodes, and A. Klug. Structure of the nucleosome core particle at °7A resolution. *Nature*, 311(5986):532–7, Jan 1984.

F. V. Rivas, N. H. Tolia, J.-J. Song, J. P. Aragon, J. Liu, G. J. Hannon, and L. Joshua-Tor. Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat Struct Mol Biol*, 12(4):340–9, Apr 2005.

F. Robert, D. K. Pokholok, N. M. Hannett, N. J. Rinaldi, M. Chandy, A. Rolfe, J. L. Workman, D. K. Gifford, and R. A. Young. Global position and recruitment of HATs and HDACs in the yeast genome. *Mol Cell*, 16(2):199–209, Oct 2004.

A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Res*, 14(10A): 1902–10, Oct 2004.

A. Rodriguez, E. Vigorito, S. Clare, M. V. Warren, P. Couttet, D. R. Soond, S. van Dongen, R. J. Grocock, P. P. Das, E. A. Miska, D. Vetrie, K. Okkenhaug, A. J. Enright, G. Dougan, M. Turner, and A. Bradley. Requirement of bic/microRNA-155 for normal immune function. *Science*, 316(5824):608–11, Apr 2007.

A. E. Rougvie and J. T. Lis. The RNA polymerase II molecule at the 5′ end of the uninduced hsp70 gene of D. melanogaster is transcriptionally engaged. *Cell*, 54(6): 795–804, Sep 1988.

J. Ruby, A. Stark, W. Johnston, M. Kellis, D. Bartel, and E. Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *Genome Res*, Nov 2007a.

J. G. Ruby, C. Jan, C. Player, M. J. Axtell, W. Lee, C. Nusbaum, H. Ge, and D. P. Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans. *Cell*, 127(6):1193–207, Dec 2006.

J. G. Ruby, C. H. Jan, and D. P. Bartel. Intronic microRNA precursors that bypass Drosha processing. *Nature*, 448(7149):83–6, Jul 2007b.

K. Ryan, O. Calvo, and J. L. Manley. Evidence that polyadenylation factor CPSF-73 is the mrna 3′ processing endonuclease. *RNA*, 10(4):565–73, Apr 2004.

S. C. Satchwell, H. R. Drew, and A. A. Travers. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol*, 191(4):659–75, Oct 1986.

A. Saunders, L. J. Core, and J. T. Lis. Breaking barriers to transcription elongation. *Nat Rev Mol Cell Biol*, 7(8):557–67, Aug 2006.

V. Scaria, M. Hariharan, B. Pillai, S. Maiti, and S. K. Brahmachari. Host-virus genome interactions: macro roles for microRNAs. *Cell Microbiol*, 9(12):2784–94, Dec 2007.

D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–98, Mar 2008.

B. E. Schwartz and K. Ahmad. Transcriptional activation triggers deposition and removal of the histone variant H3.3. *Genes Dev*, 19(7):804–14, Apr 2005.

D. S. Schwarz, G. Hutvágner, T. Du, Z. Xu, N. Aronin, and P. D. Zamore. Asymmetry in the assembly of the RNAi enzyme complex. *Cell*, 115(2):199–208, Oct 2003.

E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–8, Aug 2006.

K. Seggerson, L. Tang, and E. G. Moss. Two genetic circuits repress the Caenorhabditis elegans heterochronic gene lin-28 after translation initiation. *Dev Biol*, 243 (2):215–25, Mar 2002.

E. A. Sekinger, Z. Moqtaderi, and K. Struhl. Intrinsic histone-DNA interactions and low nucleosome density are important for preferential accessibility of promoter regions in yeast. *Mol Cell*, 18(6):735–48, Jun 2005.

G. L. Sen and H. M. Blau. Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. *Nat Cell Biol*, 7(6):633–6, Jun 2005.

U. Sheth and R. Parker. Decapping and decay of messenger RNA occur in cytoplasmic processing bodies. *Science*, 300(5620):805–8, May 2003.

A. Stark, J. Brennecke, R. B. Russell, and S. M. Cohen. Identification of Drosophila MicroRNA targets. *PLoS Biol*, 1(3):E60, Dec 2003.

A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, 123(6):1133–46, Dec 2005.

A. Stark, M. Lin, P. Kheradpour, J. Pedersen, L. Parts, M. Rasmussen, S. Roy, A. Deoras, J. Ruby, J. Brennecke, H. Curators, B. Project, E. Hodges, A. Hinrichs, A. Caspi, B. Paten, S. Park, M. Han, M. Maeder, B. Polansky, B. Robson, S. Aerts, J. van Helden, B. Hassan, D. Gilbert, D. Eastman, M. Rice, M. Weir, M. Hahn, Y. Park, C. Dewey, L. Pachter, W. Kent, D. Haussler, E. Lai, D. Bartel, G. Hannon, T. Kaufman, M. Eisen, A. Clark, D. Smith, W. Gelbart, M. Kellis, H. F. curators, M. Crosby, B. Matthews, A. Schroeder, L. S. Gramates, S. S. Pierre, M. Roark, K. W. Jr, R. Kulathinal, P. Zhang, K. Myrick, J. Antone, B. D. G. Project, J. Carlson, C. Yu, S. Park, K. Wan, and S. Celniker. Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232, Nov 2007.

I. C. Taylor, J. L. Workman, T. J. Schuetz, and R. E. Kingston. Facilitated binding of GAL4 and heat shock factor to nucleosomal templates: differential function of DNA-binding domains. *Genes Dev*, 5(7):1285–98, Jul 1991.

D. Teixeira, U. Sheth, M. A. Valencia-Sanchez, M. Brengues, and R. Parker. Processing bodies require RNA for assembly and contain nontranslating mRNAs. *RNA*, 11(4):371–82, Apr 2005.

D. P. Tran, S. J. Kim, N. J. Park, T. M. Jew, and H. G. Martinson. Mechanism of poly(A) signal transduction to RNA polymerase II in vitro. *Mol Cell Biol*, 21(21): 7495–508, Nov 2001.

A. A. Travers and A. Klug. The bending of DNA in nucleosomes and its wider implications. *Philos Trans R Soc Lond, B, Biol Sci*, 317(1187):537–61, Dec 1987.

D. J. Tremethick. Higher-order structures of chromatin: the elusive 30 nm fiber. *Cell*, 128(4):651–4, Feb 2007.

E. N. Trifonov and J. L. Sussman. The pitch of chromatin DNA is reflected in its nucleotide sequence. *Proc Natl Acad Sci USA*, 77(7):3816–20, Jul 1980.

W. G. Turnell and A. A. Travers. Algorithms for prediction of histone octamer binding sites. *Meth Enzymol*, 212:387–99, Jan 1992.

S. West, N. Gromak, and N. J. Proudfoot. Human $5'- > 3'$ exonuclease Xrn2 promotes transcription termination at co-transcriptional cleavage sites. *Nature*, 432 (7016):522–5, Nov 2004.

E. Whitelaw and N. Proudfoot. Alpha-thalassaemia caused by a poly(A) site mutation reveals that transcriptional termination is linked to $3'$ end processing in the human alpha 2 globin gene. *EMBO J*, 5(11):2915–22, Nov 1986.

H. R. Widlund, H. Cao, S. Simonsson, E. Magnusson, T. Simonsson, P. E. Nielsen, J. D. Kahn, D. M. Crothers, and M. Kubista. Identification and characterization of genomic nucleosome-positioning sequences. *J Mol Biol*, 267(4):807–17, Apr 1997.

B. Wightman, I. Ha, and G. Ruvkun. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell*, 75(5):855–62, Dec 1993.

T. Woodage, M. A. Basrai, A. D. Baxevanis, P. Hieter, and F. S. Collins. Characterization of the CHD family of proteins. *Proc Natl Acad Sci USA*, 94(21):11472–7, Oct 1997.

J. L. Workman. Nucleosome displacement in transcription. *Genes Dev*, 20(15):2009–17, Aug 2006.

L. Wu and J. G. Belasco. MicroRNA regulation of the mammalian lin-28 gene during neuronal differentiation of embryonal carcinoma cells. *Mol Cell Biol*, 25(21):9198–208, Nov 2005.

L. Wu and J. G. Belasco. Let me count the ways: mechanisms of gene regulation by miRNAs and siRNAs. *Mol Cell*, 29(1):1–7, Jan 2008.

L. Wu, J. Fan, and J. G. Belasco. MicroRNAs direct rapid deadenylation of mRNA. *Proc Natl Acad Sci USA*, 103(11):4034–9, Mar 2006.

S. Yekta, I. hung Shih, and D. P. Bartel. MicroRNA-directed cleavage of HOXB8 mRNA. *Science*, 304(5670):594–6, Apr 2004.

R. Yi, Y. Qin, I. G. Macara, and B. R. Cullen. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. *Genes Dev*, 17(24):3011–6, Dec 2003.

M. Yonaha and N. J. Proudfoot. Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol Cell*, 3(5):593–600, May 1999.

G.-C. Yuan and J. S. Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol*, 4(1):e13, Jan 2008.

G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309(5734):626–30, Jul 2005.

P. D. Zamore, T. Tuschl, P. A. Sharp, and D. P. Bartel. Rnai: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell*, 101(1):25–33, Mar 2000.

Y. Zeng, R. Yi, and B. R. Cullen. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. *Proc Natl Acad Sci USA*, 100 (17):9779–84, Aug 2003.

H. Zhang, F. A. Kolb, L. Jaskiewicz, E. Westhof, and W. Filipowicz. Single processing center models for human Dicer and bacterial RNase III. *Cell*, 118(1):57–68, Jul 2004.

H. Zhang, D. N. Roberts, and B. R. Cairns. Genome-wide dynamics of Htz1, a histone H2A variant that poises repressed/basal promoters for activation through histone loss. *Cell*, 123(2):219–31, Oct 2005.

J. Zhao, L. Hyman, and C. Moore. Formation of mRNA 3′ ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiology and Molecular Biology Reviews*, 63(2):405–445, 1999.

T. Zhao, G. Li, S. Mi, S. Li, G. J. Hannon, X.-J. Wang, and Y. Qi. A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii. *Genes Dev*, 21(10):1190–203, May 2007a.

Y. Zhao, J. F. Ransom, A. Li, V. Vedantham, M. von Drehle, A. N. Muth, T. Tsuchihashi, M. T. McManus, R. J. Schwartz, and D. Srivastava. Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. *Cell*, 129 (2):303–17, Apr 2007b.

Chapter 2

Determinants of Targeting by Endogenous and Exogenous microRNAs and siRNAs

Cydney Nielsen, Noam Shomron, Rickard Sandberg, Eran Hornstein, Jacob Kitzman and Christopher Burge

This chapter is presented in the context of its contemporary science and originally appeared in *RNA* 13:1894-1910 (2007).

Corresponding Supplementary Material can be found in Appendix 2.

Chapter 2

Determinants of Targeting by Endogenous and Exogenous microRNAs and siRNAs

Abstract

Vertebrate mRNAs are frequently targeted for post-transcriptional repression by microRNAs (miRNAs) through mechanisms involving pairing of 3′ UTR seed matches to bases at the 5′ end of miRNAs. Through analysis of expression array data following miRNA or siRNA overexpression or inhibition, we found that mRNA fold change increases multiplicatively (i.e., log-additively) with seed match count and that a single 8 mer seed match mediates down-regulation comparable to two 7 mer seed matches. We identified several targeting determinants that enhance seed match-associated mRNA repression, including the presence of adenosine opposite miRNA base 1 and of adenosine or uridine opposite miRNA base 9, independent of complementarity to the siRNA/miRNA. Increased sequence conservation in the 50 bases 5′ and 3′ of the seed match and increased AU content 3′ of the seed match were each independently associated with increased mRNA down-regulation. All of these determinants are enriched in the vicinity of conserved miRNA seed matches, sup-

porting their activity in endogenous miRNA targeting. Together, our results enable improved siRNA off-target prediction, allow integrated ranking of conserved and non-conserved miRNA targets, and show that targeting by endogenous and exogenous miRNAs/siRNAs involves similar or identical determinants.

Introduction

Precise control of mRNA and protein levels in different cell types requires regulation at multiple levels. In metazoans, a large proportion of mRNAs is targeted for post-transcriptional repression by ∼22 nucleotide (nt) microRNAs (miRNAs). Identified as developmental regulators, miRNAs are now known to play roles in diverse biological processes including control of proliferation, apoptosis, stress resistance, and metabolism (Ambros 2004; Bartel 2004; Filipowicz et al. 2005; Zamore and Haley 2005).

miRNAs were initially described as exerting their effects primarily by inhibiting productive translation of mRNAs (Lee et al. 1993; Wightman et al. 1993). More recently, several studies have demonstrated that animal miRNAs can direct accelerated decay of targeted mRNAs (Hutvagner and Zamore 2002; Bagga et al. 2005; Lim et al. 2005; Rehwinkel et al. 2005; Giraldez et al. 2006) and that siRNAs commonly direct decay of off-target mRNAs (Jackson et al. 2003). When they possess near-perfect complementarity to a targeted mRNA, miRNAs can direct endoribonucleolytic cleavage of mRNAs (slicer activity) by Argonaute2 (AGO2) (Hutvagner and Zamore 2002; Llave et al. 2002; Meister et al. 2004). This type of targeting is predominant in plants, but appears to occur only rarely for animal miRNAs (Yekta et al. 2004; Jones-Rhoades et al. 2006). For typical metazoan targets that possess comple-

mentarity only to a segment at the miRNA 5′ end, miRNAs appear to direct mRNA degradation by mechanisms that may involve AGO2 but do not appear to involve its slicer activity (Bagga et al. 2005; Schmitter et al. 2006). Instead, decay may be promoted by relocalization of targeted mRNAs to specific cytoplasmic locations, which can be sites of mRNA decapping and degradation (for review, see Valencia-Sanchez et al. 2006) and/or by acceleration of mRNA deadenylation (Giraldez et al. 2006; Wu et al. 2006).

In many studies, miRNA regulation has been assessed only at the protein level, without distinguishing the relative contributions of effects on mRNA decay and on inhibition of translation. However, for some individual targets, both mRNA-level and protein-level effects have been measured. For the classical *let-7* and *lin-4* target genes *lin-41*, *lin-14*, and *lin-28* (Lee et al. 1993; Wightman et al. 1993), a recent study found a predominant effect on mRNA stability (Bagga et al. 2005). Studies of transfected miRNAs or siRNAs using transfected reporters with moderate degrees of complementarity have typically reported significant effects on protein levels, with modest or negligible effects on mRNA levels (Zeng et al. 2002, 2003; Doench et al. 2003; Doench and Sharp 2004). However, studies that have examined changes in the expression of endogenous mRNAs in response to manipulation of miRNAs have generally observed widespread miRNA-associated changes in mRNA levels. Following miRNA overexpression, Lim and colleagues (2005) observed down-regulation of sets of mRNAs that were enriched for predicted miRNA targets and for genes with low expression levels in the tissues where the miRNAs were naturally expressed, supporting the physiological relevance of this effect. Inhibiting the expression of the critical miRNA processing enzymes Dicer and Drosha also yields specific derepression of predicted miRNA targets at the mRNA level (Rehwinkel et al. 2005; Giraldez et al.

2006; Schmitter et al. 2006). Thus, perturbations of miRNA expression commonly affect the levels of endogenous mRNAs, and effects on mRNA stability appear to be an important component of the endogenous function of miRNAs.

The special importance of the miRNA 5′ end was suggested by early studies (Lee et al. 1993; Wightman et al. 1993; Lai 2002). Since then, the critical importance of pairing to the miRNA seed, comprising bases 2-7 from the miRNA 5′ end, has been established through extensive comparative genomic and experimental studies (Lewis et al. 2003, 2005; Doench and Sharp 2004; Brennecke et al. 2005; Stark et al. 2005). The degree of conservation above background in orthologous 3′ UTRs of seed match segments having WatsonCrick (WC) complementarity (matching) to the seed regions of conserved miRNAs can be used to estimate the number of conserved targets. This approach indicated that at least one-third of mammalian mRNAs are conserved targets of one or more conserved miRNAs (Lewis et al. 2005), and related methods have indicated that a comparably large fraction of *Drosophila* mRNAs represent conserved miRNA targets (Brennecke et al. 2005; Grun et al. 2005). Recent analyses of mRNA sequence and expression patterns have detected pervasive effects of miRNAs on mRNA expression and evolution, suggesting that most mRNAs are subject either to direct miRNA regulation or to evolutionary pressure to avoid miRNA targeting (Farh et al. 2005; Stark et al. 2005).

Some targets identified genetically possess complementarity to bases at the 3′ as well as 5′ ends of miRNAs, which may confer specificity to individual members of a miRNA family. However, comparative genomic approaches have determined that "seed only" type targets comprise the vast majority of all conserved miRNA targets (Brennecke et al. 2005; Lewis et al. 2005). This conclusion is also supported by

miRNA overexpression experiments; e.g., in the study by Lim and colleagues (2005), 88% of mRNAs whose expression was significantly repressed following transfection of miR-1 contained seed matches in their 3′ UTRs, and replacing the 3′ end of the transfected miRNA by unrelated sequences yielded a largely overlapping set of down-regulated mRNAs. However, the presence of a minimal seed match is not generally sufficient to generate detectable mRNA down-regulation; e.g., only about one-tenth to one-twentieth of expressed genes containing a 6 nt seed match in the Lim study were significantly down-regulated (not shown), suggesting that miRNA regulation is strongly influenced by additional targeting determinants.

To assess a variety of potential targeting determinants, we analyzed the effects on global mRNA expression in miRNA and siRNA overexpression studies. In parallel, the effects on mRNA expression of endogenous mouse miRNAs were analyzed following knockout of the Dicer1 gene, which is essential for miRNA maturation in vertebrates. Our results uncover additional rules and determinants for targeting that hold for both endogenously expressed miRNAs and exogenous miRNAs and siRNAs.

Results and Discussion

A hierarchy of extended seed match types associated with different degrees of target down-regulation

To explore miRNA targeting determinants, we analyzed global mRNA expression data following transfection of the tissue-specific miRNAs miR-1 and miR-124 into HeLa cells reported by Lim and colleagues (2005). To assess the impact of a putative targeting determinant on down-regulation, we compared the distributions of log fold change (LFC), defined as the log base 2 of expression in miRNA-transfected cells over that in mock-transfected cells, for mRNAs containing and lacking the putative determinant. The cumulative distribution functions (CDFs) of LFCs for these two mRNA sets could then be compared and the significance of differences assessed using a Wilcoxon rank sum test (Materials and Methods). Using this approach, mRNA sets with and without specific hexanucleotides (6 mer) in their $3'$ UTRs were compared for all 4096 possible 6 mer in both miRNA transfection datasets. Following miR-1 transfection, the most significant down-regulation was observed for mRNAs containing the 6 mer CAUUCC, which has perfect WC complementarity to miR-1 bases 2-7 ($P < 10^{-34}$, Bonferroni corrected for the 4096 comparisons performed). For miR-124, the most significant down-regulation was associated with GUGCCU ($P < 10^{-58}$) and UGCCUU ($P < 10^{-26}$), which are complementary to miR-124 bases 3-8 and 2-7, respectively. These observations, obtained without the need to define significantly up- and down-regulated mRNA sets, are entirely consistent with the motif-finding analyses of significantly down-regulated mRNAs by Lim and colleagues (2005), and suggest that pairing to miRNA seed matches was a primary effector of mRNA down-regulation in this experiment.

Stronger down-regulation was observed for mRNAs containing additional matching to the transfected miRNAs in their 3′ UTRs beyond the 6-base seed match (Fig. 1B,D). As shown in Figure 1A, we use the notation m1, m2, ... to refer to miRNA bases, starting at the 5′-most base, and t1, t2, ... to refer to positions in target mRNAs opposite miRNA bases m1, m2, ..., respectively, in presumptive seed:seed match duplexes (Lewis et al. 2005). Those mRNAs that contained a seed match 6 mer flanked by a WC match to miRNA base 8 (Fig. 1, M8 7 mer; red curves) exhibited enhanced down-regulation relative to those that contained a 6 mer alone ($P < 10^{-5}$ for both miR-124 and miR-1). The presence of an adenosine at position t1 (Fig. 1, A1 7 mer; blue curves) was also associated with greater mRNA down-regulation than a seed match alone for both miRNAs ($P < 0.03$, $P < 10^{-5}$ for miR-124 and miR-1, respectively). Those mRNAs that contained seed matches flanked by both of these features (Fig. 1, M8-A1 8 mer; purple curves) exhibited greater mRNA down-regulation ($P < 0.002$ relative to A1 7 mer for both miRNAs). Modest but significant down-regulation was observed for mRNAs that contained only a seed match 6 mer not flanked by an M8 or A1 base (Fig. 1, 6 mer; green curves) for miR-1 ($P < 10^{-4}$), but not miR-124 (NS). Therefore, the highly significant down-regulation observed for the seed match 6 mer in the independent 6 mer analysis is attributable primarily to the effects of M8 and A1 7 mer and M8-A1 8 mer. We consider these 7 mer and 8 mer and the seed match 6 mer to represent distinct "seed match types" and refer to these 7 mer and 8 mer collectively as "extended seed matches". These observations suggest that the presence of these types of extended seed matches, not just of a seed match 6 mer, may be generally required for effective miRNA-directed down-regulation of mRNAs. A similar hierarchy of seed match types was observed when mRNAs containing conserved and nonconserved extended seed matches were analyzed separately (Supplemental Fig. S1). All mRNA sets in the above analyses

were mutually exclusive, and no significant differences between the distributions of expression levels of mRNAs containing different seed match types were detected by rank sum test.

There are multiple ways to think about the magnitude of the mRNA down-regulation effect attributable to a given seed match type. One perspective is to consider the set of mRNAs containing the given seed match type that were significantly down-regulated (e.g., those with LFC $< 97.5\%$ of control mRNAs lacking seed matches). By this criterion, 45% of expressed mRNAs containing 8 mer seed matches were down-regulated following miR-124 transfection. Among these genes, the average LFC was -0.97, corresponding to a $100 \times (1 - 2^{-0.97}) = 49\%$ decrease in expression. For M8 7 mer seed matches, 25% of mRNAs were significantly down-regulated, and these had a mean LFC of -0.87, a 45% decrease in expression. The fraction of mRNAs significantly down-regulated, two measures of the magnitude of down-regulation, and rank sum P-values for all of the analyses shown in Figure 1 and Supplemental Figure S1 are provided in Supplemental Tables S1 and S2, respectively.

Another perspective is to consider all of the data and to calculate the mean normalized log fold change (nLFC), defined as the mean LFC for expressed mRNAs lacking seed matches to the transfected miRNA minus the mean LFC for expressed mRNAs containing the given seed match type in their $3'$ UTRs. (As defined, the nLFC will be positive if a seed match type is associated with mRNA down-regulation.) For miR-124, the mean nLFC value of the M8-A1 8 mer seed match type was 0.56, roughly twice that for the M8 7 mer (0.25). Thus, if fold change is multiplicative (i.e., log-additive) in the number of seed matches (as will be shown below), then the fold change associated with one 8 mer seed match is roughly equivalent to that associated

with two 7 mer seed matches. Because it uses the largest possible set of mRNAs, and is less sensitive to the shape of the tail of the no-seed-match distribution, the mean nLFC is a more robust statistic for analyzing seed-match-associated effects than the fraction of significantly down-regulated mRNAs. For this reason, mean nLFC is used extensively in this study. However, by considering all seed-match-containing mRNAs, not just those with significant changes, the mean nLFC likely underestimates the true magnitude of miRNA effects on target mRNA levels, and mRNA fold change values underestimate protein-level changes (see below) because miRNAs often inhibit translation as well as mRNA stability.

Seed match hierarchy supported by siRNA, comparative genomic, and luciferase data

Exogenously added siRNAs complementary to seed match segments in mRNA 3′ UTRs have been observed to direct similar effects at both mRNA and protein levels as transfected miRNAs (Doench et al. 2003; Jackson et al. 2003, 2006). Using global mRNA expression data following transfection of siRNAs generated by Jackson and colleagues (Jackson et al. 2003, 2006) and Schwarz and colleagues (Schwarz et al. 2006), similar shifts in mRNA expression for seed-match-containing mRNAs were seen for siRNAs as were observed for transfected miRNAs. Those mRNAs that contained M8-A1 8 mer matches to the siRNAs were most strongly down-regulated, with the nLFC value for 8 mer roughly twice that seen for M8 or A1 7 mer, as was seen for transfected miRNAs. The effects of M8 7 mer were comparable to that for A1 7 mer, and both had nLFC values more than twice that of 6 mer, the same ordering of seed match types as was observed for transfected miRNAs (Fig. 1F). Thus, the targeting rules observed for transfected miRNAs generally apply to transfected siRNAs, suggesting that transfected siRNAs and miRNAs enter similar or identical silencing

complexes and mediate similar effects on their targets (Hutvagner and Zamore 2002).

Analyses of sequence conservation in mammalian 3′ UTRs have previously found that a 6 mer seed match is the minimal unit of sequence that suffices to elicit a significant conservation signal above noise for conserved vertebrate miRNAs (Lewis et al. 2005), but that requiring conservation of M8 and/or A1 bases greatly increased the signal:noise ratio. In alignments of five vertebrate genomes, the signal:noise ratio increased from 2.4:1 for 6 mer to 3.8:1 each for M8 and A1 7 mer, to 5.6:1 for M8-A1 8 mer (Lewis et al. 2005). Thus, M8 and A1 bases adjacent to conserved seed matches in mammalian 3′ UTRs are very often conserved, and the relative ordering of comparative genomic signal:noise ratios for different seed match types generally agreed with the relative magnitude of mRNA down-regulation effects observed above for transfected miRNAs (i.e., M8-A1 8 mer $>$ M8 7 mer $\geq$ A1 7 mer $>$ 6 mer). The agreement between these two orderings suggests that the miRNA effects on mRNA levels captured by microarrays are tightly correlated with the fold protein down-regulation - resulting from the product of mRNA decay and translational effects - which is presumably the effect that is under selection.

Comparing data from a panel of luciferase reporters following miRNA transfection (Farh et al. 2005) to fold change values measured by microarray (Lim et al. 2005) for the corresponding endogenous mRNAs (Supplemental Table S3), we observed a significant Spearman rank correlation of 0.63, despite the obvious differences in UTR context and whatever experimental noise was present in these assays. (As expected, average protein-level repression was somewhat larger than repression at the mRNA level.) This observation, though based on a small sample of genes, suggests that for typical targets, effects of miRNAs at the mRNA and protein levels may be reasonably

well correlated.

Data from the panel of luciferase reporters (Farh et al. 2005) could also be used to address the effects of different seed match types. We observed that those reporters that contained at least one 8 mer seed match were more strongly repressed than those that contained exclusively 7 mer seed matches ($P < 0.05$ by rank sum test). Further, among those reporters containing exclusively 7 mer seed matches, those with at least one M8 7 mer were more strongly repressed than those containing exclusively A1 7 mer ($P < 0.01$ by rank sum test). Thus, the hierarchy of seed match types observed in the mRNA array data appears to hold also when miRNA effects were assessed at the protein level.

Effects of seed matches located in regions other than the 3′ UTR were either very modest (coding regions) or not detected (5′ UTRs), and so were not further explored here (not shown).

Evidence for direct recognition of t1 adenosines by the silencing complex

Preferential conservation of adenosine residues at the t1 position adjacent to miRNA seed matches was reported previously, even for the minority of miRNAs that do not begin with U (and have no known paralogs that begin with U). This observation led to the hypothesis that t1A residues in target mRNAs can be recognized directly by the silencing complex, in a manner that does not require pairing to the m1 base of the miRNA (Lewis et al. 2005). To directly test this hypothesis, we turned to data from three siRNA transfection studies by Jackson and colleagues (Jackson et al. 2003, 2006) and Schwarz and colleagues (Schwarz et al. 2006). To distinguish between direct recognition of t1A and possible base-pairing to miRNA base m1, Figure

1F includes data only for siRNAs whose first base was not U, representing 33 of the 44 "effective" siRNAs in these studies (see Supplemental Material). Strikingly, we observed stronger mRNA down-regulation associated with A1 7 mer (which lack complementarity to base m1) than for M1 7 mer (which have a WC match to base m1) for these siRNAs (Fig. 1F, cf. solid blue curve, blue triangles and inset nLFC plot, $P < 10^{-15}$), supporting direct recognition of t1 adenosines by the silencing complex. In fact, no stronger down-regulation was observed for M1 7 mer than for 6 mer flanked by nonmatching bases other than A (Fig. 1F, solid green curve), suggesting that base-pairing between the m1 and t1 bases, if it occurs, does not contribute to targeting. Similarly, stronger down-regulation was observed for M8-A1 8 mer than for M8-M1 8 mer (Fig. 1F, cf. solid purple curve and purple triangles, $P < 10^{-13}$). Again, no stronger down-regulation was observed for M8-M1 8 mer than for M8 7 mer with nonmatching, non-A bases at position t1. Together, these observations strongly support the hypothesis that t1A residues adjacent to 6 mer or to M8 7 mer are recognized directly by a protein component of the silencing machinery in human cells, and that pairing to the m1 base, if it occurs, is of little or no consequence for targeting. This conclusion is consistent with recent structural studies of an Argonaute protein homolog in complex with dsRNA or an siRNA-like duplex, showing that the 5′ nucleotide of the guide RNA (corresponding to the m1 base in an miRNA:mRNA or siRNA:mRNA duplex) is not base paired (Ma et al. 2005; Parker et al. 2005). The predictions of widely used miRNA target prediction algorithms that reward WC matching at position 1 (e.g., John et al. [2004]) should therefore be improved by instead rewarding t1A independent of miRNA complementarity.

Stronger down-regulation of mRNAs with conserved seed matches

The widespread conservation of 3′ UTR seed matches since the divergence of rodents, carnivores, and primates (>50 million years ago [mya]) raises the issue of whether miRNA targets conserved over this time span commonly possess other determinants of miRNA targeting. To address this question, the distributions of LFCs for mRNAs containing exclusively nonconserved 3′ UTR extended seed matches to miR-1 or miR-124 were compared with those of mRNAs containing conserved 3′ UTR extended seed matches to these miRNAs (which will be greatly enriched for authentic conserved targets of these miRNAs). Notably, the mean nLFC for conserved extended seed matches was twice that seen for nonconserved extended seed matches for both miRNAs (Fig. 1C,E). This difference was significant ($P < 0.001$ for miR-124, $P < 0.01$ for miR-1, by rank sum test), when controlling for overall UTR conservation, seed match type and count, and initial mRNA expression level (Fig. 1C,E; Supplemental Table S1, with controls performed as illustrated in Supplemental Fig. S2). This observation suggests that authentic conserved miRNA targets contain additional targeting determinants that make them substantially more repressible by miRNA-programmed silencing complexes. An additional control for generic effects of 7 mer conservation was performed using data from siRNA studies (Jackson et al. 2003, 2006; Schwarz et al. 2006). Because the siRNAs used in the Jackson/Schwarz studies are unrelated in sequence to known endogenous mammalian miRNAs, any conservation of seed matches to these siRNAs is purely coincidental and unrelated to regulation by endogenous miRNAs. No significant difference in the distribution of LFC values was observed between mRNAs containing conserved rather than nonconserved extended seed matches to the transfected siRNAs, when expression, seed match type and count, and overall UTR conservation were controlled for as above (Fig. 1G; Supplemental

Table S1). These observations imply that the increased repression observed for mRNAs containing conserved miRNA seed matches results from selection to enhance miRNA-directed repression in conserved targets relative to other genes.

Inducible inhibition of endogenous miRNA expression in mouse embryonic fibroblasts

The analyses described above rely on systems in which miRNAs or siRNAs are transfected into cells in which these RNAs are not naturally expressed. Although supported by independent analyses of UTR sequence conservation, the results are therefore subject to any potential differences between the activities of exogenous and endogenously expressed miRNAs, e.g., resulting from differences in incorporation into silencing complexes, if such differences exist. Therefore, it was of interest to ascertain whether the targeting rules observed above, e.g., the differences between seed match types and between conserved and nonconserved seed matches, apply also to regulation by endogenous miRNAs.

To study the activities of endogenously expressed mammalian miRNAs, we developed a conditional Dicer knockout system. Following Drosha processing in the nucleus, ∼70 nt hairpin pre-miRNAs are exported to the cytoplasm for secondary processing to the mature ∼22 nt miRNA by the RNase III enzyme Dicer (Kim 2005). Vertebrates express only a single Dicer gene, *Dicer1*, which is essential for development in both the mouse and the zebrafish (Bernstein et al. 2003; Wienholds et al. 2003). All vertebrate miRNAs appear to require processing by the protein product of this gene. Mice homozygous for a conditional null allele of *Dicer1* were generated using a tamoxifen-inducible promoter driving Cre recombinase (Danielian et al. 1998; Hayashi and McMahon 2002) and a conditional LacZ reporter (Soriano 1999; Sup-

plemental Fig. S3). Cells were harvested from embryos at gestational day 16 and propagated in culture according to standard protocols to generate mouse embryonic fibroblasts (MEFs), which we refer to as conditional *Dicer* knockout (CDKO) MEFs.

Exposure of these cells to tamoxifen (ortho hydroxy tamoxifen; OHT) induces expression of Cre recombinase, resulting in a deletion that heritably inactivates the *Dicer1* locus. By staining the cells for LacZ, we established the minimum concentration and time required to induce Cre expression and inactivate the *Dicer1* locus in essentially all MEFs (Fig. 2A; Supplemental Fig. S4). Within 24-48 h post-induction of Cre, proliferation slowed (Fig. 2B). Visual inspection of the cells suggested minimal levels of apoptosis, and a modest level of apoptosis not substantially greater than for control cells was confirmed by Annexin V staining (Supplemental Fig. S5). In these respects, the *Dicer*-deficient MEFs bore some similarities to *Dicer*-deficient T cells, which were reported to have reduced proliferation but only modestly increased levels of apoptosis (Muljoet al. 2005).

By day 4 post-induction, Western analysis with Dicer antibodies detected an approximately threefold decrease in protein levels (Fig. 2C). Day 4 post-induction represents an average of ∼2 d post-inactivation of the *Dicer* locus (Supplemental Fig. S4). At this stage, total RNA was collected from the MEFs for microarray analysis. Untreated CDKO MEFs and MEFs derived from wild-type mice (untreated or subjected to OHT treatment) were used as controls, and each experiment was repeated twice.

Mature miRNAs were profiled in the MEFs using a spotted oligonucleotide miRNA microarray with standard miRNA probes present in quadruplicate. Using this array,

expression of 99 miRNAs was detected at more than two standard deviations (SD) above background in seven of the eight miRNA arrays (Supplemental Table S4). Among the most highly expressed miRNAs were members of the let-7 family, miR-1, miR-124, miR-15a, miR-175p, and several other miRNAs previously detected in mouse embryos (Thomson et al. 2004). Expression of several of the array-detected miRNAs was also confirmed by Northern analysis miRNA and siRNA targeting determinants (Supplemental Fig. S6). Spiked control RNA and second-channel reference RNAs were used to enable comparison of miRNA array data between control and knockout cells. Levels of most miRNAs decreased following Cre induction/*Dicer* knockout (Supplemental Fig. S6). The fold change in microarray hybridization intensity and in expression measured by Northern analysis were correlated, with the array intensity change consistently lower than the fold change measured by Northern (Fig. 2D). The expression of most miRNAs tested was reduced by approximately twofold by Northern, consistent with the about threefold reduction in Dicer protein levels and the notion that miRNAs have fairly long, but not infinite, half-lives. Variability in the fold changes of different miRNAs was observed, which could reflect differences in miRNA stability, in pre-miRNA processing efficiency in the presence of limiting amounts of Dicer protein, or perhaps changes in miRNA transcription or processing in response to reduced Dicer protein or miRNA levels.

Targeting rules inferred from derepression of mRNAs following Dicer knockout

The expression of mRNAs was profiled in control and CDKO MEFs using Affymetrix Mouse Genome 430_2 arrays. This CDKO system has certain advantages over transfection-based systems for studies of miRNA function, including the potential to study the activities of endogenous miRNAs expressed at natural levels. The CDKO system gen-

erates a modest and gradual ebbing of miRNA levels, as opposed to miRNA/siRNA transfection, which effectively floods the cell with a specific miRNA/siRNA species. Although it requires administration of tamoxifen, use of targeted gene knockout to reduce Dicer levels, rather than RNAi, has the advantages of permanently inactivating the *Dicer1* gene and avoiding addition of exogenous siRNAs, which could exert "off-target" effects like those seen in Figure 1F, complicating analysis of mRNA expression changes.

Analysis of miRNA effects on mRNAs following *Dicer* knockout is necessarily more complex than for miRNA/siRNA transfection experiments because loss of *Dicer* results in decreases in the levels of dozens of miRNAs at once. One straightforward approach uses the median LFC (where LFC is defined for *Dicer* knockout experiments as the base 2 log of hybridization intensity in treated CDKO cells over the intensity in control cells) over all mRNAs containing conserved extended seed matches to a particular miRNA, analyzing each miRNA independently. (In this analysis, untreated CDKO cells, and treated and untreated wild-type MEFs, served as controls; hybridization intensity averaged over these three types provided a control value for calculating LFC.) Applying this approach to the set of 99 miRNAs (representing 80 unique seed sequences) detected by miRNA array analysis yielded a distribution of median LFCs that was significantly shifted toward higher values than for control (random) sets of mRNAs or for mRNAs containing conserved extended seed matches to the miRNAs (representing 50 unique seeds) whose expression was not detected above background (Fig. 3A). (Median LFC was used in this analysis rather than mean because of its greater stability in the face of noise for the sometimes very small sets of conserved targets being analyzed.) This observation suggested that many of the changes in mRNA expression observed in this experiment resulted from derepression

of genes whose mRNA levels were specifically repressed by miRNA-programmed silencing complexes prior to knockout of *Dicer*. The derepression of mRNAs containing conserved seed matches to many expressed miRNAs following about twofold reduction in miRNA expression suggested that, at least in this system, many miRNAs are not expressed at saturating levels relative to their targets. mRNAs with seed matches to miRNAs not detected by microarray were shifted to an insignificant degree toward higher values relative to random mRNA sets of the same size (Fig. 3A). A list of the mRNAs whose expression changed significantly following *Dicer* knockout is provided in Supplemental Table S5.

Three previous studies have analyzed the effects of inhibition of miRNA processing enzymes on global mRNA expression, two using RNAi knockdown and one using targeted gene knockout. Rehwinkel and colleagues (2005) found that the set of mRNAs derepressed following RNAi knockdown of Drosha in *Drosophila* cells were enriched for miRNA targets predicted using the algorithm of Stark and colleagues (Brennecke et al. 2005; Stark et al. 2005), which is based on rules for targeting that (like TargetScanS) emphasize WC pairing to miRNA bases 2-8. Thus, the Rehwinkel study supported the idea that endogenous miRNAs commonly regulate their targets at the mRNA level through mechanisms involving seed match pairing. Recently, Schmitter and colleagues (2006) studied global changes in mRNA expression following RNAi knockdown of Dicer and Argonaute in cultured human cells. They observed up-regulation/derepression of overlapping sets of transcripts 2 and 6 d after knockdown of Dicer and 2 d after knockdown of Ago2, and again found enrichment for miRNA seed matches in the UTRs of derepressed mRNAs. Very modest effects were observed following knockdown of other Argonaute family genes. In the third study, Giraldez and colleagues (2006) used sophisticated gene knockout techniques to

generate "MZdicer" zebrafish embryos deficient in both maternal and zygotic Dicer activity. The set of mRNAs whose expression was significantly increased in MZdicer embryos relative to wild type were enriched for seed matches to miRNAs of the miR-430 family, the most abundantly expressed miRNA family during early zebrafish development, representing ∼50% of miRNAs cloned. These and related studies of MZdicer embryos convincingly demonstrated that miRNAs promote accelerated decay of targeted mRNAs *in vivo*.

The predominance of a single miRNA family in zebrafish embryos made this system suitable for assessing the effects of seed match type on mRNA regulation by endogenous miRNAs. Analyzing mRNAs containing different miR-430 seed match types, M8-A1 8 mer were associated with the strongest derepression, with a mean nLFC value almost twice that seen for M8 or A1 7 mer. The mean nLFC values for the two 7 mer types were similar to each other and higher than for 6 mer (Fig. 3B; Supplemental Table S6). Thus, the ordering of seed match types and the relative magnitudes of 8 mer versus 7 mer seed match effects paralleled those seen for transfected miRNAs/siRNAs (Fig. 1), indicating that the seed match hierarchy inferred from transfection data also holds for regulation by endogenous vertebrate miRNAs. For zebrafish miR-430, 6 mer had a higher nLFC value relative to 7 mer and 8 mer than in the mammalian miRNA/siRNA transfection experiments (Supplemental Tables S1, S2). The 6 mer nLFC value may be magnified by effects of other miR-430 superfamily miRNAs. Analogous seed match type comparisons were not attempted using the CDKO MEF data because most mRNAs contained a mixture of different seed match types, often to several different expressed miRNAs, so too few mRNAs containing only a single seed match type were available for effective analysis; the effect of the t1 position was not addressed in the CDKO MEF data for the same reason.

The repression of mRNAs containing conserved rather than nonconserved seed matches could be most effectively analyzed in the CDKO MEF data. In the zebrafish MZdicer data, the set of mRNAs containing conserved miR-430 seed matches was relatively small, and significant derepression relative to nonconserved seed matches was not observed (Supplemental Fig. S7). Seed match conservation is more difficult to assess in fish, as large differences in intergenic region sizes among the fishes yield less reliable genomic alignments, and classification based on seed match presence is limited by the relatively sparse 3′ UTR annotations available for orthologous fish genes. In the MEF CDKO data, far larger mRNA sets were available for this analysis. In these data, the mean nLFC for mRNAs containing extended seed matches conserved between human, mouse, rat, and dog (HMRD) to a set of 31 "responsive" miRNAs (see Supplemental Material) was ∼50% higher than that for mRNAs containing exclusively nonconserved seed matches ($P < 0.05$), controlling for overall UTR conservation, mRNA expression, and seed match count (Fig. 3C). This analysis, indicating that mRNAs containing conserved extended seed matches are preferentially repressed by endogenous miRNAs, further supports the idea that conserved miRNA targets possess additional targeting determinants that contribute to their repression by the miRNAs that naturally target them.

Fold change increases multiplicatively with seed match count for both endogenous miRNAs and exogenous miRNAs/siRNAs

Using luciferase or other reporter assays, increases in the magnitude of miRNA-directed repression have typically been observed when the number of 3′ UTR seed matches is increased (Doench and Sharp 2004; Vella et al. 2004; Pillai et al. 2005), but the quantitative relationship between seed match count and repression has not been established using large sets of targets. Grouping mRNAs based on the number

of extended seed matches to transfected miRNAs in the Lim datasets analyzed in Figure 1, mean nLFC increased approximately linearly as extended seed match count increased from one to three for both miRNAs (Fig. 4A). The dose-response relationship between extended seed match count and mRNA nLFC further supports the idea that seed matches are the primary determinant of mRNA down-regulation by miRNAs. Although the Lim (Lim et al. 2005), Jackson (Jackson et al. 2003, 2006), and Schwarz (Schwarz et al. 2006) experiments used identical protocols and concentrations of transfected RNAs, the magnitude of the mean nLFC per seed match for the two miRNAs was roughly twice that seen on average for the siRNAs (not shown), suggesting some degree of optimization of target and perhaps miRNA sequences for efficient repression.

For endogenous zebrafish miR-430, a roughly linear relationship was also seen between the mean nLFC values of mRNAs containing one to three extended seed matches in the MZdicer experiment (Fig. 4B). Here, as for the miRNA transfection data, too few mRNAs were available to extend the analysis beyond three seed matches. Because of the greater diversity of miRNAs affected, the CDKO MEF experiment allowed analysis of the effects of a larger range of seed match counts on mRNA repression. For the set of 31 "responsive" miRNAs used above, an approximately linear relationship was again observed between mean nLFC and the count of conserved extended seed matches (Fig. 4C). This relationship held for conserved extended seed match counts from one up to at least five, suggesting that miRNA regulation is tunable over a very broad range. The essentially linear relationship between seed match count and the logarithm of the fold change observed in Figure 4 indicates that each seed match contributes multiplicatively to fold change in mRNA level. Multiplicative effects could be explained if RISCs act independently and each has a

chance of interaction with a single effector site on the mRNA - such as the 5′ cap (Kiriakidou et al. 2007) - is required for RISC-mediated repression.

Evidence for A or U at position t9 as a targeting determinant

To search for additional targeting determinants, we analyzed the effects on mRNA down-regulation of nucleotides present at different target positions in the vicinity of seed matches. The Jackson/Schwarz siRNA transfection data were most suitable for this analysis because of the large number of independent siRNAs and array measurements. Although modest increases in down-regulation were associated with the presence of adenosine and/or uridine at a few other positions (not shown), the most pronounced effect was observed for the presence of A or U at position t9. Those mRNAs that had a t9W base (using the abbreviation W = A or U) were down-regulated to a greater degree following siRNA transfection than those with a t9S (S = C or G) residue (Fig. 5A). This effect was pronounced for M8 7 mer ($P < 10^{-6}$) and 8 mer ($P < 10^{-2}$) seed matches, with a marginal effect observed for A1 7 mer (not shown). The effect remained highly significant whether controlling for UTR CG content (as in Fig. 5) or not. The increased repression of extended seed matches containing t9W was observed independent of whether the base m9 of the siRNA was a match to t9 or not (Fig. 5B). No significant effect of t9 matching was observed, though the t9W match set for 8 mer in particular was quite small (n = 141), limiting statistical power to detect any effect that might exist. Taken together, these observations suggest that the presence of a t9W base adjacent to an extended seed match contributes to typical seed match targeting interactions, independent of pairing to m9.

The targeting role of t9 inferred from siRNA transfection data was corroborated by analyses of seed matches to miRNAs. Examining the composition of the t9 posi-

tion for seed matches to a large set of conserved miRNAs in mammalian UTRs, we observed an increased frequency of t9W residues adjacent to conserved seed matches relative to control sets of nonconserved seed matches in UTRs matched for UTR CG content (Fig. 5C). Signal:noise values for this miRNA set, calculated with control oligonucleotides matched for both count and CG content, were significantly higher for t9W compared with t9S seed matches (Fig. 5D). Consistent with the siRNA analyses, the difference appeared independent of the base at position m9. These observations extend previous observations of increased conservation of t9A residues, independent of complementarity to miRNA base m9 (Lewis et al. 2005), and support a role for t9W in miRNA targeting in vivo.

Increased conservation and AU content flanking siRNA seed matches associated with increased mRNA repression

Increased sequence conservation across mammals is observed in the vicinity of conserved miRNA seed matches relative to those that are not conserved, even when overall UTR conservation is controlled for (Fig. 6A). The increase in conservation extends to 50 bases 3′ and 5′ of the seed match and beyond; similar patterns of increased local conservation are associated with other conserved UTR motifs (not shown; Lewis et al. 2005). One possible explanation is that the sequence context flanking authentic conserved target sites is enriched for feature - e.g., protein binding sites or RNA structural properties - that, directly or indirectly, enhance the effectiveness of miRNA targeting. To explore these issues, we returned to the siRNA data and compared siRNA-directed mRNA repression between mRNAs having different levels of sequence conservation in the 50 bases 5′ and 3′ of the seed match, in sets matched for overall UTR conservation, expression level, seed match type, and local and global AU composition (Supplemental Fig. S9). Strikingly, substantially stronger mRNA

repression was observed for siRNA seed matches with high conservation in the 50 bases upstream of the seed match relative to seed matches with low conservation in this region (mean nLFC = 0.16 and 0.11, respectively, $P < 10^{-4}$ by rank sum test; Fig. 6C). A similar effect of conservation in the downstream 50 bases was observed (mean nLFC = 0.16 versus 0.12, $P < 0.05$; Fig. 6C). These results were not affected by whether or not the siRNA seed match itself was conserved (not shown).

In vertebrate genomes, AU-rich sequences have higher levels of average sequence conservation than CG-rich sequences, at least in part because of the high mutation rate of CpG dinucleotides (Hwang and Green 2004), so it was important to determine whether the increased repression associated with conserved flanking regions resulted from an effect of base composition. The 30-50 bases just upstream of and downstream from conserved miRNA seed matches are indeed biased toward higher AU composition (Fig. 6B), suggesting that local AU composition itself might contribute to targeting. However, the effect of sequence conservation on mRNA repression was significant whether AU content was controlled for (as in Fig. 6C) or not. Conversely, significantly increased repression was observed for siRNA seed matches flanked by high AU content in the 50 bases either $3'$ or $5'$ of the seed match relative to those with low AU content in these regions (not shown). When this analysis was controlled for the effects of sequence conservation, overall UTR AU content, expression level, and seed match type, seed matches with high AU content in the $3'$ 50 bases had significantly increased down-regulation relative to those with low AU content in this region (mean nLFC = 0.17 and 0.10, respectively, $P < 10^{-4}$ by rank sum test), but the effects of $5'$ AU content were no longer significant (Fig. 6D). Thus, conservation immediately $5'$ and $3'$ and AU content $3'$ of seed matches are independently associated with increased mRNA down-regulation by siRNAs, and $5'$ AU content may also

enhance down-regulation. Choice of 50 base pairs (bp) as the size of the region to analyze was based on the distributions shown in Figure 6, A and B; however, the magnitude of the effects shown in Figure 6, C and D, were little changed when the size of the analyzed region was expanded or reduced by 20 bp.

In terms of the magnitude of nLFC change, each of these variables contributed to targeting to at least the same extent as the identity of the t9 base (e.g., Fig. 5B, compare mean nLFC value differences). The conservation of more distal 50-bp windows (e.g., bases 151-200 downstream from the seed match) was associated with increased repression when the controls on conservation in other regions were relaxed, but disappeared when either overall UTR conservation or conservation in the 50 bases immediately 3′ of the seed match were controlled for (not shown). This observation suggests that effects for such distal windows observed in the uncontrolled analysis derive from the (fairly strong) positive correlation between conservation in nearby UTR regions and that the proximal 50 bp is of central importance. Similar results were obtained for AU content (not shown), again supporting the importance of the regions immediately adjacent to the seed match.

There are at least two plausible ways in which high AU content might increase effectiveness of adjacent seed matches. AU-rich sequences could be recognized directly by a component of the RISC or an auxiliary activating factor; a number of protein families are known that have affinity for A-rich, U-rich, or AU-rich RNA sequences (Barreau et al. 2005), and functional connections between AU-rich binding factors and miRNA regulation have been reported (Bhattacharyya et al. 2006; Vasudevan and Steitz 2007). Alternatively, AU-richness might enhance targeting by reducing the tendency for formation of stable RNA secondary structures that could interfere

with RISC binding. Previously, it has been reported that predicted local folding of the mRNA in the vicinity of seed matches is a negative predictor of miRNA targeting (e.g., Robins et al. 2005). Consistently, we have observed that seed matches in regions of lower predicted thermodynamic stability using standard algorithms are associated with increased mRNA-level repression, but we have found that this effect disappears when the AU content of the region is controlled for (not shown). Thus, the effect of AU content we observe may contribute to targeting by reducing the potential for inhibitory mRNA structures, but other effects of AU-richness are also consistent with the data.

Since siRNAs are not naturally expressed, siRNA seed matches tend to be distributed essentially randomly in UTRs and do not experience selection related to targeting. The magnitude of the effect of local conservation on siRNA seed match efficacy (Fig. 6C) was similar in magnitude when siRNA seed matches falling within 100 bp of conserved miRNA seed matches were included (as in Fig. 6C) or excluded (not shown). Therefore, the enhanced repression observed for siRNA seed matches that occur in regions of high local conservation (Fig. 6C) is likely to represent a side effect of mRNA features that are conserved for reasons unrelated to targeting by the RISC, such as RNA binding sites for factors involved in other aspects of mRNA biology (mRNA processing, stability, localization, translation, etc.). For example, presence of proteins bound to sites nearby the seed match might increase accessibility to RISC by interfering with formation of local RNA secondary structures. Alternatively, the long coevolution of RISC and non-RISC factors binding nearby in mRNAs may have engendered more direct interactions, with common mRNA binding factors functioning to facilitate or stabilize RISC binding to nearby seed matches. Two very recent studies observed that, with the exception of the first ∼20 bases of the mRNA,

the density of conserved seed matches increases as proximity to the stop codon or poly-A tail increases, (Gaidatzis et al. 2007; Majoros and Ohler 2007); however, signal:noise in these regions was not assessed. Another recent study reported synergy between nearby seed matches located 13-35 bases apart, providing support for the idea that RISC activity is modulated by the presence of proteins or complexes in nearby flanking regions (Saetrom et al. 2007). Given the complexities of conservation effects on miRNA target analysis, including the phenomenon that mRNAs with lower overall UTR conservation have substantially higher signal:noise values for conserved miRNAs (Lewis et al. 2005), analyses of the effects of local conservation and AU content on miRNA targeting were deferred pending availability of additional relevant data. Further investigation is clearly needed to understand the mechanisms underlying these phenomena.

Perspectives and applications to target/off-target prediction

Here, we have described specific rules for miRNA/siRNA targeting, including a hierarchy of seed match types, the multiplicative effects of multiple seed matches, and targeting determinants outside of the seed match, including t9W and local conservation and AU content effects. All of these rules and determinants were supported for exogenous siRNAs and/or miRNAs by direct effects on mRNA levels, and for endogenous miRNAs through direct effects and/or comparative genomic data. Thus, in the available data these rules and determinants appear to be applicable to targeting by exogenous siRNA/miRNAs and endogenous miRNAs.

Current miRNA target predictions have relied very heavily on seed match conservation, ignoring potential species-specific miRNA targeting. However, we observe that nonconserved 8 mer seed matches on average exhibit stronger repression than

conserved 7 mer (Supplemental Fig. S8). For studies of miRNA function, it would be extremely useful to be able to predict which mRNAs will experience the strongest repression to facilitate choice of targets for in-depth characterization, and similar considerations apply to the design and interpretation of experiments involving siRNAs. The new rules and determinants identified here can be combined to produce an expected nLFC score for a seed match by summing the mean nLFC of the seed match type (Fig. 1F) plus the residual contribution to mean nLFC of the t9 base (Fig. 5A), and of flanking AU content and conservation (Fig. 6). Following the results of Figure 4, if multiple seed matches are present, their scores are added (scoring details are given in Supplemental Material). Because this score can be used to rank potential siRNA off-target effects, and to generate an integrated ranking of conserved and nonconserved miRNA targets, we call it the TargetRank score. Application of TargetRank scoring to sets of mRNAs, each with a single 7 mer seed match to transfected siRNAs (with parameters derived from a held-out set of siRNA transfections) yielded a dramatic separation between the LFC distributions of the bottom 20% and top 20% of TargetRanked mRNAs in the test set (Fig. 7A), with mean nLFC increasing from 0.07 to 0.26, and the fraction of significantly down-regulated mRNAs increasing from 5% to 20% (Supplemental Table S7). Because all of the mRNAs in this analysis contained single 7 mer seed matches, the separation of the two distributions results from the additional determinants identified in this study (t9W, flanking conservation, and AU content), demonstrating their combined utility for siRNA off-target prediction.

Applying the same scoring system to expression data for Th1 thymocytes from mir-155 knockout mouse versus wild-type cells (Rodriguez et al. 2007) also yielded a strong degree of separation. When scoring mRNAs containing exclusively nonconserved single 7 mer seed matches, mean nLFC increased from 0.01 for the bottom

20% of TargetRanked mRNAs (NS relative to no-seed-match mRNAs) to 0.09 for the top 20% of mRNAs ($P < 0.01$ relative to no-seed-match or bottom 20% mRNAs) (Fig. 7B; Supplemental Table S7), demonstrating the applicability of these additional determinants to regulation by endogenous miRNAs, and suggesting an approach for identification of important species-specific miRNA targets. In practice, scoring of 6 mer, 7 mer, and 8 mer seed matches and messages containing multiple seed matches yields a broader range of TargetRank scores and a correspondingly greater separation between the distributions of higher and lower ranked genes. Grouping siRNA and nonconserved miR-155 targets into five bins of TargetRank scores demonstrated a strong and approximately log-linear relationship between mRNA down-regulation and TargetRank score (Fig. 7C,D). The relative ranking given by TargetRank is probably more useful than the score itself, since the overall magnitude of miRNA- or siRNA-associated repression will vary in different systems, as seen above.

Unlike purely conservation-based methods, TargetRank scoring of the expressed mRNAs in a cell type yields an integrated ranking of conserved and nonconserved targets, and should therefore be particularly helpful in identifying important species- or clade-specific miRNA targeting relationships. These results should also aid in interpretation of RNAi phenotypes and in prediction of the miRNA targeting effects of mutations and polymorphisms in human genes.

Materials and Methods

Conditional Dicer knockout mice and MEFs

Male mice carrying one copy of the pCAGGCre-ER allele (Hayashi and McMahon 2002) and one Dicer floxed allele (Harfe et al. 2005) were crossed to Dicer floxed/floxed females harboring also a LacZ reporter (R26R) for detection of Cre activity (Soriano 1999). Timed-pregnant females were sacrificed at embryonic day 16 and embryos were dissected and dissociated to generate mouse embryonic fibroblast (MEF) primary culture (following Abbondanzo et al. [1993]). After 72 h of incubation, cells were frozen in aliquots. Mice were housed and handled in accordance with protocols approved by the Institutional Animal Care and Use Committee of Harvard Medical School.

Cell culture and treatments

MEFs were thawed prior to experiments, grown in DMEM supplemented with 10% FCS, penicillin/streptomycin and glutamine, split once, and induced for loss of functional Dicer by addition of 4-orthohydroxy Tamoxifen (0.5 mM; OHT; Sigma). Following 4 d (and daily change of media and drug), total RNA and total protein were extracted. Control MEFs derived from wild-type mice were subjected to the same treatments.

RNA extraction

Total RNA was extracted using TRIzol reagent (Sigma), and RNA quality was measured using an Agilent Bioanalyzer.

MEF miRNA microarray analysis

MiRNA microarrays were printed using a Cartesian PixSys 5500 Arrayer on epoxy slides (Corning) using Ambions miRvana amine-modified DNA oligonucleotide probe set (version 1564V1) and scanned using an Axon Scanner GenPix 4000 (see Supplemental Material for further details).

Northern analysis

Thirty micrograms of total RNA was separated in 15% TBE-UREA gels (Bio-Rad), transferred to a GeneScreen Plus membrane (Perkin Elmer) using semidry electroblot apparatus (Owl) in 13 TBE (90 mM Tris-base, 2 mM EDTA, 90 mM Boric acid) at 350 mA for 35 min. The membrane was then UV cross-linked at 1000 mJ (Stratagene) and heated for 2 h at 80°C. Prehybridization and hybridization were carried out in PerfectHyb Plus Hybridization Buffer (Sigma) supplemented with Salmon Sperm DNA (20 mg/mL) for 2 and 16 h, respectively, at 42°C, with a radiolabeled probe added to the latter. Washes were done in 23 SSC + 0.2% SDS (twice), then 13 SSC + 0.2% SDS (once) for 5 min at 50°C. Membranes were exposed to a PhosphorImager cassette for 3 d, then scanned (PhosphorImager, Molecular Dynamics, 445 SI) and quantitated (ImageQuant, Molecular Dynamics).

MEF mRNA microarrays

Affymetrix GeneChip Mouse Genome 430_2 Array labeling, hybridization, and scanning were performed according to the manufacturers instructions. The data discussed in this publication have been deposited in NCBIs Gene Expression Omnibus (GEO, `http://www.ncbi.nlm.nih.gov/geo/`) and are accessible through GEO Series accession number GSE6046. To map probes on the Affymetrix Mouse 430_2 array to Ref-

seq transcripts, we used custom CDF file MM430_MM_REFSEQ_6, downloaded from `http://brainarray.mbni.med.umich.edu/Brainarray/Database/CustomCDF`, the custom CDF project site. Refseq transcript expression levels were then calculated using GCRMA (GCRMA package, Bioconductor in R environment) using default settings. Genes with normalized $\log_2$ intensity below 3 were excluded from analysis.

3′ UTR datasets

Genome coordinates for 3′ UTRs were obtained using Refseq annotations and alignments of hg17 with 16 other vertebrate genomes, available from UCSC (`http://hgdownload.cse.ucsc.edu`) for human (hg17, May 2004), mouse (mm5, May 2004), and zebrafish (danRer3, May 2005). Only Refseq transcripts mapping uniquely to the genome were considered. Annotated 3′ UTRs shorter than 50 nt were excluded.

miRNA seed match counts and conservation

The 3′ UTR sequences were searched for nonoverlapping seed matches to relevant miRNAs or siRNAs of the types shown in Figure 1A. For human and mouse analyses, multiple alignments were obtained for each 3′ UTR by extracting the relevant region from genomic alignments available in multiple alignment format (MAF) from UCSC (`http://hgdownload.cse.ucsc.edu`, hg17 alignments of 17 vertebrate genomes). Seed matches with perfect conservation in aligned human, mouse, rat, and dog (HMRD) UTRs were labeled conserved.

miRNA and siRNA transfection datasets

Microarray expression data for miR-1 and miR-124 HeLa transfection experiments (Lim et al. 2005) were obtained from GEO accession GSE2075. Array platform information for these experiments was obtained from GEO accession GPL1749. Probes

were mapped to the human genome using BLAST, and subsequently mapped to Refseq annotated 3′ UTRs using Refseq genomic mapping files available from UCSC (`http://hgdownload.cse.ucsc.edu/goldenPath/hg17/database/`). Microarray expression data for siRNA HeLa transfection experiments (Jackson et al. 2003) were obtained from `http://www.rii.com/publications/2003/nbt831.html` and from GEO accession GSE5814 (Jackson et al. 2003, 2006; Schwarz et al. 2006) and GSE5291 (Jackson et al. 2003, 2006; Schwarz et al. 2006). Only values with Refseq IDs were used for this analysis. To remove poorly expressed genes, we excluded genes with $\log_2$ intensity < -4.0 for both datasets. The analyses reported here are based on 24-h data where repression was typically stronger.

Zebrafish embryo Dicer knockout datasets

Microarray expression data for zebrafish wild-type and MZdicer mutant embryos (Giraldez et al. 2006) were obtained from GEO (accession GSE4201). Probe information for the Affymetrix GeneChip Zebrafish Genome Array was also obtained from GEO (accession GPL1319). Probes were mapped to Refseqs using genomic mapping information for zebrafish Affymetrix Exemplar sequences from the UCSC annotation database. Only probes with a present (P) call were used for analysis.

bic/mir-155 knockout datasets

Microarray expression data for mouse wild-type and miR-155 deficient Th1 cells (Rodriguez et al. 2007) were obtained from ArrayExpress (accession E-TABM-232). Only probes mapping to mouse Refseqs were used for analysis.

Log fold change (LFC)

The mean normalized log fold change value for miRNA/siRNA transfection experiments was defined as the difference between the mean LFC for mRNAs lacking seed matches to the transfected miRNA/siRNA and the mean LFC for mRNAs containing seed matches of the given type (median nLFC was defined analogously). For Dicer miRNA knockout experiments, the nLFC was defined as the difference between the mean LFC for mRNAs containing seed matches to the relevant miRNAs (e.g., miR-430) and the mean LFC for mRNAs lacking seed matches to any relevant miRNA. The variability of the nLFC value due to experimental noise was estimated for 12 effective siRNAs where duplicate array data were available (MAPK14-M1, -M2as, -M4as, -M5as, -M6as, -M15, -M18, MPHOSPH12692, PRKCE-1295, SOS11582as, VHL-2651as, VHL-2652, where as indicates the strand antisense to the targeted mRNA). The nLFCs for expressed mRNAs containing one or more extended seed matches to the relevant siRNA were calculated for each array. The average Pearson correlation between nLFC values from duplicate array pairs was 0.83. The Pearson correlation among mean nLFC values across the duplicate array pairs was 0.91. These data indicate that while there is some variation among nLFCs for individual mRNAs, the mean nLFC is highly reproducible.

Statistical analyses

All test statistics were calculated using R (`http://www.r-project.org`). The Wilcoxon rank sum test was chosen over the t-test because it does not assume normality of the underlying distributions, and because it is more intuitive and familiar than non-parametric alternatives such as the KolmogorovSmirnov (KS) test. t-tests and KS tests using these data gave generally similar results. A P-value cutoff of 0.05 was

used for all analyses.

Software Availability

A TargetRank Web server is available at `http://genes.mit.edu/targetrank/`.

Acknowledgments

We thank R. Friedman and P.A. Sharp for helpful discussions and comments on the manuscript. Several of the analyses in this work relied on data published by scientists at Rosetta Inpharmatics and at the Schier, Turner, and Bradley laboratories, for which we are grateful. This work was supported by fellowships from NSERC (C.B.N.), and from the Knut and Alice Wallenberg Foundation (R.S.), and by a Searle Scholar Award and a grant from the NIH (C.B.B.).

Received July 26, 2007; accepted August 8, 2007.

References

Abbondanzo, S.J., Gadi, I., and Stewart, C.L. 1993. Derivation of embryonic stem cell lines. Methods Enzymol. 225: 803-823.

Ambros, V. 2004. The functions of animal microRNAs. Nature 431: 350-355.

Bagga, S., Bracht, J., Hunter, S., Massirer, K., Holtz, J., Eachus, R., and Pasquinelli, A.E. 2005. Regulation by *let-7* and *lin-4* miRNAs results in target mRNA degradation. Cell 122: 553-563.

Barreau, C., Paillard, L., and Osborne, H.B. 2005. AU-rich elements and associated factors: Are there unifying principles? Nucleic Acids Res. 33: 7138-7150.

Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. Cell 116: 281-297.

Bernstein, E., Kim, S.Y., Carmell, M.A., Murchison, E.P., Alcorn, H., Li, M.Z., Mills, A.A., Elledge, S.J., Anderson, K.V., and Hannon, G.J. 2003. Dicer is essential for mouse development. Nat. Genet. 35: 215-217.

Bhattacharyya, S.N., Habermacher, R., Martine, U., Closs, E.I., and Filipowicz, W. 2006. Relief of microRNA-mediated translational repression in human cells subjected to stress. Cell 125: 1111-1124.

Brennecke, J., Stark, A., Russell, R.B., and Cohen, S.M. 2005. Principles of microRNA-target recognition. PLoS Biol. 3: e85.

Danielian, P.S., Muccino, D., Rowitch, D.H., Michael, S.K., and McMahon, A.P. 1998. Modification of gene activity in mouse embryos in utero by a tamoxifen-inducible form of Cre recombinase. Curr. Biol. 8: 1323-1326.

Doench, J.G. and Sharp, P.A. 2004. Specificity of microRNA target selection in translational repression. Genes & Dev. 18: 504-511.

Doench, J.G., Petersen, C.P., and Sharp, P.A. 2003. siRNAs can function as miRNAs. Genes & Dev. 17: 438-442.

Farh, K.K., Grimson, A., Jan, C., Lewis, B.P., Johnston, W.K., Lim, L.P., Burge, C.B., and Bartel, D.P. 2005. The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. Science 310: 1817-1821.

Filipowicz, W., Jaskiewicz, L., Kolb, F.A., and Pillai, R.S. 2005. Post-transcriptional gene silencing by siRNAs and miRNAs. Curr. Opin. Struct. Biol. 15: 331-341.

Gaidatzis, D., van Nimwegen, E., Hausser, J., and Zavolan, M. 2007. Inference of miRNA targets using evolutionary conservation and pathway analysis. BMC Bioinformatics 8: 69.

Giraldez, A.J., Mishima, Y., Rihel, J., Grocock, R.J., Van Dongen, S., Inoue, K., Enright, A.J., and Schier, A.F. 2006. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science 312: 75-79.

Grun, D., Wang, Y.L., Langenberger, D., Gunsalus, K.C., and Rajewsky, N. 2005. microRNA target predictions across seven Drosophila species and comparison to mammalian targets. PLoS Comput. Biol. 1: e13.

Harfe, B.D., McManus, M.T., Mansfield, J.H., Hornstein, E., and Tabin, C.J. 2005. The RNaseIII enzyme Dicer is required for morphogenesis but not patterning of the vertebrate limb. Proc. Natl. Acad. Sci. 102: 10898-10903.

Hayashi, S. and McMahon, A.P. 2002. Efficient recombination in diverse tissues by a tamoxifen-inducible form of Cre: A tool for miRNA and siRNA targeting determinants temporally regulated gene activation/inactivation in the mouse. Dev. Biol. 244: 305-318.

Hutvagner, G. and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. Science 297: 2056-2060.

Hwang, D.G. and Green, P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc. Natl. Acad. Sci. 101: 13994-14001.

Jackson, A.L., Bartz, S.R., Schelter, J., Kobayashi, S.V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P.S. 2003. Expression profiling reveals off-target gene regulation by RNAi. Nat. Biotechnol. 21: 635-637.

Jackson, A.L., Burchard, J., Leake, D., Reynolds, A., Schelter, J., Guo, J., Johnson, J.M., Lim, L., Karpilow, J., Nichols, K., et al. 2006. Position-specific chemical modification of siRNAs reduces "off-target" transcript silencing. RNA 12: 1197-1205.

John, B., Enright, A.J., Aravin, A., Tuschl, T., Sander, C., and Marks, D.S. 2004. Human MicroRNA targets. PLoS Biol. 2: e363.

Jones-Rhoades, M.W., Bartel, D.P., and Bartel, B. 2006. MicroRNAs and their regulatory roles in plants. Annu. Rev. Plant Biol. 57: 19-53.

Kim, V.N. 2005. MicroRNA biogenesis: Coordinated cropping and dicing. Nat. Rev. Mol. Cell Biol. 6: 376-385.

Kiriakidou, M., Tan, G.S., Lamprinaki, S., De Planell-Saguer, M., Nelson, P.T., and Mourelatos, Z. 2007. An mRNA m(7)G cap binding-like motif within human Ago2 represses translation. Cell 129: 1141-1151.

Lai, E.C. 2002. MicroRNAs are complementary to 3′ UTR sequence motifs that mediate negative post-transcriptional regulation. Nat. Genet. 30: 363-364.

Lee, R.C., Feinbaum, R.L., and Ambros, V. 1993. The C. elegans heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell 75: 843-854.

Lewis, B.P., Shih, I.H., Jones-Rhoades, M.W., Bartel, D.P., and Burge, C.B. 2003. Prediction of mammalian microRNA targets. Cell 115: 787-798.

Lewis, B.P., Burge, C.B., and Bartel, D.P. 2005. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 120: 15-20.

Lim, L.P., Lau, N.C., Garrett-Engele, P., Grimson, A., Schelter, J.M., Castle, J., Bartel, D.P., Linsley, P.S., and Johnson, J.M. 2005. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 433: 769-773.

Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. 2002. Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA. Science 297: 2053-2056.

Ma, J.B., Yuan, Y.R., Meister, G., Pei, Y., Tuschl, T., and Patel, D.J. 2005. Structural basis for 5′-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature 434: 666-670.

Majoros, W.H. and Ohler, U. 2007. Spatial preferences of microRNA targets in 3′ untranslated regions. BMC Genomics 8: 152.

Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. 2004. Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. Mol. Cell 15: 185-197.

Muljo, S.A., Ansel, K.M., Kanellopoulou, C., Livingston, D.M., Rao, A., and Rajewsky, K. 2005. Aberrant T cell differentiation in the absence of Dicer. J. Exp. Med. 202: 261-269.

Parker, J.S., Roe, S.M., and Barford, D. 2005. Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. Nature 434: 663-666.

Pillai, R.S., Bhattacharyya, S.N., Artus, C.G., Zoller, T., Cougot, N., Basyuk, E., Bertrand, E., and Filipowicz, W. 2005. Inhibition of translational initiation by Let-7 MicroRNA in human cells. Science 309: 1573-1576.

Rehwinkel, J., Behm-Ansmant, I., Gatfield, D., and Izaurralde, E. 2005. A crucial role for GW182 and the DCP1:DCP2 decapping complex in miRNA-mediated gene silencing. RNA 11: 1640-1647.

Robins, H., Li, Y., and Padgett, R.W. 2005. Incorporating structure to predict microRNA targets. Proc. Natl. Acad. Sci. 102: 4006-4009.

Rodriguez, A., Vigorito, E., Clare, S., Warren, M.V., Couttet, P., Soond, D.R., van Dongen, S., Grocock, R.J., Das, P.P., Miska, E.A., et al. 2007. Requirement of bic/microRNA-155 for normal immune function. Science 316: 608-611.

Saetrom, P., Heale, B.S., Snove Jr, O., Aagaard, L., Alluin, J., and Rossi, J.J. 2007. Distance constraints between microRNA target sites dictate efficacy and cooperativity. Nucleic Acids Res. 35: 2333-2342.

Schmitter, D., Filkowski, J., Sewer, A., Pillai, R.S., Oakeley, E.J., Zavolan, M., Svoboda, P., and Filipowicz, W. 2006. Effects of Dicer and Argonaute down-regulation on mRNA levels in human HEK293 cells. Nucleic Acids Res. 34: 4801-4815.

Schwarz, D.S., Ding, H., Kennington, L., Moore, J.T., Schelter, J., Burchard, J., Linsley, P.S., Aronin, N., Xu, Z., and Zamore, P.D. 2006. Designing siRNA that distinguish between genes that differ by a single nucleotide. PLoS Genet. 2: e140.

Soriano, P. 1999. Generalized lacZ expression with the ROSA26 Cre reporter strain. Nat. Genet. 21: 70-71.

Stark, A., Brennecke, J., Bushati, N., Russell, R.B., and Cohen, S.M. 2005. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3′ UTR evolution. Cell 123: 1133-1146.

Thomson, J.M., Parker, J., Perou, C.M., and Hammond, S.M. 2004. A custom microarray platform for analysis of microRNA gene expression. Nat. Methods 1: 47-53.

Valencia-Sanchez, M.A., Liu, J., Hannon, G.J., and Parker, R. 2006. Control of translation and mRNA degradation by miRNAs and siRNAs. Genes & Dev. 20: 515-524.

Vasudevan, S. and Steitz, J.A. 2007. AU-rich-element-mediated upregulation of translation by FXR1 and Argonaute 2. Cell 128: 1105-1118.

Vella, M.C., Choi, E.Y., Lin, S.Y., Reinert, K., and Slack, F.J. 2004. The C. elegans microRNA let-7 binds to imperfect let-7 complementary sites from the *lin-41* 3′UTR. Genes & Dev. 18: 132-137.

Wienholds, E., Koudijs, M.J., van Eeden, F.J., Cuppen, E., and Plasterk, R.H. 2003. The microRNA-producing enzyme Dicer1 is essential for zebrafish development. Nat. Genet. 35: 217-218.

Wightman, B., Ha, I., and Ruvkun, G. 1993. Post-transcriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in C. elegans. Cell 75: 855-862.

Wu, L., Fan, J., and Belasco, J.G. 2006. MicroRNAs direct rapid deadenylation of mRNA. Proc. Natl. Acad. Sci. 103: 4034-4039.

Yekta, S., Shih, I.H., and Bartel, D.P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. Science 304: 594-596.

Zamore, P.D. and Haley, B. 2005. Ribo-gnome: The big world of small RNAs. Science 309: 1519-1524.

Zeng, Y., Wagner, E.J., and Cullen, B.R. 2002. Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol. Cell 9: 1327-1333.

Zeng, Y., Yi, R., and Cullen, B.R. 2003. MicroRNAs and small interfering RNAs can inhibit mRNA expression by similar mechanisms. Proc. Natl. Acad. Sci. 100: 9779-9784.

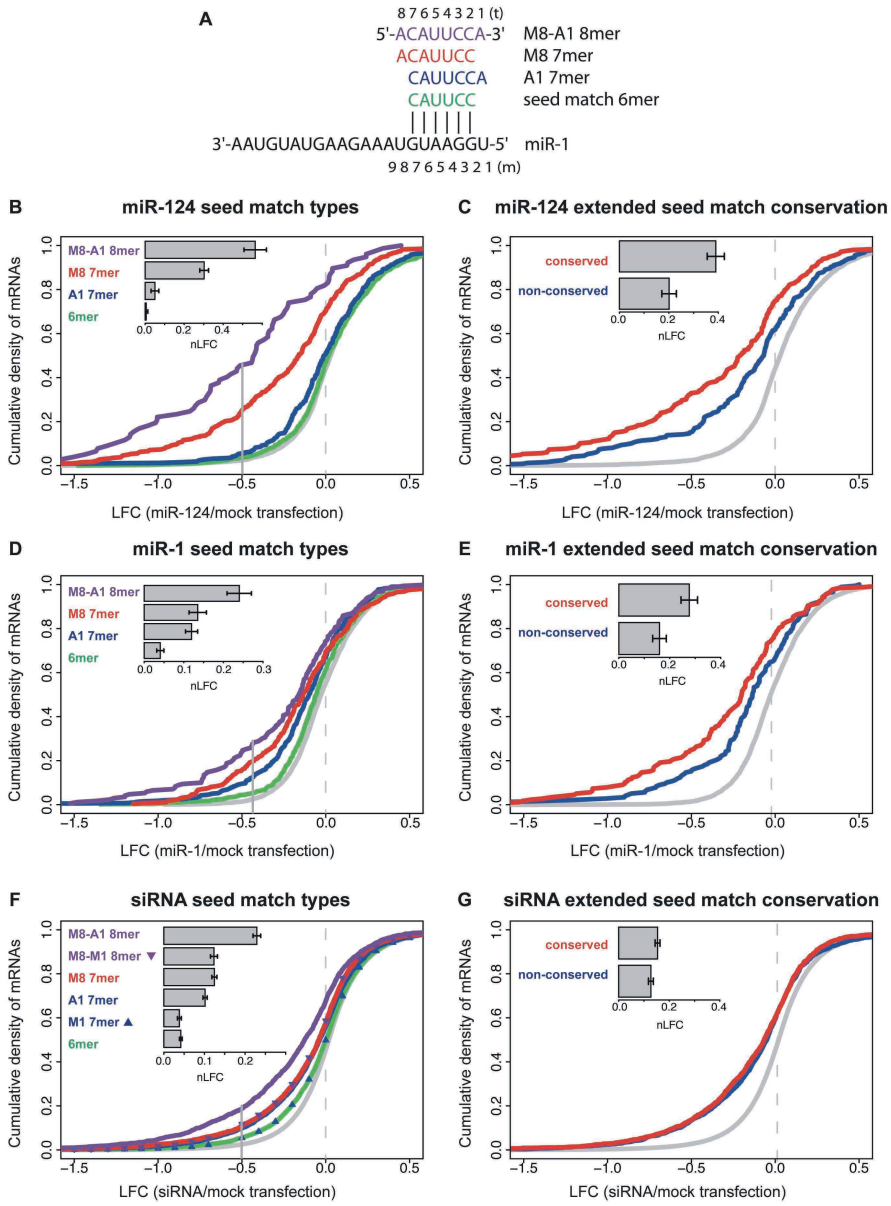

Figure 1: Effects of seed match type and conservation on mRNA repression for miRNAs and siRNAs. (A) Seed match types and numbering system, illustrated for miR-1. Positions in the miRNA are numbered 5′-3′. (Seed match 6 mer) WC inverse complement of miRNA bases 2-7; (A1) presence of adenosine opposite miRNA base 1; (M8) WC match to miRNA base8. (B) CDFs (cumulative distribution functions) of LFCs ($\log_2$ fold changes) for mRNAs containing indicated miR-124 seed match types (colored lines and labels) or no miR-124 seed matches (gray line) following transfection of miR-124. (Solid vertical gray line) The LFC above which 97.5% of the no-seed-match mRNA set falls. (Inset bar plot) nLFC (normalized $\log_2$ fold change) values for each seed match type with error bars indicating standard error. Data for panels B-E are from Lim et al. (2005). Plots include only mRNAs containing exactly one miR-124 seed match, and thus the seed match type sets are mutually exclusive. The distribution of mRNA expression values did not differ significantly between seed match type sets ($P > 0.05$ by rank sum test). All seed match types except the 6 mer have distributions significantly different from the no-seed-match class ($P < 0.005$ by rank sum test). (C) CDFs of LFCs for mutually exclusive mRNA sets containing conserved (red) or nonconserved (blue) extended seed matches to miR-124, or no seed matches (gray); the conserved and nonconserved sets are significantly different ($P < 0.001$ by rank sum test). The "nonconserved" mRNA set contains exclusively nonconserved seed matches; the "conserved" mRNAs may also contain nonconserved seed matches. The nonconserved set was sampled to match the conserved set in seed match type and count, overall UTR conservation, and initial mRNA expression level (Supplemental Fig. S2). (D) Same as B for miR-1. All seed match type classes are significantly different from the no-seed-match class ($P < 10^{-4}$ by rank sum test). (E) Same as C for miR-1. The CDFs of conserved and nonconserved mRNA sets are significantly different ($P < 0.01$ by rank sum test). (F) Same as B for pooled set of 33 "effective" siRNAs that begin with non-U bases (Supplemental Material). Additional seed match classes containing M1 are shown (triangles). All seed match types have distributions significantly different from the no-seed-match class ($P < 10^{-14}$ by rank sum test). (G) Same analysis and controls as C for pooled set of siRNAs. The CDFs of conserved and nonconserved mRNA sets are not significantly different ($P > 0.05$ by rank sum test). See Supplemental Table S1 for additional statistics.

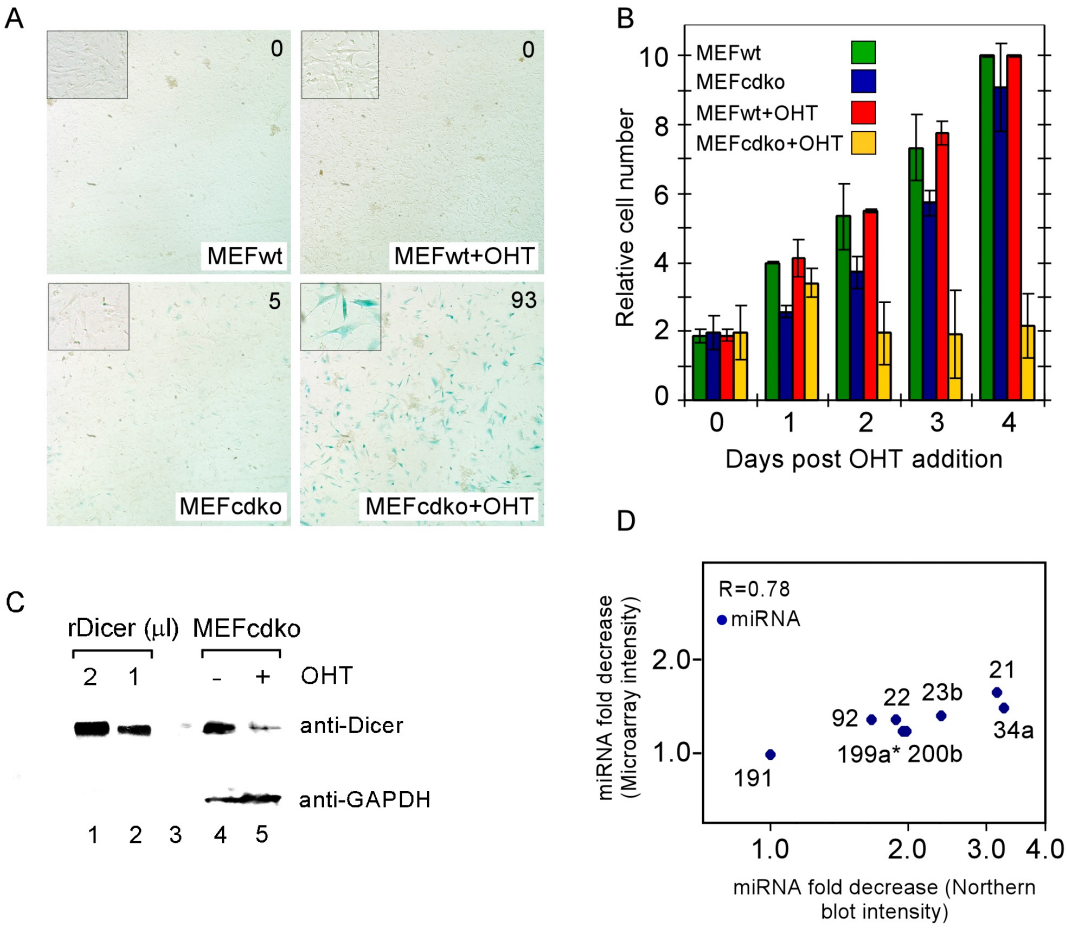

Figure 2: Characterization of conditional Dicer knockout (CDKO) mouse embryonic fibroblasts (MEFs). (A) Wild-type (wt) and CDKO MEFs are shown, untreated (left panels) or 4 d after ortho hydroxy tamoxifen (OHT) treatment (right panels). Cells were stained for LacZ, and the percentage of LacZ-positive cells is shown (upper right). (B) Proliferation of wt and CDKO MEFs, untreated or following addition of OHT. Error bars represent standard deviation of three independent counts. (C) Western analysis of CDKO MEFs, untreated (lane 4) or after OHT addition (lane 5), showing a three- to fourfold reduction in Dicer protein levels following OHT addition. GAPDH is a loading control. Westerns using different concentrations of recombinant Dicer protein are shown as a positive control (lanes 1,2). (D) Microarray hybridization intensity change and expression level change measured by Northern analysis (log scale, both axes) for eight miRNAs (miR-21, miR-22, miR-23b, miR-34a, miR-92, miR-191, miR-199a*, and miR-200b).

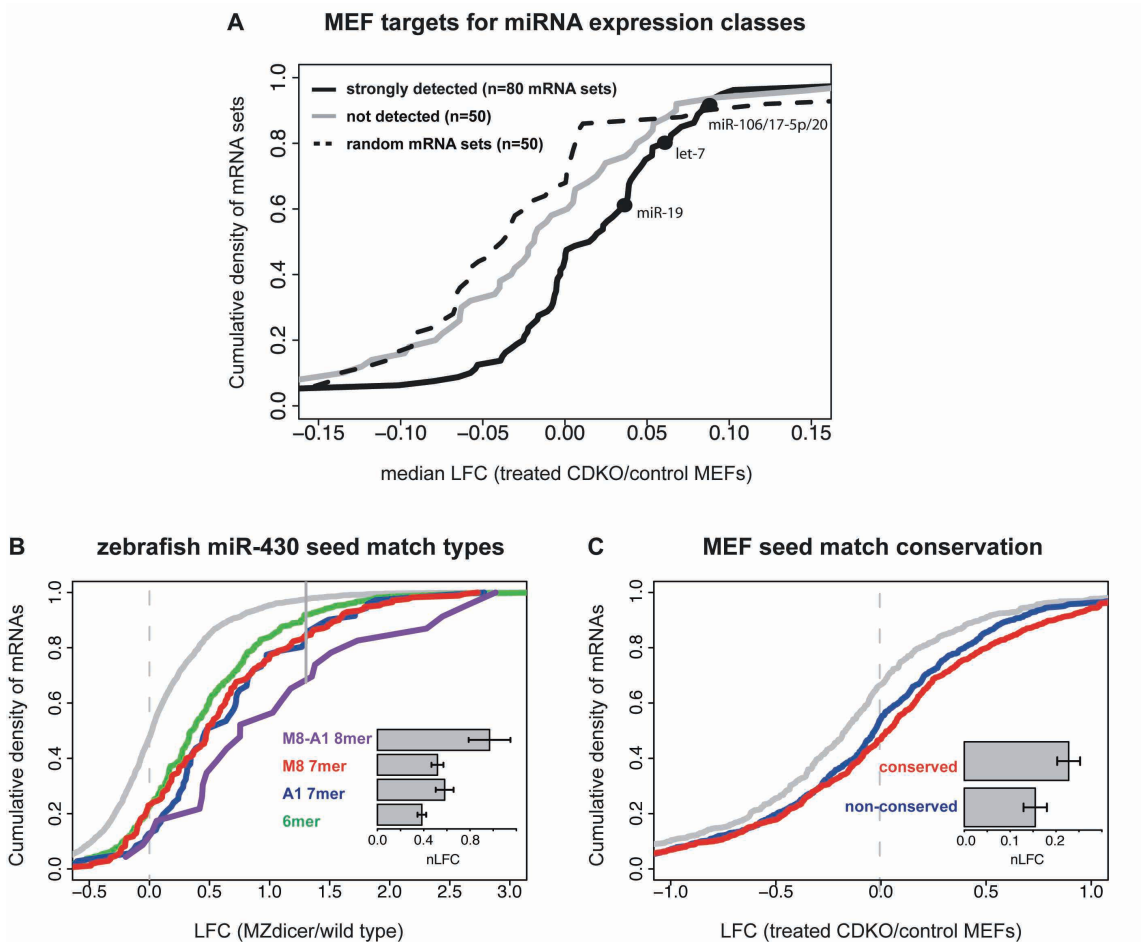

Figure 3: mRNA derepression following Dicer knockout varies with seed match type and conservation status. (A) CDFs of median LFC for three classes of mRNA sets. The expression classes were: (1) mRNA sets containing extended seed matches to the 80 miRNA families whose expression was detected above background by microarray (black curve, selected miRNA family names shown); (2) mRNA sets containing extended seed matches to the 50 miRNA families that were not detectably expressed (gray curve); (3) randomly selected mRNA sets (dotted line). Distributions of detected and nondetected sets are significantly different ($P < 0.01$ by rank sum test), but distributions of nondetected and random mRNA sets are not. (B) CDFs of LFCs for mRNAs containing the indicated miR-430 seed match types - or no miR-430 seed matches (gray curve) - for MZdicer zebrafish embryo data (Giraldez et al. 2006); only mRNAs with exactly one miR-430 seed match were included (seed match type mRNA sets are mutually exclusive). All seed match type LFC distributions differed significantly from the no-seed-match class ($P < 10^{-7}$ by rank sum test), and the pooled extended seed match types differed from the 6 mer class ($P < 0.01$ by rank sum test). (Solid vertical gray line) The LFC below which 97.5% of the no-seed-match mRNA set falls. (Inset bar plot) LFC values for each seed match type, with error bars indicating standard error. (C) CDFs of LFCs for mRNAs containing conserved (red) or exclusively nonconserved (blue) extended seed matches, or no seed matches (gray) to the set of 31 "responsive" miRNAs in the CDKO MEF experiment (Supplemental Material). Seed match count, overall UTR conservation, and mRNA expression level were controlled between the sets. The distributions of mRNA with conserved and nonconserved seed matches are significantly different ($P < 0.05$ by rank sum test). See Supplemental Table S6 for additional statistics.

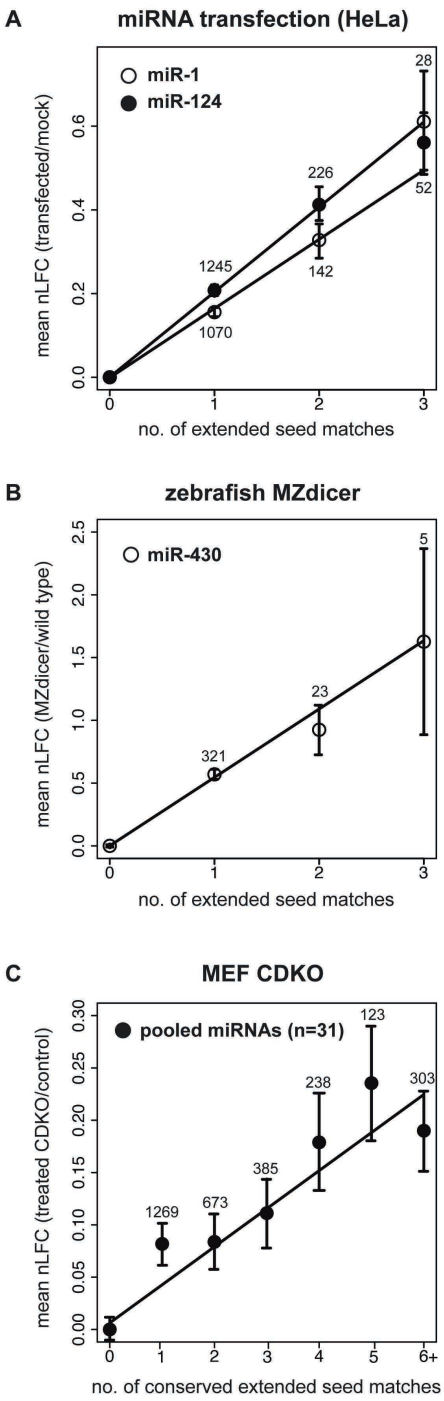

Figure 4: mRNA fold change increases multiplicatively with extended seed match count. (A) For miR-1 (open circles) and miR-124 (solid black circles), the total number of extended seed matches was enumerated for each mRNA, and the mean nLFCs in the Lim transfection experiments were determined for sets of mRNAs grouped by seed match count (set sizes indicated above or below points). (Solid lines) Least squares fit for the whole data set. Error bars correspond to standard error. For each of these plots, the proportions of different seed match types for different seed match counts remained fairly constant. (B) Same as A for miR-430 extended seed match counts following Dicer knockout in zebrafish embryos (Giraldez et al. 2006). (C) Same as A for conserved extended seed match counts to 31 "responsive" mouse miRNAs (see Supplemental Material) in CDKO MEFs.

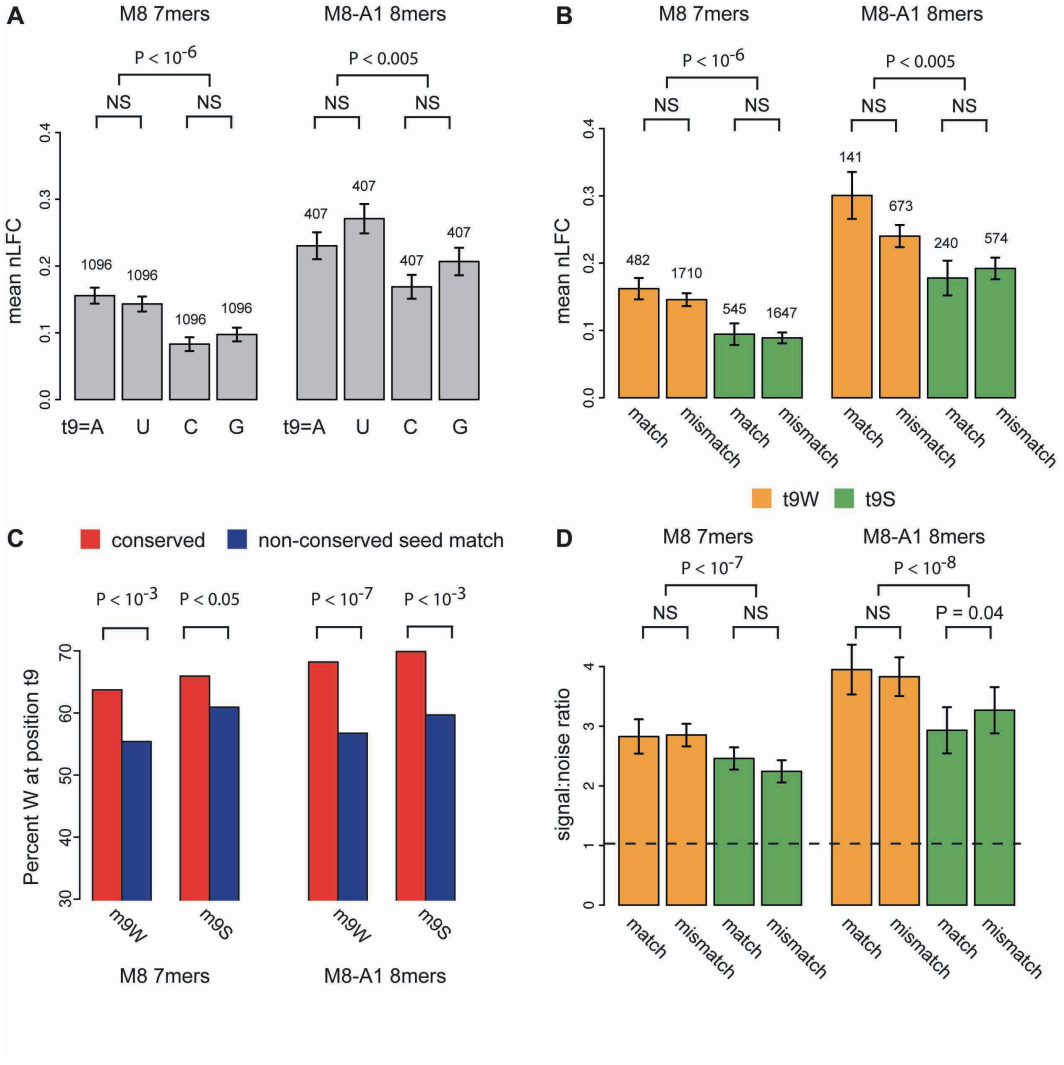

Figure 5: Increased down-regulation of mRNAs with adenosine or uridine at position t9. (A) Mean nLFC for mRNAs containing the indicated nucleotide at position t9 flanking siRNA M8 7 mer and M8-A1 8 mer (rank sum test P-values; NS = not significant at P-value cutoff 0.05). Error bars indicate standard error, and the numbers of mRNAs are indicated above the bars. Each mRNA contained exactly one seed match to any given siRNA (i.e., t9 sets are mutually exclusive), and mRNAs in each of the four t9 sets were controlled for 3′ UTR GC content. Other variables, such as mRNA expression, 3′ UTR conservation, or m9 composition, did not differ significantly between t9 sets. (B) Same as A, but reclassifying the controlled mRNA sets by whether the t9 base pairs with the siRNA m9 (match) or not (mismatch). (C) Enrichment of t9W nucleotides flanking conserved versus nonconserved miRNA M8 7 mer and M8-A1 8 mer in human 3′ UTRs (χ^2 test P-values). The miRNA set consisted of conserved human miRNAs used for target prediction by Lewis et al. (2005) after removal of miRNAs with common m2-m8 seed regions but different m9 nucleotides, and pairs of miRNAs in the same superfamily. The nonconserved seed matches were sampled to match the seed match type, miRNA, and overall UTR CG content of the conserved set. (D) Mean signal:noise ratios for M8 7 mer and M8-A1 8 mer with t9W or t9S in match and mismatch configurations based on cohorts of control oligonucleotides (Lewis et al. 2005) matched for both count and exact CG content (error bars indicate standard deviation based on 14 control cohorts). (Dashed line) Baseline S:N value of 1. P-values based on Wilcoxon rank sum tests between indicated sets (NS = not significant at P-value cutoff 0.05).

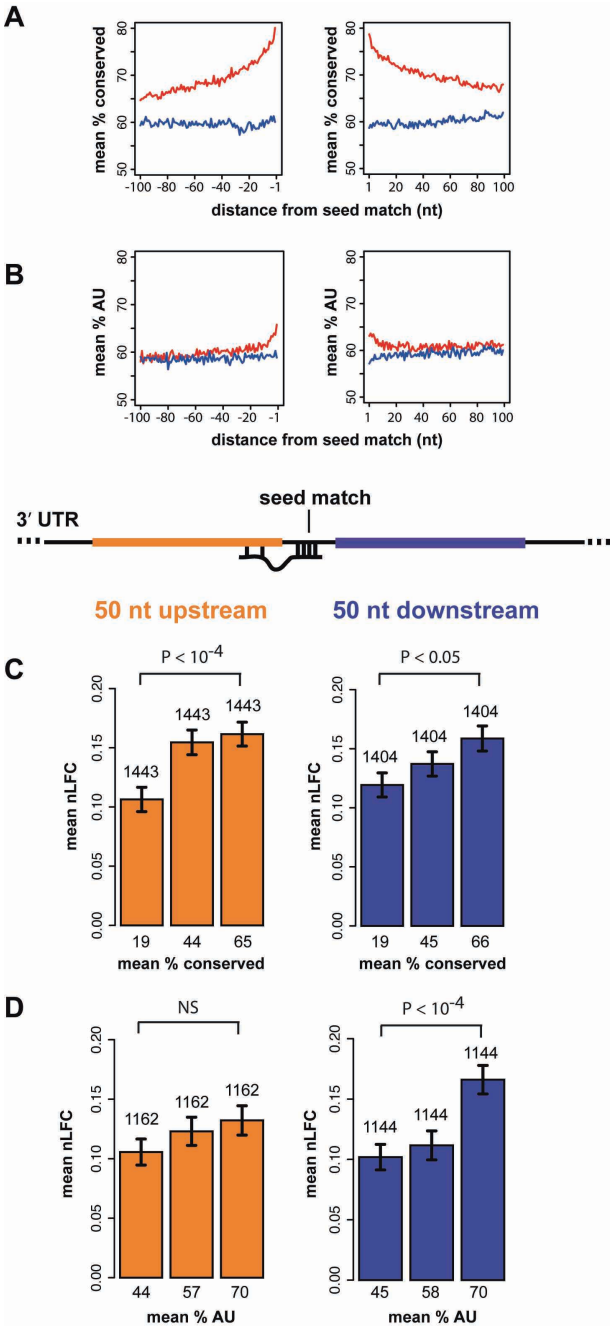

Figure 6: siRNA-directed mRNA repression is enhanced by local conservation and AU content. (A) Mean percent conservation at UTR positions within 100 bp 5′ (left) and 3′ (right) of conserved (red) or nonconserved (blue) extended seed matches to the set of conserved vertebrate miRNAs used in Lewis et al. (2005); overall UTR conservation was controlled for in the comparison of conserved and nonconserved seed matches. The average conservation differs significantly for the 100 bases 5′ and 3′ of the seed match ($P < 10^{-200}$ by rank sum test for both). (B) Mean AU content (sets controlled for UTR AU content); average AU content is significantly different both 5′ and 3′ of the seed match ($P < 10^{-18}$, $P < 10^{-30}$), respectively. (C) Mean nLFC for three equal-sized mRNA sets binned by percent conserved positions (in HMRD) in the 50-nt region immediately 5′ (orange) or 3′ (purple) of siRNA seed matches (5′ region ends at position t10; 3′ region begins one base 3′ of position t1) for mRNAs containing single extended seed match to the relevant siRNA. Bars indicate standard error of the mean. Set size and mean percent conservation for each set are reported above and below each bar, respectively. P-values are for two-sided rank sum tests between the first and third bins. For both 5′ and 3′ conservation, the three bins have been sampled such that their distributions of overall UTR conservation, 5′ (or 3′) AU content, overall UTR AU content, seed match type, and initial expression level are not significantly different ($P \geq 0.05$) (Supplemental Fig. S9). (D) Same as C, but with UTRs binned by AU content in the same 50-nt regions. Bins are sampled to control for UTR AU content, 5′ (or 3′) conservation, overall UTR conservation, seed match type, and initial expression level (NS=not significant at P-value cutoff 0.05).

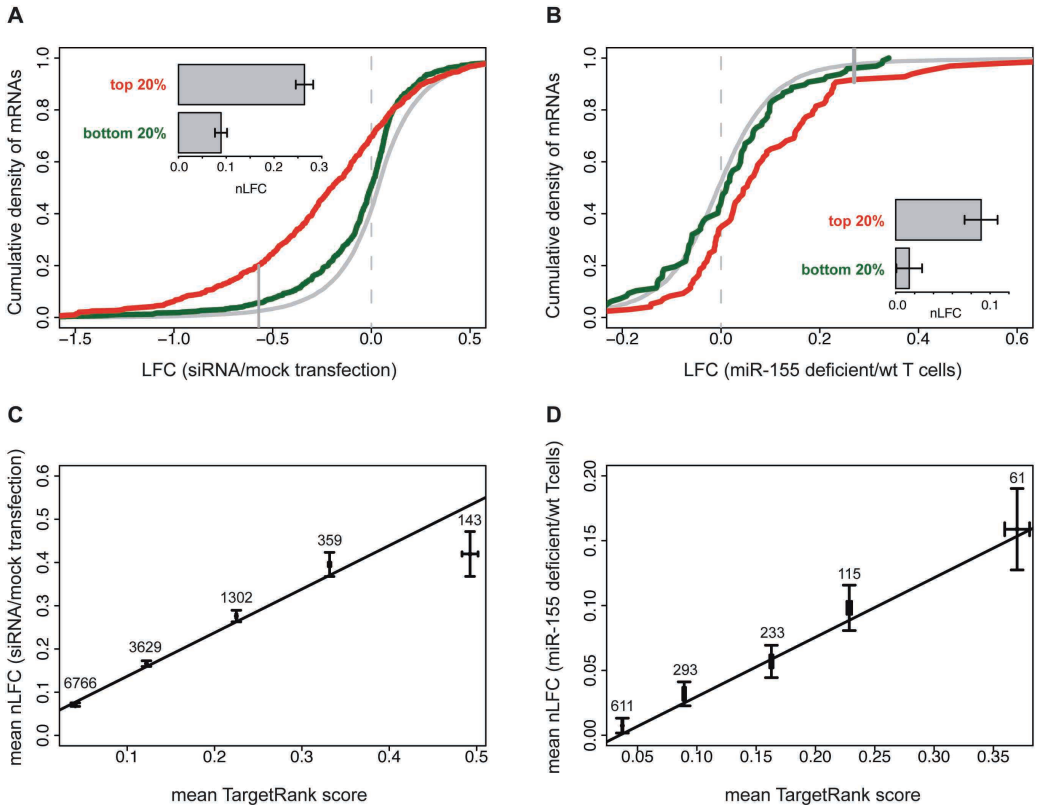

Figure 7: TargetRank scoring separates strongly and weakly down-regulated mRNAs. (A) CDFs of LFCs (as in Fig. 1) for the top 20% (red) and bottom 20% (green) of mRNAs to the relevant siRNA in the test set of eight randomly chosen siRNA transfection experiments ranked by TargetRank score. Only expressed mRNAs containing exactly one 7 mer seed match and no other seed matches of any type were used. For reference, the CDF for mRNAs lacking seed matches to the relevant siRNAs is shown (gray). (Solid vertical gray line) The LFC above which 97.5% of the no-seed-match mRNA set falls. (Inset bar plot) See Fig. 1. See Supplemental Table S7 for additional statistics. (B) Same as A, but for miR-155 knockout T cell data from Rodriguez et al. (2007). (C) All mRNAs containing seed matches (of any type or count) to the relevant siRNA in the eight test siRNA transfection experiments (same sets as in A) were scored using TargetRank. The mean TargetRank score and mean nLFC are plotted with standard error bars for mRNAs sets binned by TargetRank score (mRNA set sizes indicated above points). Line corresponds to least squares fit for entire data set (ANOVA $P < 10^{-100}$); r = 0.23 (Pearson correlation). (D) Same as C, but for miR-155 knockout T cell data from (Rodriguez et al. 2007). Line corresponds to least squares fit for entire data set (ANOVA $P < 10^{-18}$); r = 0.24 (Pearson correlation).

Chapter 3

Nucleosome Positioning at Gene 3′ Ends

Data processing and analysis software for this work was developed in collaboration with Noah Spies.

Corresponding Supplementary Material can be found in Appendix 3.

Chapter 3

Nucleosome Positioning at Gene 3′ Ends

Abstract

Nucleosomes are the basic repeating unit of packaged DNA (chromatin) and their positioning around transcriptional start sites (TSSs) can be dynamic and play important roles in transcriptional regulation. However, nucleosome distributions across other gene structures remain mostly unexplored. Using recently published high-throughput Illumina sequencing data of human nucleosome boundaries (Barski et al., 2007; Schones et al., 2008), we explored patterns of nucleosome positioning at the 3′ ends of genes. We observed a strong depletion of nucleosome density at the terminal ends of genes and demonstrate that the canonical poly(A) signal (PAS), AATAAA, which forms part of a bipartite motif important for co-transcriptional cleavage and polyadenylation of the emerging RNA transcript, has strong nucleosome positioning effects. To investigate whether such nucleosome positioning may influence recognition of the poly(A) site, we considered poly(A) sites with high or low usage, as defined by relative supporting-EST counts, and found evidence for greater nucleosome depletion across high usage sites. A nucleosome affinity model was developed that recapitulates known patterns of nucleosome occupancy in the vicinity of TSSs and PASs based on

primary sequence features. Application of this model genome-wide demonstrated that poly(A) sites with higher nucleosome affinity downstream, consistent with nucleosome occupancy data, appear to have higher relative usage. Our results suggest a connection between DNA accessibility and recognition of transcript sequences associated with RNA processing.

Introduction

In recent years, evidence has emerged for a close coupling of mRNA precursor transcription to subsequent mature mRNA processing. In particular, the dependencies between polyadenylation (poly(A)) site recognition and downstream events such as transcriptional termination and nuclear export have clarified the impact of these RNA processing events on cytoplasmic transcript levels (reviewed by Buratowski (2005); Vinciguerra and Stutz (2004)). The first *in vitro* cleavage and polyadenylation systems provided early mechanistic insights, and the core RNA level signals that guide cleavage and subsequent polyadenylation (addition of ∼200 terminal adenosines) at the 3′ end of transcripts are now well characterized (**?**). However, much remains to be discovered regarding how recognition of such sites is regulated and greater consideration for the co-transcriptional context of this process is warranted.

Cleavage and polyadenylation requires recognition of two core RNA sequence elements: (i) the poly(A) signal (PAS), characterized by an AAUAAA hexamer or close variant, and (ii) the downstream element (DSE), a less well defined U-rich region (Figure 1A). The PAS, located 10-30 nucleotides upstream of the site of cleavage and polyadenylation, is recognized by the cleavage and polyadenylation specificity factor (CPSF), a multimeric protein complex consisting of 160, 100, 73, and 30 kDa

subunits (Zhao et al., 1999). The cleavage stimulation factor (CstF), composed of 77, 64, and 50 kDa protein components, binds to the U-rich DSE approximately 30 or fewer nucleotides downstream of the poly(A) site. Direct protein-protein interactions result in mutual stabilization of the CPSF-CstF-RNA complex, and recent evidence indicates that CPSF-73 acts as the nuclease that catalyzes the cleavage reaction (Ryan et al., 2004; Mandel et al., 2006). Cleavage factors (CF) I_m and II_m and poly(A) polymerase (PAP) are also required to form a cleavage-competent complex on the transcript. Following cleavage, CstF, CFI_m and $CFII_m$ dissociate and the $3'$ cleavage product is degraded by a $5' \rightarrow 3'$ exonuclease, Xrn2, leaving CPSF and PAP to complete the polyadenylation step, together with newly recruited poly(A)-binding protein II (PAB II), required for PAP to achieve its full processive activity.

While cleavage and polyadenylation can occur independently of transcription *in vitro*, there is strong evidence that the two processes are coupled *in vivo* (reviewed by Proudfoot (2004); Zorio and Bentley (2004)). An important molecular link is the C-terminal domain (CTD) of the RNA polymerase II (Pol II) large subunit. With its 52 tandem heptad repeats in mammals (consensus YSPTSPS), the CTD becomes phosphorylated at Ser5 and then preferentially at Ser2, recruiting components of the poly(A) machinery as it proceeds along the gene length (Zhang and Corden, 1991; Komarnitsky et al., 2000). In addition to being a platform for assembly of the poly(A) machinery, the CTD is an essential component of the cleavage reaction and is required *in vitro* in the absence of transcription (McCracken et al., 1997; Hirose and Manley, 1998). The kinetics of transcriptional elongation also appear to have a role in poly(A) site recognition as downstream pause sites or defects in elongation factors can lead to enhanced usage of upstream poly(A) sites (Aranda and Proudfoot, 1999; Birse et al., 1997; Yonaha and Proudfoot, 1999, 2000; Cui and Denis, 2003). Recent reports sug-

gest that recognition of the AAUAAA PAS in emerging nascent transcripts by CPSF, which binds to the Pol II body, triggers transcriptional pausing and enables assembly of an active cleavage complex (Nag et al., 2006, 2007). While the details of assembly mechanics are still emerging, initial ChIP experiments support a role for downstream pausing in many genes (Glover-Cutter et al., 2008). Taken together, these data reveal that through Pol II, DNA level events during transcription can impact processing at the RNA level.

Eukaryotic DNA is packaged into chromatin and its core repeating unit, the nucleosome, is composed of 146 bp of DNA wrapped around an octamer of histone proteins, two copies of each H3, H4, H2A, and H2B. While promoters contain nucleosome free regions, nucleosomes are found across the length of genes (Lee et al., 2004), and provide a continual barrier to Pol II elongation (reviewed by Armstrong (2007)). The cell employs several methods to overcome this barrier, such as H2A/H2B removal and reassembly during transcription facilitated by nucleosome assembly factors, such as FACT (facilitates chromatin transcription) (Belotserkovskaya et al., 2003). In addition, ATP-dependent chromatin remodeling factors are thought to mediate nucleosome sliding and are implicated in a number of steps of transcriptional elongation (reviewed by Armstrong (2007)). Modifications to histone proteins influence their interactions with DNA and in particular, acetylation, commonly associated with actively transcribing genes, may facilitate Pol II passage (Wittschieben et al., 1999). A few studies have suggested intriguing connections between chromatin and co-transcriptional RNA processing events. In particular, Nogues et al. (2002) showed that treatment with trichostatin A, a potent inhibitor of histone deacetylation, correlates with skipping of alternative exons, possibly as a result of hyper-acetylation of core histones allowing passage of transcribing polymerase. More recently, Batsché

et al. (2006) demonstrated a role for SWI/SNF nucleosome remodeling complex in alternative splicing, proposing that it functions by recruiting splicing factors to nascent RNA. In the work described here, we have extended this theme and explored the relationship between nucleosome positioning and 3′ end processing.

Results

Poly(A) signals can influence nucleosome positioning

Recent high-throughput sequencing experiments provide genome-wide nucleosome positioning information at an unprecedented resolution. Much attention as been given to nucleosome positioning around transcriptional start sites (TSSs), however details of positioning patterns across other gene regions remain mostly unexplored.

We used recently published human $CD4^+$ T cell Illumina sequencing data (Barski et al., 2007; Schones et al., 2008) to examine nucleosome distributions flanking poly(A) sites. In both studies, chromatin was digested with micrococcal nuclease (MNase) to produce predominantly mononucleosome-sized DNA. While Schones and coworkers (2008) gel purified ∼150 bp digested fragments for sequencing, Barski and coworkers (2007) sequenced the output of chromatin IP (ChIP) experiments separately using antibodies targeting 20 different histone modifications. Illumina reads with unique genome matches were considered, and the densities of nucleosomes centered at each genomic position were estimated as the average of the read densities derived from each nucleosome end (+/- 75 bp; see Methods). Using read density as a measure of nucleosome occupancy, we observed a striking depletion of nucleosomes +/- 75 bp spanning the canonical PAS, AATAAA, upstream of human EST-supported poly(A) sites (Figure 1B; red curve). Similar patterns were observed using nucleosome data from Schones et al. (2008) (Figure 1B) or pooled data from the 20 histone ChIP experiments in Barski et al. (2007) (not shown). Differences in nucleosome binding affinity have been reported for distinct genomic sequences and, in particular, poly(dA:dT) stretches are known to have poor nucleosome affinity resulting from their resistance to curvature (Drew and Travers, 1985; Satchwell et al., 1986; Peckham et al., 2007).

Nucleosome density plots across control AATAAA hexamer occurrences in intergenic regions (Figure 1B; black curve) confirm that the hexamer alone has a nucleosome positioning effect. Similar distributions were observed around the most common PAS variant, ATTAAA, and other A/T-rich motifs (not shown). The controls showed an enrichment of nucleosomes centered approximately at -150 to -100 and at +100 to +150, suggesting that the AATAAA hexamer leads to nucleosome phasing in the neighboring regions. In Figure 1B, read density per genomic position has been normalized to the mean read density across all plotted positions of both sets (displayed on a $\log_2$ scale). Normalization by this global mean allows above average (>1.0) and below average (<1.0) density regions to be clearly distinguished, while preserving the magnitude of read density differences between the sets. Both control AATAAA hexamers and true PASs showed near average normalized nucleosome density ($\sim$1.0) at +/- 300 bp from the central hexamer. True PASs however did not display a clear phasing of neighboring nucleosomes and showed an even lower nucleosome occupancy across the PAS region. These difference may be in part due to additional avoidance effects from the heterogeneous U-rich DSE, located at a variable downstream distance from the PAS, together with possible U-rich upstream enhancer elements (Hu et al., 2005). These observations led us to ask whether nucleosome positioning near the PAS is associated with differences in PAS activity.

Highly used poly(A) sites are flanked by more precisely positioned nucleosomes

To investigate the relationship between apparent poly(A) site recognition and nucleosome localization, we defined sets of poly(A) sites with either low or high usage. Briefly, we built a database of human poly(A) sites identified using expressed sequence

tags (ESTs) from diverse tissues. ESTs were filtered for evidence of a non-genomically derived poly(A) tail and a canonical or variant PAS in the -1 to -40 region upstream of the aligned poly(A) site (Figure S1; similar to Tian et al. (2005); Yan and Marr (2005)). This approach identified ∼10,000 alternative poly(A) events from ∼5,000 genes. Using an equivalent method to those of Legendre and Gautheret (2003) and Hu et al. (2005), sites with high usage were defined as those supported by greater than 70% of the gene's mapped poly(A) ESTs, whereas low usage sites were defined as those having less than 30% of the supporting ESTs (unless stated otherwise, all considered poly(A) sites had at least 2 supporting poly(A) ESTs). For our analyses, we used tandem sets of poly(A) sites with the most common upstream PASs (AAUAAA, AUUAAA), such that the high and low usage sites were derived from the same gene set.

High and low usage tandem poly(A) sites showed distinct patterns of nucleosome occupancy (Figure 2; blue and red curves respectively). In particular, the high usage sites displayed a more dramatic depletion of nucleosomes in the nucleosome-sized window (+/- 75 bp) flanking poly(A) sites ($P < 10^{-10}$). Greater nucleosome enrichment was observed downstream of high usage sites, between +75 and +375, as compared to low usage sites ($P < 10^{-7}$ in 150 bp windows). Consistent with previous studies (Legendre and Gautheret, 2003; Hu et al., 2005), we observed stronger (closer to consensus) core poly(A) motifs around high usage sites compared to low usage sites, scored using weight matrix models (see Methods). To address whether differences in these core elements, which function at the RNA level, were influencing the differences in nucleosome distributions between high and low usage poly(A) sites, we sampled the poly(A) sets to match their poly(A) motif scores. Figure 2 shows nucleosome densities for the sampled sets, and controlling for the motif scores of core poly(A) elements

had little effect on the initial distribution (not shown). These analyses suggest that high usage sites have more precisely positioned nucleosomes in their flanking regions as compared to less frequently used sites.

A model of nucleosome affinity

High-resolution measurements of nucleosome positions across entire chromosomes, obtained by tiling-microarray hybridization or cloning and sequencing of MNase digested chromatin, offered initial characterizations of genome-wide nucleosome distributions and sequence preferences (Satchwell et al., 1986; Yuan et al., 2005; Segal et al., 2006). Early models, based on dinucleotide content of nucleosome associated sequences, estimated that 54% of *in vivo* nucleosome positions could be correctly predicted from sequence alone compared to 39% expected by chance, i.e. within +/- 35 bp of reported positions (Segal et al., 2006), and use of larger training sets improves performance (J. Widom, personal communication). These studies suggest that while eukaryotic cells contain abundant ATP-dependent nucleosome remodeling complexes, whose activity may over-ride nucleosome sequence preferences in some situations or lead to nucleosome removal from specific regions, it is clear that intrinsic sequence affinities are probably the major contributor to global nucleosome localization.

Nucleosome read densities from recent Illumina sequencing experiments (Barski et al., 2007; Schones et al., 2008) are orders of magnitude higher than those achieved using earlier methods. These studies offer ∼10-fold coverage of all nucleosomes, assuming one nucleosome every 200 bp, however the number of reads per genomic position remains low (0.05) and cannot be used to infer nucleosome positioning within a unique region. Observed nucleosome occupancy, inferred from read densities, is

influenced by both intrinsic nucleosome sequence affinities and by localization of DNA-binding proteins or chromatin remodeling enzymes. To clearly separate these effects and enable assessment of individual 3′ UTRs not well represented in the Illumina data, we sought an objective scoring function for nucleosome sequence affinity. Several groups have constructed such models. However, most were trained using sequences from the yeast *Sacchromyces cerevisiae*, and due to the small training set size, they only considered dinucleotide content (Segal et al., 2006; Ioshikhes et al., 2006; Peckham et al., 2007; Yuan and Liu, 2008). Given the remarkably large number of nucleosome boundary sequences captured by the more recent human T cell experiments (Barski et al., 2007; Schones et al., 2008), we set out to devise a model trained using human sequence and leveraging the large training set size to capture higher order nucleotide composition.

Using ∼84 million randomly chosen reads, representing 75% of the perfectly and uniquely mapping reads in the Barski et al. (2007) data set, we trained a fifth-order Markov model for each position in the 146 bp downstream of mapped read starts. The choice of a fifth-order model was motivated by both the magnitude of available training data, and by observations that hexamers with similar dinucleotide content showed different nucleosome occupancy distributions in flanking regions (not shown). A background model was trained in the same fashion using an equal-sized set of randomly selected positions from neighboring regions (see Methods). The first 15 nt of the models were excluded due to the strong sequence bias in the nucleosome start data, which appears related to technical issues, such as the preference for MNase to cleave NpA and NpT diester bonds (Figure S2). MNase cleavage occurs preferentially at nucleosome boundaries, however, in addition to its predominant endonucleolytic activity, it can also subsequently function as an exonuclease, leading to heterogeneity

in the cleavage position (Hörz and Altenburger, 1981). While the effects will vary with digestion conditions, earlier studies used electrophoresis and cloning and estimated such terminal variations to be typically only a few base pairs (Johnson et al., 2006). High affinity nucleosome sequences are known to contain AT-dinucleotides with ∼10-bp periodicity, ∼5 bp out of phase with periodic GC-dinucleotides. These AT and GC periodicities are thought to facilitate bending at positions where the major groove of DNA faces toward or away from the histone octamer, respectively. To capture these periodicities, Segal et al. (2006) trained their nucleosome affinity model using center-aligned nucleosome bound sequences from yeast. Having only terminal reads, we cannot apply this alignment approach. However, Segal and coworkers (2006) recorded dinucleotide frequencies using a 3 bp moving average, which generated smoother distributions from their small training set and which they justified by citing experimental evidence that small changes in spacing of key nucleosome DNA sequence motifs can occur with relatively small cost to the free energy of histone-DNA interactions. Given that the anticipated variability in MNase cleavage is on this same order of magnitude (+/- a few base pairs), we anticipate that global patterns will be accurately captured by our model.

To test our model, we compared our predicted nucleosome affinity scores to well-characterized nucleosome occupancy patterns, inferred from Illumina read densities, for regions surrounding TSSs and found that our model qualitatively reproduces observed distributions (Figure S3). We then used our model to score regions surrounding poly(A) sites of low and high usage. Comparisons of nucleosome affinity scores (Figure 3A) with observed nucleosome occupancy (Figure 2) indicated that our model yields a distribution of affinity scores that resembles the measured nucleosome distribution in overall shape. Reduced nucleosome affinity is predicted near high usage

poly(A) sites (0 to +75 bp) and greater nucleosome affinity is predicted downstream (+75 to +450) as compared to low usage sites (red and blue curves, Figure 3A). One limitation of our model is that it scores nucleosome affinity per position along a sequence, but does not consider spatial constraints, such as the inability of two nucleosomes to occupy overlapping positions. Such positional constraints certainly influence observed nucleosome occupancy, and while we plan to make appropriate modifications to next generation models, these effects may in part explain current discrepancies between our model and the data.

Higher downstream nucleosome affinity is associated with increased usage independent of core poly(A) motifs

Using our scoring method, we explored whether intrinsic nucleosome affinity is predictive of poly(A) site usage. In particular, we hypothesized that for genes with multiple alternative poly(A) sites, the relative nucleosome affinity of a site compared to its competing sites may influence its usage. To test whether larger differences in nucleosome affinity are associated with larger differences in usage, we collected a set of alternative sites (see Methods) and calculated pairwise differences in nucleosome affinity score and usage values. This was done by taking the value for the upstream poly(A) site in a pair and subtracting off the value for the downstream site. Negative differences were thus obtained when the downstream site had higher affinity or usage. To control for the effects of the core poly(A) motifs, we chose to examine nucleosome affinity downstream of the motifs (from +100 to +246 relative to the poly(A) site). This analysis controlled for pairwise differences in poly(A) motif scores by subtracting the upstream score from the downstream score, as done for the other measures.

Figure 3B illustrates that pairs of alternative poly(A) sites binned by relative nucleosome affinity (leftmost panel), and sampled to control for relative poly(A) motif scores, had different relative usage values (rightmost panel). In particular, site pairs with the most negative change in relative affinity (i.e. the downstream site had higher affinity than the upstream partner), showed the greatest change in relative usage. For all relative affinity bins however, the mean relative usage values were negative, suggesting a strong underlying preference for the downstream site. The modest yet significant effect observed here ($P < 10^{-4}$) may underestimate the magnitude of the true effect, as we have applied a very stringent test that considers only nucleosome affinity at a downstream location (+100) in order to exclude the potential impact of known RNA level signals.

Impact of $3'$ UTR context

In tandem sets, internal poly(A) sites are typically used less frequently than the terminal site (Tian et al., 2005). As a result, low and high usage sites typically differ in their flanking nucleotide compositions, with high usage sites often next to downstream intergenic sequence and low usage sites adjacent to $3'$ UTR internal sequence. To address whether such sequence context could influence nucleosome density, we examined tandem poly(A) sites from internal or terminal positions separately (Figure 4A).

Considering only terminal poly(A) sites, nucleosome distributions for high usage (red curve) and low usage (blue curve) poly(A) sites displayed similar global patterns to those seen previously (compare Figure 4B with Figure 2). However, comparison of regions of nucleosome depletion (-75 to +75 bp) or enrichment (+75 to +375 bp)

for high versus low usage sites, revealed that the differences in nucleosome occupancy were now more modest, indicating that sequence context is a factor. The nucleosome distributions observed flanking internal tandem poly(A) sites had more pronounced differences (Figure 4C). Again, high usage sites showed more extensive depletion of nucleosomes across the poly(A) site (-75 to +75 bp), however, instead of downstream enrichment, we observed nucleosome peaks upstream of high usage poly(A) sites, but not low usage sites ($P < 0.01$; 150 nt windows). Both these analyses (Figure 4B,C) were controlled for poly(A) core motif scores. Poly(A) sites with annotated stop codons within 150 nt were excluded to eliminate effects of protein coding sequence, and these peaks persisted despite more stringent filters using a +/- 300 nt window (not shown). One pattern that emerges from these studies is that high usage sites display more dramatic nucleosome depletion in the region spanning the poly(A) site, independent of RNA level PASs or 3′ UTR context.

Discussion

In this study, we sought to explore patterns of nucleosome positioning at the 3′ ends of genes, using recently published high-throughput sequencing data from nucleosome boundaries (Barski et al., 2007; Schones et al., 2008). We observed that the PAS has strong nucleosome positioning effects, and consistent with early studies of nucleosome localization across runs of poly(dA:dT) (Drew and Travers, 1985), the PAS appears to be avoided in the nucleosome core. By considering poly(A) sites with high or low usage, as defined by relative supporting-EST counts, we found evidence for lower nucleosome occupancy across high usage sites and higher nucleosome occupancy downstream of the PAS, independent of poly(A) motif scores or 3′ UTR context. We trained a nucleosome affinity model using data sampled from across the genome, and demonstrated that reduced relative nucleosome affinity adjacent to the poly(A) site and higher relative downstream affinity are characteristic of higher relative usage. This observation suggests that differences in intrinsic nucleosome affinity underly some of the differences in nucleosome occupancy.

A handful of experiments point to connections between chromatin structure and RNA processing. In two studies from the Proudfoot lab, mutations in the chromatin-remodeling enzymes, Chd1 and Isw1, were both shown to lead to defects in transcriptional termination in yeast (Alén et al., 2002; Morillon et al., 2003). These factors were implicated in enhancing Pol II pausing during the switch from elongation to termination. Pol II pausing also has a documented role in regulating alternative RNA splicing. Slowing the elongation rate of Pol II, as a result of point mutation, led to increased inclusion of the human alternative fibronectin EDI exon with weak splice sites (de la Mata et al., 2003). Together with observations that downstream Pol II

pause sites induced greater poly(A) site usage (Aranda and Proudfoot, 1999; Birse et al., 1997; Yonaha and Proudfoot, 1999, 2000), these experiments support a kinetic model where Pol II pausing leads to greater recognition time between the RNA signals in the nascent transcript and CTD associated RNA processing enzymes. Recently, qRT-PCR and ChIP experiments demonstrated that overexpression of the human SWI/SNF subunit Brm led to increased inclusion of and Pol II accumulation on a variable exon in the CD44 gene, implicating this chromatin-remodeling protein in Pol II pausing (Batsché et al., 2006). Nucleosomes themselves can impede transcription *in vitro* (Knezetic and Luse, 1986). In considering our observations of nucleosome density flanking poly(A) sites, one model is that downstream regions with greater nucleosome affinity pose a barrier to the elongating Pol II, and induce slowing/pausing, leading to more frequent recognition of the emerging PASs. However, it is difficult to reconcile this model with the observed peaks in nucleosome density upstream of high usage internal sites (Figure 4C; red curve). Allemand and coworkers (2008) proposed that histones, with their N-terminal tails rich in basic residues, may sequester nascent RNA directly and facilitate recruitment of RNA processing enzymes. In this case, perhaps an abundance of nucleosomes on either side of the poly(A) site is sufficient to promote recognition. However, the most consistent pattern we observed was a greater depletion of nucleosomes across the poly(A) region in high usage versus low usage sites (Figure 2, Figure 4 B,C). This drop in nucleosome occupancy is reminiscent of evicted nucleosomes immediately upstream of TSSs at promoters, which is thought to provide greater access to transcription factors. It is possible that association of DNA-binding proteins with sequences flanking or overlapping the genomic poly(A) site could compete with nucleosomes for occupancy of this region, and either through interactions with Pol II or the Pol II associated poly(A) machinery, facilitate recognition of the PAS on the nascent transcript.

There are several areas where our analyses could be extended. First, emerging high-throughput cDNA sequencing data (mRNA-SEQ) from a spectrum of normal tissues has the potential to provide more accurate relative expression values for alternative regions such as tandem 3′ UTRs, and should be explored in the future. Using those data to identify alternative poly(A) sites may also lead to a larger 3′ UTR set, as compared to our current method which generates ∼10,000 poly(A)-EST supported alternative sites. Second, our model of nucleosome affinity only considers primary sequence and does not consider spatial constraints. It is possible that we may predict a region with a high score to be occupied by a nucleosome, however, this site may be unoccupied in the presence of an overlapping region with even higher affinity. Capturing aspects of these spatial constraints, either by iteratively searching for the highest affinity site in the remaining unoccupied sequence space, or by more sophisticated methods (Segal et al., 2006), should be explored. Considering our careful controls, we anticipate that these improvements will add statistical power and accuracy to our analyses, but are unlikely to fundamentally change the results presented here.

Recent studies have demonstrated specific examples of significant nucleosome repositioning in different cell states. In particular, Schones and coworkers (2008) provided evidence for stronger phasing of nucleosomes relative to the TSS in transcriptionally active genes compared to inactive genes, correlated with Pol II binding. It is interesting to consider whether such nucleosome repositioning occurs around poly(A) sites. For example, during B cell differentiation into plasma cells, cells change their relative usage of two IgM heavy chain alternative poly(A) sites, switching from preferentially processing a downstream poly(A) site in the B cell stage to favoring an upstream, intronic site in plasma cells (reviewed by Edwalds-Gilbert et al. (1997)). The truncated transcript produced in plasma cells lacks two downstream transmembrane

domains and thus is responsible for the developmental switch from membrane-bound to secreted IgM. It would be interesting to explore whether differences in nucleosome positioning occur across alternative poly(A) sites in different cell states. While not addressing this question directly, preliminary analyses using constitutive poly(A) sites in transcriptionally active and inactive genes do not show dramatic nucleosome repositioning as observed at TSSs (Figure S4). It seems reasonable that the binding of Pol II to the TSS of transcriptionally active genes may be responsible for this repositioning. As poly(A) factors recognize their target motifs at the RNA level, it is possible that there is no equivalent competition between a DNA-binding protein and nucleosomes to induce such changes near poly(A) sites, however the nature of nucleosome dynamics at distinct gene locations remains to be explored.

Methods

Illumina read data

Genome mappings (to human genome release hg18) of Illumina sequencing reads (mostly 25 bp in length) were downloaded from `http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgTcell.html` and from `http://dir.nhlbi.nih.gov/papers/lmi/epigenomes/hgtcellnucleosomes.html`. Only uniquely mapping reads were considered, producing 118 million nucleosome derived reads from Barski et al. (2007), and 254 million reads from Schones et al. (2008) where one and two mismatches were tolerated. Reads with greater than 10 occurrences were removed to avoid rare outlier cases likely a result of technical biases. Given an arbitrarily defined sense strand (the direction of transcription for gene regions), reads matching the sense strand were labeled 5′ and reads matching the antisense strand were labeled 3′ to reflect their positions on either side of the nucleosome. For all analyses, total read counts at a given genomic position for a region set were normalized to the number of regions, to generate mean read counts per position (density). At each genomic position, the mean nucleosome density was estimated by taking the average of the 5′ read density at -75 bp and the 3′ read density at +75 bp, positions roughly corresponding to the centered nucleosome ends. These normalize values were divided by a global average in the query region to facilitate distinguishing enriched and depleted nucleosome densities.

Poly(A) sites: database construction and subsets for analysis

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 non-genomically derived poly(A) tail and a canonical or variant PAS (Beaudoing and Gautheret, 2001) in the -1 to -40 region upstream of the aligned poly(A) site (Figure S1). The resulting set was then mapped to a comprehensive and non-redundant set of reference sequence (Refseq) transcripts (Pruitt et al., 2005) and clustered to create a database of poly(A) sites.

Tandem 3′ UTRs used in Figure 2 and Figure 3A were selected to contain either an AATAAA or ATTAAA PAS in the -40 to -1 region upstream of the poly(A) site. Larger poly(A) site sets used for analysis in Figure 3B followed poly(A) definitions outlined in Tian et al. (2005), as used for the generation of PolyA_DB (sites with only single poly(A)-EST support were included). Intergenic control hexamers were selected from regions without any cDNA/EST coverage and at least 500 nt from the nearest gene annotation. The 200 nt around each AATAAA hexamer was filtered to not contain any repeat elements as detected by RepeatMasker.

Poly(A) motif scoring

Weight matrix models of core poly(A) motifs described in 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 sum of the score for the CUE2 element, corresponding the PAS, and the average score for the CDE1-CDE4 elements, corresponding to the U-rich downstream signals, was reported as the core poly(A) motif score.

Sampling procedures and statistics

Sampling between poly(A) site sets was done by selecting a set as a reference (typically the smaller set), and then for each poly(A) site in the reference set, selecting at random and without replacement a poly(A) site with a similar value for the query variable (within some predefined window) from each of the other sets. If such a poly(A) site could not be found, the initial reference poly(A) site was skipped. Sampled sets were tested by Wilcoxon rank sum test to ensure they did not differ in the query variable ($P > 0.1$).

Nucleosome affinity model

~84 million Illumina read starts, representing 75% of the perfectly and uniquely mapping reads in the 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 $P(X_n = x|X_{n-1} = x_{n-1}, ..., X_{n-5} = x_{n-5})$ for every $n = 1...146$. Due to bias in the first 15 nt downstream of the read starts (Figure S2), positions 1 to 15 were subsequently excluded from the model. This effect is likely a result of MNase bias to cleave NpA and NpT diester bonds (Hörz and Altenburger, 1981) and has been observed previously (Johnson et al., 2006).

Nucleosome affinity scores were calculated as the $\log_2$ ratio of P(seq|nucleosome model) to P(seq|background model). As only positions 16 to 146 of the nucleosome bound

sequence were considered due to sequence bias mentioned above, the coordinate for an affinity score was corrected by -15 nt. Scores were plotted at a 73 bp offset to reflect the center of the corresponding nucleosome.

References

C. Alén, N. A. Kent, H. S. Jones, J. O'Sullivan, A. Aranda, and N. J. Proudfoot. A role for chromatin remodeling in transcriptional termination by RNA polymerase II. *Mol Cell*, 10(6):1441–52, Dec 2002.

E. Allemand, E. Batsché, and C. Muchardt. Splicing, transcription, and chromatin: a ménage à trois. *Curr Opin Genet Dev*, Mar 2008.

A. Aranda and N. J. Proudfoot. Definition of transcriptional pause elements in fission yeast. *Mol Cell Biol*, 19(2):1251–61, Feb 1999.

J. A. Armstrong. Negotiating the nucleosome: factors that allow RNA polymerase ii to elongate through chromatin. *Biochem Cell Biol*, 85(4):426–34, Aug 2007.

A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007.

E. Batsché, M. Yaniv, and C. Muchardt. The human SWI/SNF subunit Brm is a regulator of alternative splicing. *Nat Struct Mol Biol*, 13(1):22–9, Jan 2006.

E. Beaudoing and D. Gautheret. Identification of alternate polyadenylation sites and analysis of their tissue distribution using EST data. *Genome Research*, 11: 1520–1526, 2001.

R. Belotserkovskaya, S. Oh, V. A. Bondarenko, G. Orphanides, V. M. Studitsky, and D. Reinberg. FACT facilitates transcription-dependent nucleosome alteration. *Science*, 301(5636):1090–3, Aug 2003.

C. E. Birse, B. A. Lee, K. Hansen, and N. J. Proudfoot. Transcriptional termination signals for RNA polymerase II in fission yeast. *EMBO J*, 16(12):3633–43, Jun 1997.

S. Buratowski. Connections between mRNA 3′ end processing and transcription termination. *Current Opinion in Cell Biology*, 17:257–261, 2005.

Y. Cui and C. L. Denis. In vivo evidence that defects in the transcriptional elongation factors RPB2, TFIIS, and SPT5 enhance upstream poly(A) site utilization. *Mol Cell Biol*, 23(21):7887–901, Nov 2003.

M. de la Mata, C. R. Alonso, S. Kadener, J. P. Fededa, M. Blaustein, F. Pelisch, P. Cramer, D. Bentley, and A. R. Kornblihtt. A slow RNA polymerase II affects alternative splicing in vivo. *Mol Cell*, 12(2):525–32, Aug 2003.

H. R. Drew and A. A. Travers. DNA bending and its relation to nucleosome positioning. *J Mol Biol*, 186(4):773–90, Dec 1985.

G. Edwalds-Gilbert, K. L. Veraldi, and C. Milcarek. Alternative poly(A) site selection in complex transcription units: means to an end? *Nucleic Acids Research*, 25(13): 2547–2561, 1997.

K. Glover-Cutter, S. Kim, J. Espinosa, and D. L. Bentley. RNA polymerase II pauses and associates with pre-mRNA processing factors at both ends of genes. *Nat Struct Mol Biol*, 15(1):71–8, Jan 2008.

Y. Hirose and J. L. Manley. RNA polymerase II is an essential mRNA polyadenylation factor. *Nature*, 395(6697):93–6, Sep 1998.

W. Hörz and W. Altenburger. Sequence specific cleavage of DNA by micrococcal nuclease. *Nucleic Acids Res*, 9(12):2643–58, Jun 1981.

J. Hu, C. S. Lutz, J. Wilusz, and B. Tian. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA*, 11(10): 1485–93, Oct 2005.

I. P. Ioshikhes, I. Albert, S. J. Zanton, and B. F. Pugh. Nucleosome positions predicted through comparative genomics. *Nat Genet*, 38(10):1210–5, Oct 2006.

S. M. Johnson, F. J. Tan, H. L. McCullough, D. P. Riordan, and A. Z. Fire. Flexibility and constraint in the nucleosome core landscape of Caenorhabditis elegans chromatin. *Genome Res*, 16(12):1505–16, Dec 2006.

J. A. Knezetic and D. S. Luse. The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell*, 45(1):95–104, Apr 1986.

P. Komarnitsky, E. J. Cho, and S. Buratowski. Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription. *Genes Dev*, 14(19):2452–60, Oct 2000.

C.-K. Lee, Y. Shibata, B. Rao, B. D. Strahl, and J. D. Lieb. Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet*, 36(8):900–5, Aug 2004.

M. Legendre and D. Gautheret. Sequence determinants in human polyadenylation site selection. *BMC Genomics*, 4:7, 2003.

C. R. Mandel, S. Kaneko, H. Zhang, D. Gebauer, V. Vethantham, J. L. Manley, and L. Tong. Polyadenylation factor CPSF-73 is the pre-mRNA 3′-end-processing endonuclease. *Nature*, 444(7121):953–6, Dec 2006.

S. McCracken, N. Fong, K. Yankulov, S. Ballantyne, G. Pan, J. Greenblatt, S. D. Patterson, M. Wickens, and D. L. Bentley. The C-terminal domain of RNA polymerase II couples mRNA processing to transcription. *Nature*, 385(6614):357–61, Jan 1997.

C. L. Moore and P. A. Sharp. Site-specific polyadenylation in a cell-free reaction. *Cell*, 36(3):581–91, Mar 1984.

C. L. Moore and P. A. Sharp. Accurate cleavage and polyadenylation of exogenous rna substrate. *Cell*, 41(3):845–55, Jul 1985.

A. Morillon, N. Karabetsou, J. O'Sullivan, N. Kent, N. Proudfoot, and J. Mellor. Isw1 chromatin remodeling ATPase coordinates transcription elongation and termination by RNA polymerase II. *Cell*, 115(4):425–35, Nov 2003.

A. Nag, K. Narsinh, A. Kazerouninia, and H. G. Martinson. The conserved AAUAAA hexamer of the poly(A) signal can act alone to trigger a stable decrease in RNA polymerase II transcription velocity. *RNA*, 12(8):1534–44, Aug 2006.

A. Nag, K. Narsinh, and H. G. Martinson. The poly(A)-dependent transcriptional pause is mediated by CPSF acting on the body of the polymerase. *Nat Struct Mol Biol*, 14(7):662–9, Jul 2007.

G. Nogues, S. Kadener, P. Cramer, D. Bentley, and A. R. Kornblihtt. Transcriptional activators differ in their abilities to control alternative splicing. *J Biol Chem*, 277 (45):43110–4, Nov 2002.

H. E. Peckham, R. E. Thurman, Y. Fu, J. A. Stamatoyannopoulos, W. S. Noble, K. Struhl, and Z. Weng. Nucleosome positioning signals in genomic DNA. *Genome Res*, 17(8):1170–7, Aug 2007.

N. Proudfoot. New perspectives on connecting messenger RNA 3′ end formation to transcription. *Curr Opin Cell Biol*, 16(3):272–8, Jun 2004.

K. D. Pruitt, T. Tatusova, and D. R. Maglott. Ncbi reference sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Research*, 33(Database issue):D501–D504, 2005.

K. Ryan, O. Calvo, and J. L. Manley. Evidence that polyadenylation factor CPSF-73 is the mrna 3′ processing endonuclease. *RNA*, 10(4):565–73, Apr 2004.

S. C. Satchwell, H. R. Drew, and A. A. Travers. Sequence periodicities in chicken nucleosome core DNA. *J Mol Biol*, 191(4):659–75, Oct 1986.

D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–98, Mar 2008.

E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–8, Aug 2006.

B. Tian, J. Hu, H. Zhang, and C. S. Lutz. A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic Acids Research*, 33(1):201–212, 2005.

P. Vinciguerra and F. Stutz. mRNA export: an assembly line from genes to nuclear pores. *Current Opinion in Cell Biology*, 16:285–292, 2004.

B. O. Wittschieben, G. Otero, T. de Bizemont, J. Fellows, H. Erdjument-Bromage, R. Ohba, Y. Li, C. D. Allis, P. Tempst, and J. Q. Svejstrup. A novel histone acetyltransferase is an integral subunit of elongating RNA polymerase II holoenzyme. *Mol Cell*, 4(1):123–8, Jul 1999.

J. Yan and T. G. Marr. Computational analysis of 3′-ends of ESTs shows four classes of alternative polyadenylation in human, mouse, and rat. *Genome Research*, 15(3): 369–375, 2005.

M. Yonaha and N. J. Proudfoot. Specific transcriptional pausing activates polyadenylation in a coupled in vitro system. *Mol Cell*, 3(5):593–600, May 1999.

M. Yonaha and N. J. Proudfoot. Transcriptional termination and coupled polyadenylation in vitro. *EMBO J*, 19(14):3770–7, Jul 2000.

G.-C. Yuan and J. S. Liu. Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol*, 4(1):e13, Jan 2008.

G.-C. Yuan, Y.-J. Liu, M. F. Dion, M. D. Slack, L. F. Wu, S. J. Altschuler, and O. J. Rando. Genome-scale identification of nucleosome positions in S. cerevisiae. *Science*, 309(5734):626–30, Jul 2005.

J. Zhang and J. L. Corden. Identification of phosphorylation sites in the repetitive carboxyl-terminal domain of the mouse RNA polymerase II largest subunit. *J Biol Chem*, 266(4):2290–6, Feb 1991.

J. Zhao, L. Hyman, and C. Moore. Formation of mRNA 3′ ends in eukaryotes: Mechanism, regulation, and interrelationships with other steps in mRNA synthesis. *Microbiology and Molecular Biology Reviews*, 63(2):405–445, 1999.

D. A. R. Zorio and D. L. Bentley. The link between mRNA processing and transcription: communication works both ways. *Exp Cell Res*, 296(1):91–7, May 2004.

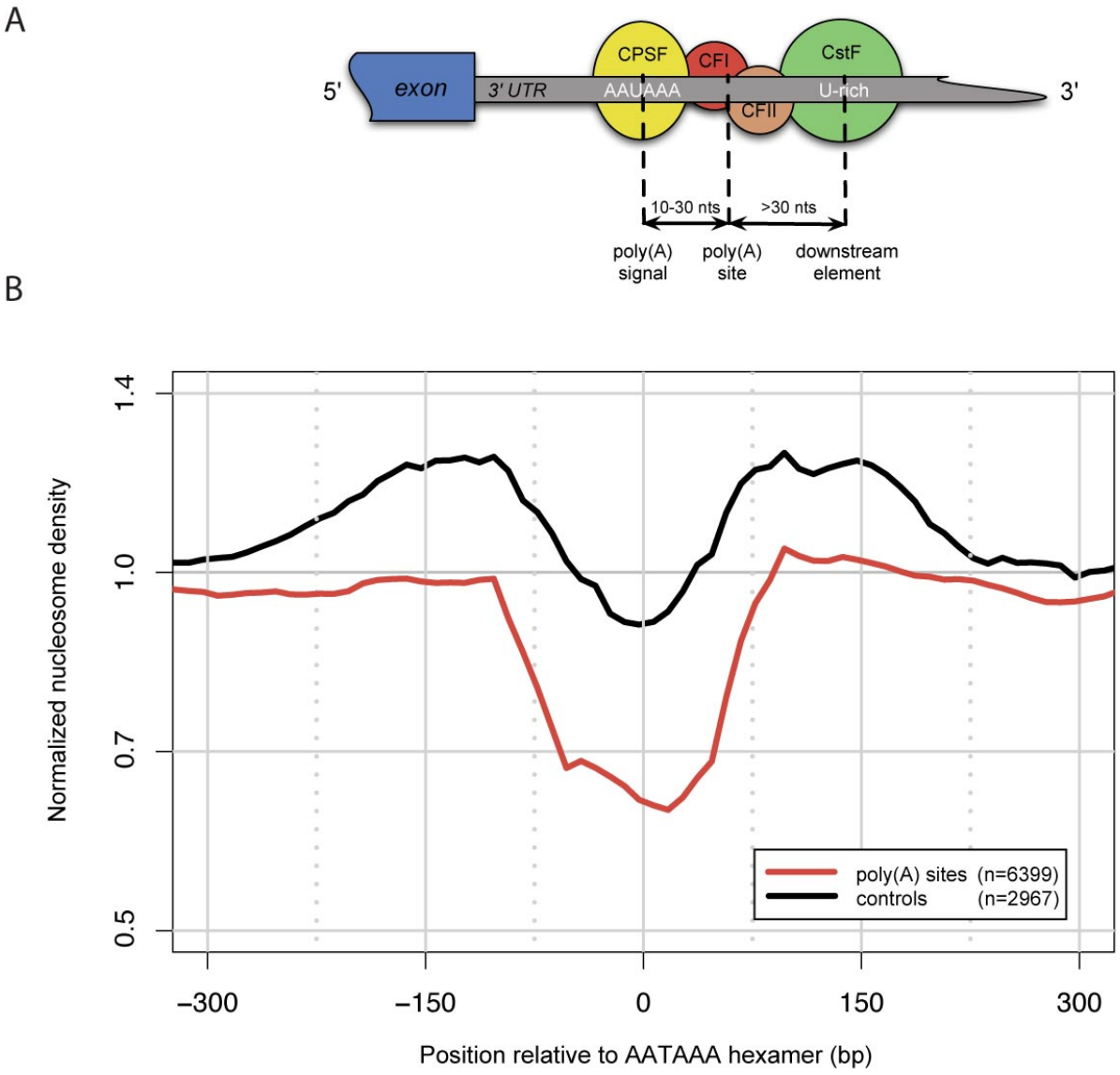

Figure 1: **A**. Schematic representation of mammalian PASs based on Zhao et al., 1999. **B**. Nucleosome density around human PASs (red) or control intergenic AATAAA hexamers (black). Position 0 corresponds to the first base of the AATAAA hexamer. Only poly(A) sites containing at least one occurrence of this hexamer in the 40 nt upstream of the poly(A) site were included, and sites were filtered to be at least 500 nt from the upstream stop codon and any other alternative poly(A) site. Nucleosome density was calculated as the average of the 5′ and 3′ read densities (Schones et al., 2008 data), normalized to the mean density for both sets in the entire window of +/- 500 bp, and plotted on a $\log_2$ scale (see Methods). Normalized density values were smoothed by plotting the average value from 50 nt sliding windows positioned every 10 nt.

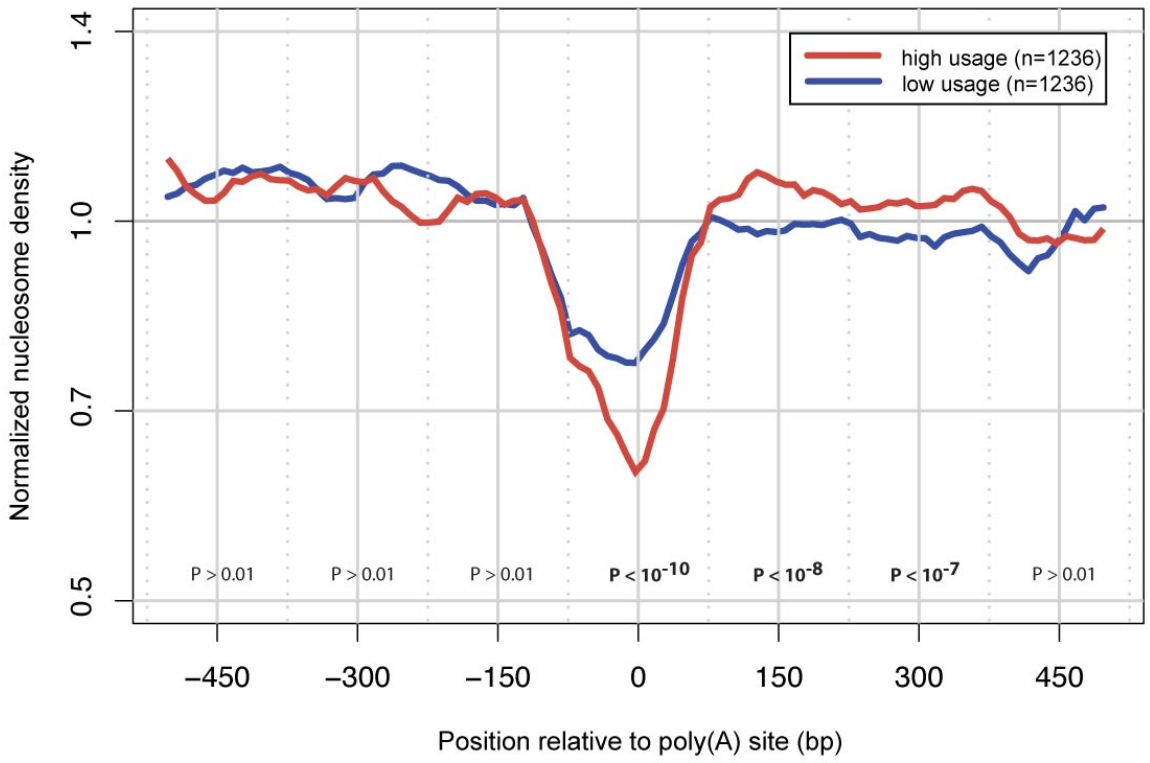

Figure 2: Nucleosome density around human poly(A) sites of low or high usage. Tandem poly(A) sites with either AATAAA or ATTAAA upstream PASs were used. Sites supported by less than 30% of the gene's polyadenylated ESTs were considered to have low usage (blue), and those with greater than 70% of the supporting ESTs were considered to have high usage (red). Density values (Barski et al., 2007 and Schones et al., 2008 data combine) were normalized and smoothed as in Figure 1. Low and high usage poly(A) site sets have been sampled to control for core poly(A) motif scores. Wilcoxon rank sum P-values shown for 150 bp windows centered on indicated positions.

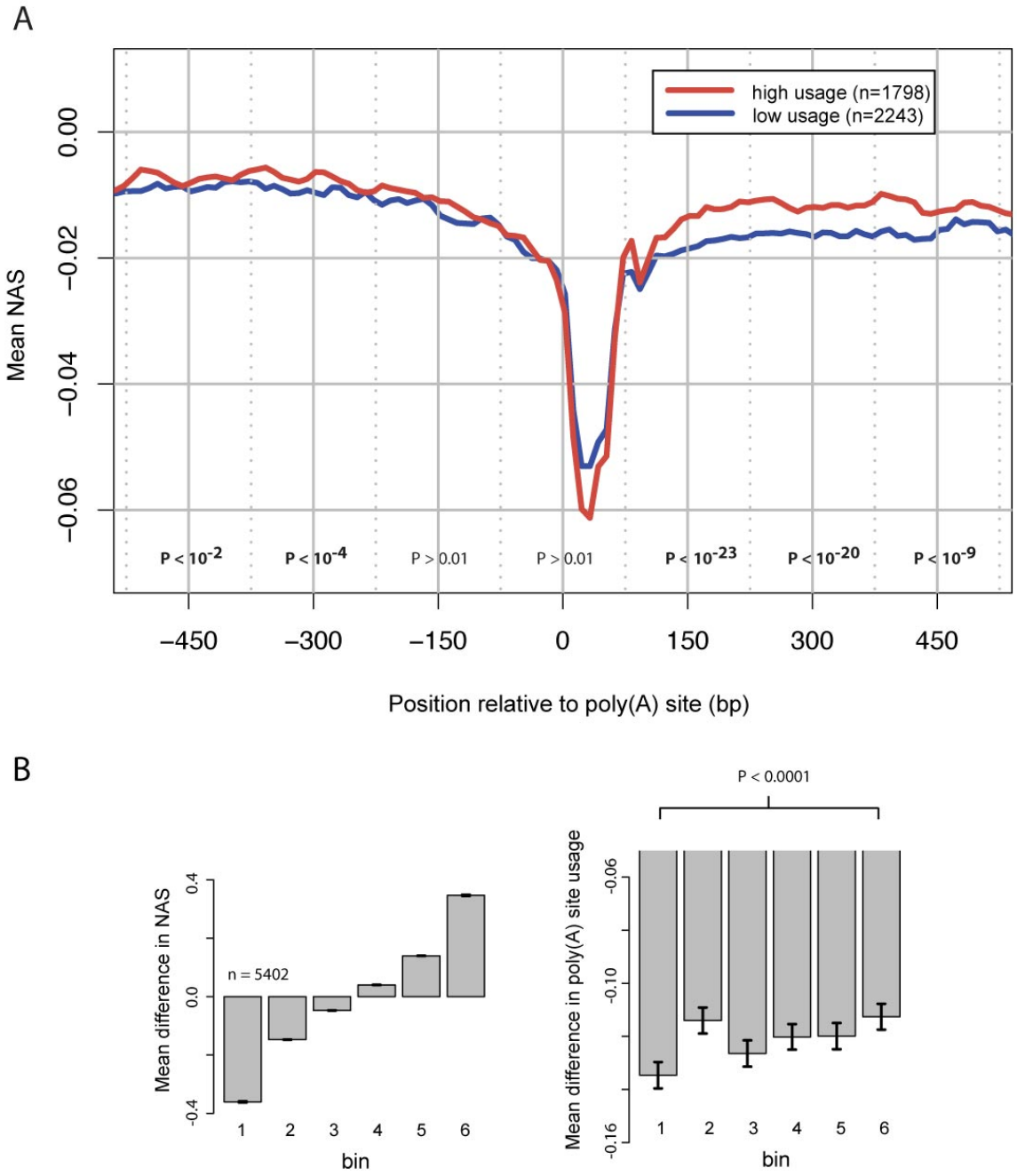

Figure 3: **A**. Mean nucleosome affinity scores (NAS) for positions around human poly(A) sites of low (blue) or high (red) usage as in Figure 2. **B**. Pairs of alternative poly(A) sites, where a pair belongs to the same gene, were binned by their differences in nucleosome affinity scores for the sequence starting at +100 downstream of the poly(A) site (outside the range of known poly(A) signals). These sets of pairs were sampled to control for differences in core poly(A) motif scores (same sampling method as performed in Figure 2; see Methods). Mean differences in nucleosome affinity scores (leftmost panel) and site usage (rightmost panel) are shown for the sampled sets together with standard error bars and Wilcoxon rank sum P-values for extreme bins. To maximize power in this stringent test, poly(A) sites supported by single poly(A)-ESTs were included.

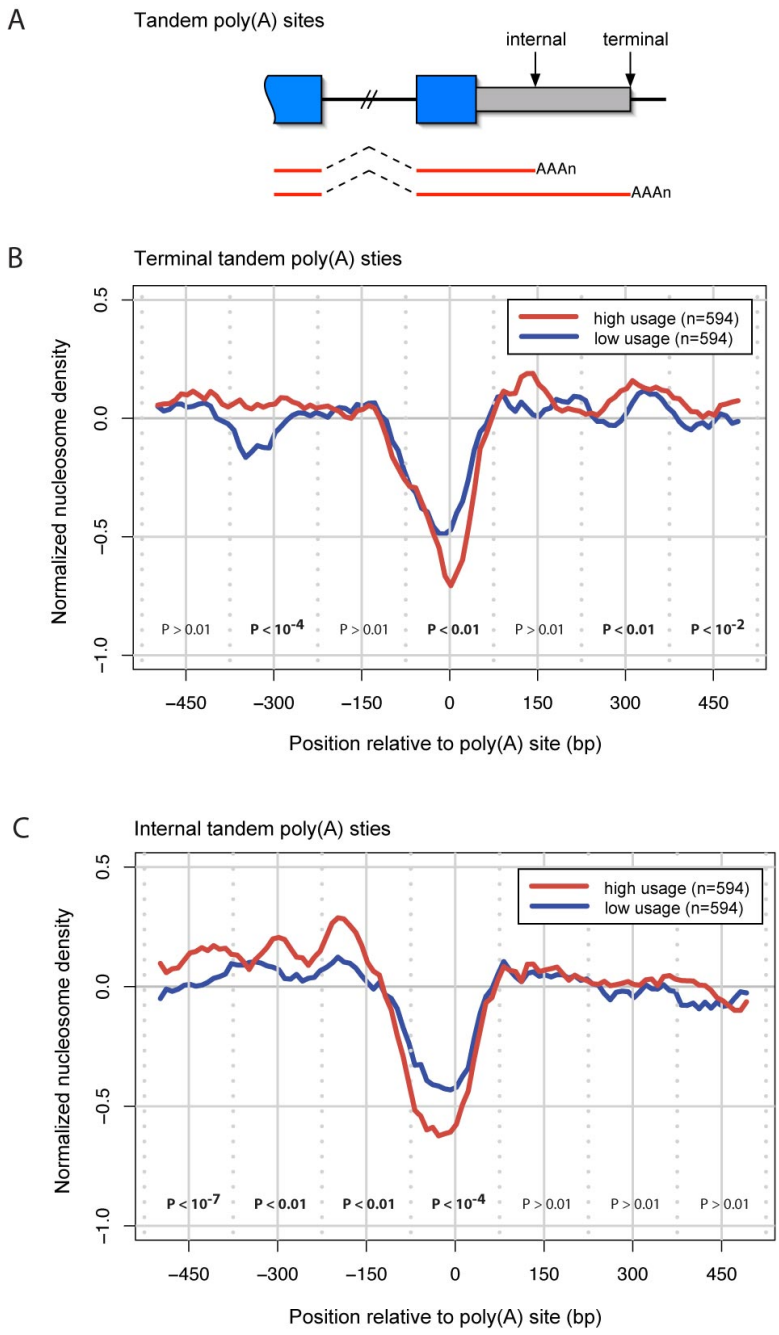

Figure 4: **A**. Schematic representation of tandem poly(A) sites with poly(A) supporting ESTs mapped below (blue boxes: exons; gray box: 3′ UTR). **B**. Nucleosome density around terminal human poly(A) sites. Sites supported by less than 30% of the gene's polyadenylated ESTs were considered to have low usage (blue), and those with greater than 70% of the supporting ESTs were considered to have high usage (red). Density values (data from Schones et al., 2008) were normalized and smoothed as in Figure 1. Low and high usage poly(A) site sets have been sampled to control for core poly(A) motif scores. Wilcoxon rank sum P-values shown for 150 bp windows centered on indicated positions. All sites are supported by at least two ESTs, and filtered to be at least 150 nt from the upstream stop codon and any other alternative poly(A) sites. **C**. Same as (B) only using internal poly(A) sites.

Chapter 4

Concluding Comments

Chapter 4

Concluding Comments

Computational miRNA target prediction is a valuable step in uncovering miRNA functions. An early example of the power of genome-wide computational predictions came from the bantam miRNA in *Drosophila* where initial methods predicted the pro-apoptotic gene *Wrinkled*, often referred to as *head involution defective* (*hid*), as a target and helped to define a role for bantam in stimulating cell proliferation and preventing apoptosis (Brennecke et al., 2003). In addition to identifying critical individual targets, target prediction can reveal enrichment of functional classes of genes which together help to define an *in vivo* role, such as the over-representation of cell cycle genes among miR-16 targets pointing to its function in negative regulation of cell cycle progression (Linsley et al., 2007).

We have made several significant contributions to improved target prediction, presented in detail in Chapter 2. Unlike most commonly used target prediction algorithms which rely heavily on conservation of 3′ UTR seed matches complementary to miRNA seeds (John et al., 2004; Lewis et al., 2005; Krek et al., 2005), our method considers the regulatory contributions of both conserved and non-conserved target sites. To do so, we used several novel targeting determinants uncovered through analyses

of mRNA expression profiles following miRNA overexpression or disruption. These features included the importance of an A or U at the position opposite miRNA base 9, independent of complementarity to the miRNA, potentially revealing a position-specific sequence preference of the silencing complex. We also found evidence that both increased AU content and conservation in the regions flanking seed matches independently are associated with greater down-regulation of the target mRNA. These features suggest a model whereby the sequence context, either through its intrinsic secondary structure or affinity for RNA-binding proteins, influences recognition by the silencing complex. Our work provides a better characterization of the relative contributions of different seed match types, which together with the features mentioned above, was incorporated into a miRNA target ranking scheme, TargetRank. Our method successfully separates responsive and non-responsive miR-155 targets detected in a miR-155 knockout T cell system (Rodriguez et al., 2007), validating the relevance of our rankings *in vivo*. At about the same time as publication of this work, Grimson and coworkers (2007) published an independent study corroborating many of the targeting determinants presented here, and provided luciferase assay data further demonstrating the importance of seed match context. We have provided a web interface (`http://genes.mit.edu/targetrank/`) which we anticipate will serve as a valuable resource of prioritized, predicted miRNA-mRNA interactions. Our studies have also revealed the applicability of miRNA targeting rules to prediction of siRNA off-target effects. Accordingly, we have enabled target prediction of arbitrary input siRNA sequences through our web interface, which we hope will assist a large community of researchers in the design and interpretation of RNAi experiments.

Our work provides evidence that the sequence context of a miRNA seed match is important for its proper recognition. An important area for future work will be in

better understanding how miRNA-mRNA interactions are regulated. A recent study of the miR-430 target *nanos1* in zebrafish embryos demonstrated that expression of Dnd1, which recognizes U-rich sequences, disrupts association of miR-430 with U-rich flanked miR-430 seed matches in the 3′ UTR of the *nanos1* transcript (Kedde et al., 2007). This mechanism enables *nanos1* expression despite high miR-430 levels. In an independent study, an AU-rich-element binding protein, HuR, was shown to relieve the CAT-1 mRNA from miR-122 repression under stress conditions (Bhattacharyya et al., 2006). Characterization of target levels under diverse conditions, together with studies of common 3′ UTR motifs and their co-occurrence with miRNA target sites, will likely shed light on these regulatory relationships. It seems likely that such insights will enable future computational target prediction to consider not only sequence features of the 3′ UTR, but also information regarding expression of regulatory proteins, thus predicting target activity in a given cellular state.

Biochemical approaches to miRNA target identification have been reported recently and offer promise as complementary methods for computational prediction. These methods involved stable expression of tagged Argonaute proteins in *Drosophila* or human cells, overexpression of a particular query miRNA, and subsequent microarray profiling of RNA from am immunoprecipitation (IP) targeting tagged silencing complexes (Easow et al., 2007; Karginov et al., 2007). One challenge of this approach is that the silencing complex may induce deadenylation of its bound mRNA, such that the transcripts targeted for IP enrichment are also subject to degradation. To compensate for these effects, Easow and coworkers (2007) also profiled total RNA levels and calculated a Net IP enrichment value per target, that combines both the RNA level change and the IP enrichment in response to the miRNA. The degree to which a miRNA directs translational repression or mRNA deadenylation appears to

vary across mRNA targets, and these IP methods may be particularly useful in the identification of the target subset that is primarily regulated at the translational level. Refined methods that are able to achieve greater fold enrichment, may provide high-throughput approaches to not only identify targets, but to also better understand how RISC association is differentially regulated across cell types or conditions. These approaches present valuable complementary methods to computational analyses.

We have also provided preliminary evidence for a relationship between chromatin structure and RNA processing at poly(A) sites, discussed in Chapter 3. Using recently published high-throughput Illumina sequencing of nucleosome boundaries detected by MNase digestion of human T cells (Barski et al., 2007; Schones et al., 2008), we observed a striking depletion of nucleosomes across the canonical poly(A) signal, AATAAA, or common variant, ATTAAA. While nucleosome occupancy flanking control hexamers in intergenic regions suggested that the hexamer itself has strong nucleosome positioning characteristics, likely due to its reduced flexibility in wrapping around the nucleosome core, we found evidence that increased nucleosome occupancy in flanking regions, and nucleosome depletion across the poly(A) site itself, are correlated with increased usage of the poly(A) site. Sequence composition clearly influences nucleosome association, and we have constructed a fifth-order Markov model that successfully reproduces features of the nucleosome distribution surrounding transcriptional start sites and poly(A) sites. Application of this affinity scoring method provided additional support for a relationship between local nucleosome positioning and poly(A) site usage.

A large body of work documents the influence of chromatin on transcription (reviewed by Li et al. (2007)), but its potential role in co-transcriptional RNA processing

is mostly unexplored. Our preliminary results regarding the association between nucleosome positioning and poly(A) site usage could benefit from direct experimental tests of this relationship. Relative poly(A) site usage could be tested in a tandem poly(A) site reporter, with poly(A) sites having identical known RNA-level motifs but different chromatin contexts, as a result of sequences with different nucleosome affinities at more distant positions (e.g. +/- 100 bp of the poly(A) signal) to avoid disruption of the poly(A) site itself. For such studies, accurate detection of poly(A) site usage, nucleosome positioning, and Pol II accumulation (pausing) are needed. Such a system would be extremely valuable in helping to sort out possible mechanisms suggested by computational analyses.

In the work presented here, we have explored two types of RNA signals, miRNA seed matches and poly(A) signals, and characterized how their sequence contexts can influence their regulation. While considered separately, exploration of the connections between miRNA targeting capacity and alternative 3′ UTRs resulting from differential poly(A) site usage poses an interesting area for future work. Computational studies using ESTs demonstrated that miRNA-targeted isoforms are present at reduced levels in tissues expressing the corresponding miRNA (Legendre et al., 2006). Stark and coworkers presented evidence that expression of a muscle-specific isoform of *Tropomyosin 1* in *Drosophila* corresponded to a switch from a different 3′ UTR containing miR-1 seed matches, a muscle-specific miRNA (Stark et al., 2005). They proposed a model whereby the negative effects of possible aberrant splicing to the alternative, non-muscle isoform are suppressed by the interaction of miR-1 with target sites in the non-muscle transcript. One challenge with these analyses is in distinguishing between the choice in poly(A) site and miRNA-driven isoform down-regulation. As with the *Tropomyosin 1* example, these events are likely coordinated

and reinforce a common signal. Global analyses of alternative 3′ UTR usage in the context of miRNA targeting will likely reveal its wide-spread importance.

We have applied a unified statistical approach that involves controlling for potentially confounding variables through sampling, which has proven useful in resolving contextual features from high-throughput data sets. These approaches can be readily applied to other regulatory signals, such as splicing motifs in RNA or transcription factor binding sites in DNA, and it would be interesting to investigate the effects of sequence context on these diverse recognition processes using similar approaches. Improved understanding of the influence of sequence context on regulatory motifs will not only assist in better prediction of functional sites, but will also be valuable in identifying mutations that may lead to misregulation and underlie disease phenotypes.

References

A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, May 2007.

S. N. Bhattacharyya, R. Habermacher, U. Martine, E. I. Closs, and W. Filipowicz. Relief of microRNA-mediated translational repression in human cells subjected to stress. *Cell*, 125(6):1111–24, Jun 2006.

J. Brennecke, D. R. Hipfner, A. Stark, R. B. Russell, and S. M. Cohen. bantam encodes a developmentally regulated microRNA that controls cell proliferation and regulates the proapoptotic gene hid in Drosophila. *Cell*, 113(1):25–36, Apr 2003.

G. Easow, A. A. Teleman, and S. M. Cohen. Isolation of microRNA targets by miRNP immunopurification. *RNA*, 13(8):1198–204, Aug 2007.

A. Grimson, K. K.-H. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell*, 27(1):91–105, Jul 2007.

B. John, A. J. Enright, A. Aravin, T. Tuschl, C. Sander, and D. S. Marks. Human microRNA targets. *PLoS Biol*, 2(11):e363, Nov 2004.

F. Karginov, C. Conaco, Z. Xuan, B. Schmidt, J. Parker, G. Mandel, and G. Hannon. A biochemical approach to identifying microRNA targets. *Proc Natl Acad Sci USA*, Nov 2007.

M. Kedde, M. J. Strasser, B. Boldajipour, J. A. F. O. Vrielink, K. Slanchev, C. le Sage, R. Nagel, P. M. Voorhoeve, J. van Duijse, U. A. Ørom, A. H. Lund, A. Perrakis, E. Raz, and R. Agami. RNA-binding protein Dnd1 inhibits microRNA access to target mRNA. *Cell*, 131(7):1273–86, Dec 2007.

A. Krek, D. Grün, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky. Combinatorial microRNA target predictions. *Nat Genet*, 37(5):495–500, May 2005.

M. Legendre, W. Ritchie, F. Lopez, and D. Gautheret. Differential repression of alternative transcripts: a screen for miRNA targets. *PLoS Comput Biol*, 2(5):e43, May 2006.

B. P. Lewis, C. B. Burge, and D. P. Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell*, 120(1):15–20, Jan 2005.

B. Li, M. Carey, and J. L. Workman. The role of chromatin during transcription. *Cell*, 128(4):707–19, Feb 2007.

P. S. Linsley, J. Schelter, J. Burchard, M. Kibukawa, M. M. Martin, S. R. Bartz, J. M. Johnson, J. M. Cummins, C. K. Raymond, H. Dai, N. Chau, M. Cleary, A. L. Jackson, M. Carleton, and L. Lim. Transcripts targeted by the microRNA-16 family cooperatively regulate cell cycle progression. *Mol Cell Biol*, 27(6):2240–52, Mar 2007.

A. Rodriguez, E. Vigorito, S. Clare, M. V. Warren, P. Couttet, D. R. Soond, S. van Dongen, R. J. Grocock, P. P. Das, E. A. Miska, D. Vetrie, K. Okkenhaug, A. J. Enright, G. Dougan, M. Turner, and A. Bradley. Requirement of bic/microRNA-155 for normal immune function. *Science*, 316(5824):608–11, Apr 2007.

D. E. Schones, K. Cui, S. Cuddapah, T.-Y. Roh, A. Barski, Z. Wang, G. Wei, and K. Zhao. Dynamic regulation of nucleosome positioning in the human genome. *Cell*, 132(5):887–98, Mar 2008.

A. Stark, J. Brennecke, N. Bushati, R. B. Russell, and S. M. Cohen. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell*, 123(6):1133–46, Dec 2005.

Appendix 1

—

Patterns of Intron Gain and Loss in Fungi

Open access, freely available online PLOS BIOLOGY

Patterns of Intron Gain and Loss in Fungi

Cydney B. Nielsen[1,2☯], Brad Friedman[1,3☯], Bruce Birren[2], Christopher B. Burge[1*], James E. Galagan[2*]

1 Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America, **2** The Broad Institute of Massachusetts Institute of Technology and Harvard University, Cambridge, Massachusetts, United States of America, **3** Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts, United States of America

Little is known about the patterns of intron gain and loss or the relative contributions of these two processes to gene evolution. To investigate the dynamics of intron evolution, we analyzed orthologous genes from four filamentous fungal genomes and determined the pattern of intron conservation. We developed a probabilistic model to estimate the most likely rates of intron gain and loss giving rise to these observed conservation patterns. Our data reveal the surprising importance of intron gain. Between about 150 and 250 gains and between 150 and 350 losses were inferred in each lineage. We discuss one gene in particular (encoding 1-phosphoribosyl-5-pyrophosphate synthetase) that displays an unusually high rate of intron gain in multiple lineages. It has been recognized that introns are biased towards the 5′ ends of genes in intron-poor genomes but are evenly distributed in intron-rich genomes. Current models attribute this bias to 3′ intron loss through a poly-adenosine-primed reverse transcription mechanism. Contrary to standard models, we find no increased frequency of intron loss toward the 3′ ends of genes. Thus, recent intron dynamics do not support a model whereby 5′ intron positional bias is generated solely by 3′-biased intron loss.

Citation: Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE (2004) Patterns of intron gain and loss in fungi. PLoS Biol 2(12): e422.

Introduction

Over a quarter of a century after the discovery of introns, fundamental questions about their function and evolutionary origins remain unanswered. Although intron density differs radically between organisms, the mechanisms by which introns are inserted and deleted from gene loci are not well understood. A correlation has been observed between intron density and positional bias (Mourier and Jeffares 2003). Introns are evenly distributed within the coding sequence of genes in intron-rich organisms, but are biased toward the 5′ ends of genes in intron-poor organisms. This bias is particularly pronounced in the yeast *Saccharomyces cerevisiae*. It has been suggested that both the paucity and positional bias of introns in yeast may be due to intron loss through a mechanism of homologous recombination of spliced messages reverse-transcribed from the 3′ poly-adenylated tail (Fink 1987). This reverse transcription mechanism was first demonstrated in experiments with intron-containing Ty elements in yeast (Boeke et al. 1985). More recently, Mourier and Jeffares (2003) concluded that homologous recombination of cDNAs is the simplest explanation for the positional bias observed in all intron-poor eukaryotes. However, few data exist concerning the actual mechanisms and dynamics of intron evolution.

Fungal genomes are in many ways ideal for exploring questions of intron evolution. The fundamental aspects of intron biology are shared between fungi and other eukaryotes, making fungi appropriate model organisms for intron study. They are gene dense with relatively simple gene structures compared with plants and animals, making gene prediction more accurate. Fungi also display a wide diversity of gene structures, ranging from far less than one intron per gene for *S. cerevisiae,* to approximately 1–2 introns per gene on average for many recently sequenced ascomycetes (including the organisms in this study), to roughly seven introns per gene on average for some basidiomycetes (e.g., *Cryptococcus*). Finally, fungi display a strong 5′ bias in intron positions, enabling us to investigate the processes underlying this phenomenon.

In principle, a 5′ intron bias could arise through various combinations of intron gain and loss, and a complete understanding of intron positional bias requires an assessment of the contributions of both of these processes. A number of studies demonstrate the occurrence of intron gain and loss in individual genes or gene families. Logsdon et al. (1995) offered early examples of well-supported intron gain by comparing triose-phosphate isomerase genes from diverse eukaryotes and demonstrated that numerous introns could be most parsimoniously explained by a single gain with no subsequent losses. O'Neill et al. (1998) later provided evidence for de novo intron insertion into the otherwise intron-less mammalian sex-determining gene *SRY*. Evidence for the occurrence of multiple independent intron losses has also been reported in studies such as those by Robertson (2000), who inferred gain and loss events in a family of chemoreceptors in *Caenorhabditis elegans*.

More recently, a number of genome-wide studies of intron dynamics have been conducted. Roy et al. (2003) described genome-wide comparisons between human and mouse (with *Fugu* as an outgroup) and between mouse and rat (with human as an outgroup), and observed a sparseness of intron loss and complete absence of intron gain in these closely related organisms. On the other hand, Rogozin et al. (2003)

Received June 22, 2004; Accepted October 5, 2004; Published November 30, 2004
DOI: 10.1371/journal.pbio.0020422

Abbreviations: PRPP, 1-phosphoribosyl-5-pyrophosphate

Academic Editor: Ken H. Wolfe, University of Dublin

*To whom correspondence should be addressed. E-mail:cburge@mit.edu (CBB), jgalag@mit.edu (JEG)

☯These authors contributed equally to this work.

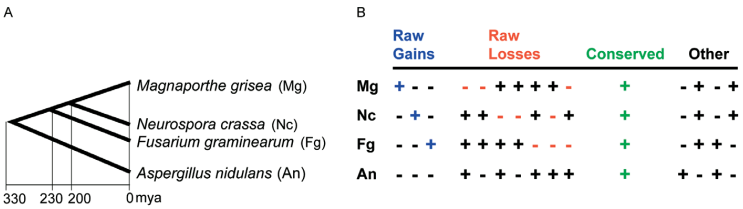

Figure 1. Phylogenetic Tree and Intron Conservation Patterns

(A) Phylogenetic tree of the four fungal organisms studied *(M. grisea, N. crassa, F. graminearum,* and *A. nidulans)* with estimated time scale in millions of years ago. The rooted organismal tree was constructed using an unweighted pair group method using arithmetic averages based on a concatenated alignment of 2,073 orthologous gene sets. Estimated dates of divergence from Taylor et al. (1999), Berbee and Taylor (2000), and Heckman et al. (2001).(B) Classification of intron presence (+) and absence (–) patterns across the four fungal species. A blue "+" indicates a raw intron gain in the corresponding organism, a red "–" indicates a raw intron loss in the corresponding organism, a green "+" indicates a conserved intron, and all other introns are indicated in black.

DOI: 10.1371/journal.pbio.0020422.g001

observed an abundance of lineage-specific intron loss and gain when analyzing clusters of orthologous genes in deeply branching eukaryotes. Similarly, Qiu et al. (2004) analyzed ten protein families in distantly related eukaryotes, with a single prokaryotic outgroup, and obtained evidence that extant introns are predominantly the result of intron gains. In search of clues to understand the mechanism of intron gain, Fedorov et al. (2003) aligned introns from various eukaryotes, and Coghlan and Wolfe (2004) applied a similar approach in a comparative study of nematodes. None of these studies addressed the positional bias of intron gain and loss events. Here we report the results of a genome-wide comparative analysis of intron evolution in organisms that have a strong 5′ bias in intron location and are at an appropriate evolutionary distance to reveal positional trends in intron gain and loss.

Results

To investigate the roles of both gain and loss in intron evolution, we compared the genomes of four recently sequenced fungi spanning at least 330 million years of evolution (Taylor et al. 1999; Berbee and Taylor 2000; Heckman et al. 2001) (Figure 1): *Aspergillus nidulans*, *Fusarium graminearum*, *Magnaporthe grisea*, and *Neurospora crassa*. Ortholog sets composed of one gene from each of the four genomes were identified as pairwise best bidirectional BLAST hits satisfying stringent overlap criteria. Orthologs in each set were subsequently aligned, and the locations of introns were marked. These intron positions (regions of the multiple sequence alignment containing an intron in at least one of the four sequences) were subjected to rigorous alignment quality filtering to eliminate alignment and annotation errors (Figure 2A). To set the filtering thresholds, we manually classified ten residue alignment windows on either side of 181 randomly selected intron positions as "clearly homologous," "possibly homologous," or "non-homologous." Requiring 30% identity and 50% similarity in these windows captured 92% of the clearly homologous positions, 29% of the possibly homologous positions, and only 2% of the non-homologous positions (Figure 2B). Passing intron positions were split into five quintiles according to their relative position within the annotated coding sequence.

Genome-Wide Characterization of Intron Conservation

We applied our analysis protocol to 2,073 putative ortholog sets that included 9,352 intron positions. Of these initial intron positions, 5,811 were removed because of low conservation surrounding the intron, or because of an adjacent gap, or both. It is possible that some of the positions neighboring gaps may in fact reflect intron gain or loss events that occurred simultaneously with coding sequence insertion or deletion (Llopart et al. 2002). However, removing these positions did not significantly impact our results, as the number of positions adjacent to gaps was only about one-tenth of the number of positions that passed the quality filter, and the removal of these introns did not alter the apparent positional bias of the overall distribution (Figure 3). An additional 92 introns had nearby introns with insufficient conservation between the two introns and were thus also rejected.

In the end, a total of 3,450 intron positions (roughly 37% of intron positions considered) passed the quality filter. The complete set of aligned orthologs with passing and failing intron positions is provided in Table S1. These data constitute a genome-wide survey of high-confidence aligned intron positions and their patterns of conservation over at least 330 million years of evolution.

An example of an alignment of putative orthologs with three passing intron positions is shown in Figure 4A. In each passing intron position (black-edged rectangles), individual introns are labeled according to the classes previously outlined in Figure 1B. One intron position is conserved across all four species (green rectangle), one is a raw gain in *N. crassa* (blue box), and the third is present only in *A. nidulans,* and, because of the ambiguity in inferring gain or loss in this case, is classified as "Other" (black-edged gray rectangle). Examining the region around the one raw gained intron in *N. crassa* at the nucleotide level (Figure 4B) reveals a clean insertion of the intron sequence within a highly conserved region. The gained intron has consensus terminal dinucleotides (GT…AG) and a putative branch point sequence that matches the yeast consensus (TACTAAC) at six of seven positions. In addition, this set of orthologs contained one poorly aligned intron position (Figure 4A, unedged gray rectangle) that was excluded by our filters. All three passing positions (black-edged rectangles) display high amino acid sequence conservation on both sides flanking the intron, supporting the correctness of the alignment. In contrast, the failing intron position (unedged gray rectangle) is adjacent to a region of the alignment that lacks significant conservation. The 3′ flank of this intron position displays considerable variation, especially with respect to the *M. grisea* gene, which was predicted to have a much longer 3′ coding region. In such an alignment region, it is difficult to distinguish true differences in intron conservation from potential annotation or alignment errors. Our filtering process thus eliminated this position from further analysis.

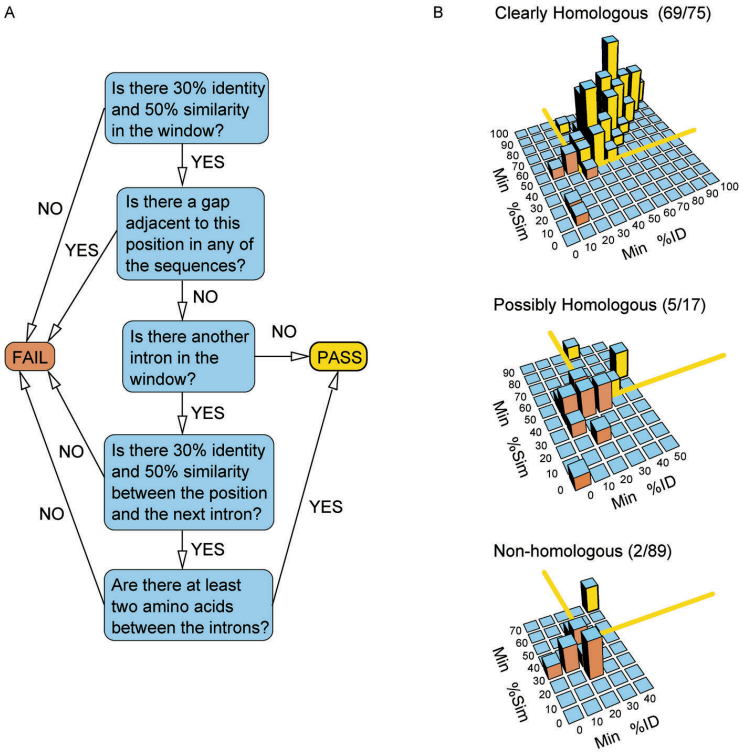

Figure 2. Alignment Filtering Protocol

(A) Schematic of filtering protocol applied to a ten-residue window on each side of every intron position. If either side failed the filter, the position was discarded.

(B) Distributions of minimum percent identity and similarity in ten-residue windows around 181 randomly selected intron positions, for three manual classifications. The minima were taken between the left and right windows. The yellow lines indicate the chosen thresholds of at least 50% similarity and 30% identity, and bars are colored yellow if they fall above the thresholds (pass) or orange if they fall below the thresholds (fail). Parentheses indicate the number of introns in each class that pass the cutoff and the total number of introns in that class. The five lowest-percent identity and similarity bars, containing 77 positions, in the "non-homologous" plot are omitted so as to not obscure the rest of the histogram.

DOI: 10.1371/journal.pbio.0020422.g002

Calculation of Raw Gains and Losses

We calculated "raw gains" and "raw losses" by positional quintile for each organism other than the outgroup, *A. nidulans*. We defined raw gains as those introns present in only a single organism (see Figure 1B). We defined raw losses as those introns that are absent in the organism in question, present in some other descendant of the organism's parent (a "sibling"), and present in some non-descendant of the parent (a "cousin") (Figure 1B). Intron positions are considered conserved if present across all four organisms. Patterns of intron presence and absence that are not captured by the above definitions were excluded from the raw counts because of the ambiguity in inferring intron gain or loss events in such cases (marked as "Other" in Figure 1B).

Probabilistic Model of Intron Gain and Loss

Raw gain and loss counts are based on parsimony and may differ somewhat from the true number of gain and loss events. The set of raw gains may include introns that were lost in multiple lineages, thus overcounting the true number of gains in a given lineage. Similarly, the set of raw losses excludes introns lost in the given organism and also lost in all cousins or siblings (marked as "Other" in Figure 1B).

We used a probabilistic model to correct for these inaccuracies. Our model assumes that all loss and gain events occur independently and uniformly within each quintile. In particular, we assume Dollo's postulate (Dollo 1893): any introns that align to the same position must have a common ancestor (no "double gains"), as in Nei and Kumar (2000) and Rogozin et al. (2003). Our method differs from the Dollo parsimony method described in Farris (1977) and applied in Rogozin et al. (2003) in that we do not artificially minimize loss events by assuming that gains occurred at the latest possible point in evolution. It also differs in that we allow different branches of the phylogenetic tree to have different rates of loss and gain. We applied our method separately to each of the five positional quintiles for each organism other than the outgroup, *A. nidulans*.

First we estimate two types of intron loss rates. The organismal loss rate, q, is calculated by dividing the number of raw losses in an organism by the total number of introns present in at least one sibling and at least one cousin. This represents the fraction of introns in the parent that did not survive to the present day in that organism. For instance, the organismal loss rate in *F. graminearum* is given by

$$q = (AM + AN + AMN)/((AM + AN + AMN) + (AFM + AFN + AFMN)) \quad (1)$$

where AM, for example, represents the number of intron positions with an intron present in *A. nidulans* (A) and *M. grisea* (M) but absent from *F. graminearum* (F) and *N. crassa* (N).

The sibling loss rate, r, is defined for a given organism as the fraction of introns in the parent that did not survive in any sibling. We define "sibling raw losses" for an organism as the number of introns that are present in the organism and at least one cousin but in no sibling. This quantity is then

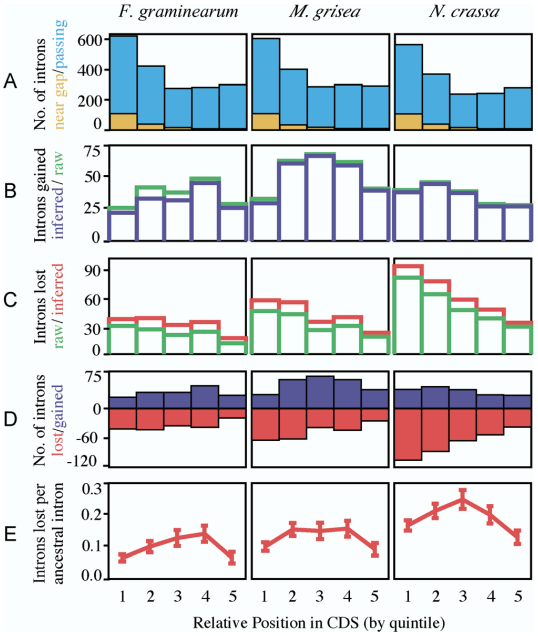

Figure 3. Positional Biases in Intron Gain and Loss

Relative intron positions were defined as the number of bases in the coding sequence upstream of the intron divided by the total length of the coding sequence. These relative positions were binned into five categories (quintiles), each representing one-fifth of the coding sequence length (quintiles numbered 1–5 on the x-axis).

(A) Introns passing quality filter (light blue, back) and introns adjacent to gaps in the protein alignment that were removed by our quality filter (orange, front).

(B) Raw and inferred gains. Raw gains (green, back) are those introns present in exactly one organism (excluding the outgroup, *A. nidulans*). Inferred gains (blue, front) are corrected for the estimated number of cases that arose by other combinations of gain and loss events. Inferred gains are thus slightly lower than raw gains.

(C) Raw and inferred losses. Raw losses (green, front) are those introns absent in the organism in question but present in at least one of its siblings (descendants of its parent in the phylogenetic tree) and one of its cousins (non-descendants of its parent). Inferred losses (red, back) are corrected for the estimated number of introns lost along multiple lineages, or gained and then lost. Inferred losses are thus slightly higher than raw losses.

(D) Number of introns gained (blue) and lost (red) since last common ancestor (losses shown as negative numbers).

(E) Intron loss rate at each position since last common ancestor (introns lost per ancestral intron). Error bars represent binomial standard deviation.

DOI: 10.1371/journal.pbio.0020422.g003

divided by the number of introns present in that organism and at least one cousin to give the sibling loss rate. For example, the sibling loss rate for *F. graminearum* is given by

$$r = (AF)/(AF + AFM + AFN + AFMN). \quad (2)$$

We next correct the raw gains for each organism. Raw gains include some introns that were in fact lost in all but one lineage. We use the loss rates to calculate the expected number of these multiple losses, m, and subtract this quantity from the raw gains to obtain "inferred gains." To calculate m we first count B_0, the number of introns conserved in the organism and at least one sibling, but in no cousin. The quantities m and B_0 are related through the variable n_0, the number of introns present in an organism's parent but not in any cousin, by the equations

$$m = n_0 r(1 - q) \quad (3)$$

and

$$B_0 = n_0(1 - r)(1 - q). \quad (4)$$

This follows from our assumption of independent gains and losses. Thus, we can calculate the expected number of multiple losses as

$$m = B_0 r/(1 - r). \quad (5)$$

We use the loss rates to estimate the number of introns in each organism's parent. To do so, we estimate separately the number of parental introns present in at least one cousin n_1, and the number not present in any cousin n_0 (introduced above). To estimate the size of the set of parental introns present in at least one cousin, we first count the subset of these introns that are presently observable. An intron is in this set if it is present in at least one cousin and at least one sibling, or is present in at least one cousin and in the organism in question. We call this number of introns B_1. By the assumption that gains and losses are independent, we have

$$B_1 = n_1(1 - qr). \quad (6)$$

Using this relation and the one in equation 4 above, we calculate the number of introns in the phylogenetic parent as

$$n_{\text{total}} = n_1 + n_0 = \frac{B_1}{(1 - qr)} + \frac{B_0}{(1 - q)(1 - r)}. \quad (7)$$

Finally, we correct raw losses. Our definition of raw losses undercounts the true number by omitting those introns not conserved in at least one cousin and at least one sibling. Taking *F. graminearum* as an example, the true number of losses would also include some introns conserved in the patterns *A, M, N,* and *MN*. We calculate the number of inferred losses as $n_{\text{total}}q$.

This method can be extended to any phylogenetic tree and to any organism with at least one cousin.

Abundance of Intron Gains

One immediate conclusion stemming from our analysis is the importance of intron gain. A summary of all raw and inferred gains and losses is shown in Figure 3. Substantial numbers of gained introns were observed in all three organisms—more than 100 independent inferred gains in each lineage, with over 200 in *M. grisea* (Figure 3B). The total numbers of gains that have occurred in each genome are likely to be substantially higher, since only predicted orthologs in all four species were considered, and roughly a third of the introns in these genes passed our quality filters. Differences in intron dynamics between lineages are also apparent, with the numbers of gained and lost introns approximately balanced in *M. grisea* and *F. graminearum*, but with roughly twice as many losses as gains in *N. crassa* (Figure 3D). It is thus apparent from these data that the process of intron gain plays a significant role in intron evolution.

Figure 4. Example Ortholog Alignment

(A) Alignment of protein sequences for orthologs MG04228, NCU05623, FG06415, and AN1892 with intron characters inserted. "0," "1," and "2" indicate the phase of an intron. A black-edged rectangle indicates an intron position passing our quality filters; an unedged gray rectangle indicates an intron position that was removed by our filter. The green rectangle indicates conserved introns, the blue box marks a raw intron gain, and the gray boxes within black-edged rectangles highlight all other introns. The consensus (bottom) line characters are as follows: asterisk, identical residue in all four sequences; colon, similar residue; and period, neutral residue.

(B) Nucleotide alignment of the region flanking the gained intron in (A). Putative 5′ and 3′ splice sites and a branch point sequence are highlighted in blue.

DOI: 10.1371/journal.pbio.0020422.g004

1-Phosphoribosyl-5-Pyrophosphate Synthetase Genes Display Lineage-Specific Increases in Intron Gain Rate

A striking example of intron gain occurs in a set of putative orthologous 1-phosphoribosyl-5-pyrophosphate (PRPP) synthetase genes. These genes encode a widely conserved protein that catalyzes the production of PRPP, a precursor in the nucleotide biosynthesis pathway. In contrast to the majority of orthologs that displayed fewer than two gained introns, the set of PRPP synthetase genes displayed a total of 22 raw gains (Figure 5A, blue boxes) that passed our alignment quality filters: six in *N. crassa,* 14 in *M. grisea,* and two in *F. graminearum.* The number of raw gains in the PRPP synthetase genes in *M. grisea* and *N. crassa* was significantly higher ($p < 3 \times 10^{-22}$ and $p < 4 \times 10^{-9}$, respectively) than the average for other genes analyzed, resulting in unusually large numbers of introns in these genes (Figure 5B). In comparison, the numbers of introns in PRPP synthetase genes in available animal genomes were within the typical range for the respective organisms, e.g., five in *C. elegans,* and six in fruitfly, human, mouse, rat, and *Fugu.* Thus the rate of intron gain for the PRPP synthetase gene in some fungi is unusually high. This gene represents an extreme example of the impact of intron gain and illustrates the variability of gain rates in different lineages.

Fungal Introns Display Phase Bias, but Lack Observable Sequence Preference

For each in-group lineage *(M. grisea, N. crassa, F. graminearum),* we determined the frequency of phase 0, 1, and 2 introns in the set of all intron positions (Table 1). In contrast to recent reports based on a much smaller sample size indicating that phase frequencies for extant fungal introns do not differ significantly from a uniform distribution (Qiu et al. 2004), our genome-wide dataset demonstrates a clear bias for phase 0 introns in each of the three fungal in-group lineages examined ($p < 4 \times 10^{-9}$ for *N. crassa* and $p < 1 \times 10^{-12}$ for *M. grisea* and *F. graminearum;* in Table 1, "all passing," and similar biases were seen in the unfiltered set). The phase distributions of raw gains and raw losses for each of the three organisms are not significantly different from a uniform distribution at $p < 0.01$; however, the datasets for these subclasses were much smaller (Table 1). Finally, we examined the exon sequences flanking gained introns, and observed no clear sequence bias (Table 2).

Absence of 3′ Bias in Intron Losses

To determine whether the pattern of intron loss in these fungi might account for the observed bias in intron position, we examined the pattern of loss as a function of position within the gene (see Figure 3E). Contrary to what would be expected if intron loss primarily involved homologous recombination of poly-adenosine-primed reverse transcripts, the rate of intron loss tends to be lower, rather than higher, at the 3′ ends of genes. Moreover, the highest rates of intron loss occur in the middles of genes in all three organisms. We found no evidence that this pattern was affected by our filtering methods. These findings suggest either other mutational mechanisms (e.g., reverse transcription primed internally) or the presence of selective pressure to preferentially conserve introns near the 5′ and 3′ ends of genes.

Discussion

We developed a system that automatically identifies evolutionary and positional patterns of intron conservation on a genome-wide scale. The core of the system is a process for stringently filtering alignments of orthologous genes to

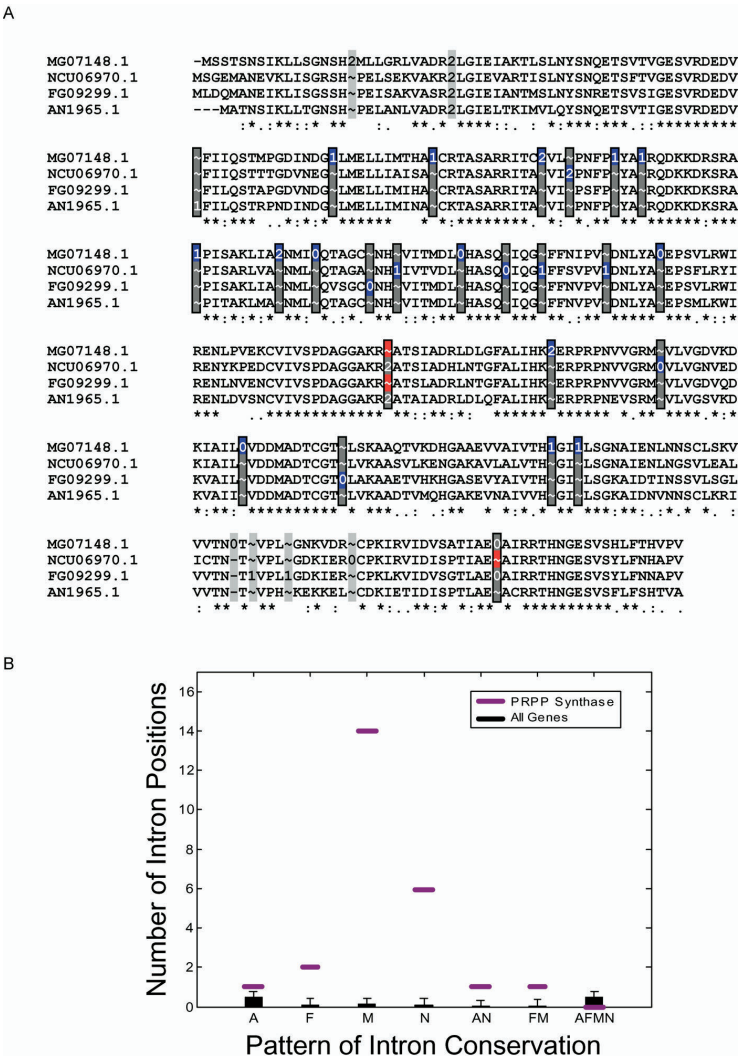

Figure 5. Intron Conservation in the PRPP Synthetase Gene

(A) Alignment of PRPP synthetase putative orthologs MG07148, NCU06970, FG09299, and AN1965. A black-edged rectangle indicates an intron position passing our quality filters, whereas an unedged gray rectangle indicates an intron position that was removed by our filter. Blue boxes mark raw intron gains, red boxes indicate raw intron losses, and gray boxes within black-edged rectangles highlight all other introns. We manually corrected an annotation error in the first intron of the last row of the alignment.

(B) Phylogenetic conservation pattern of introns in the PRPP sythetase gene. Each passing intron position was categorized as being present in *A. nidulans* (A), *F. graminearum* (F), *M. grisea* (M), *N. crassa* (N), *A. nidulans* and *N. crassa* (AN), *F. graminearum* and *M. grisea* (FM), or all four organisms (AFMN). There are no passing cases of conservation in three or four species. The number of introns in each category is shown with a purple line. The black error bar plot shows the mean and standard deviation for each category for all 2,008 ortholog sets after fitting to a Poisson distribution (see Materials and Methods). The number of introns in *M. grisea* and *N. crassa* is significantly higher, at the $p < 1 \times 10^{-9}$ level.

DOI: 10.1371/journal.pbio.0020422.g005

exclude potential annotation or alignment errors. The result of the filtering process is a high-confidence set of aligned intron positions. Differences in intron conservation at each individual position can be characterized as gains or losses (or ambiguous) based on parsimony. However, this does not accurately account for the possibility of multiple gain or loss events. We have developed a probabilistic model that allows for multiple events, providing a corrected estimate of the total number of gains and losses within the dataset. Our probabilistic method allows for a more accurate assessment of rates of gain and loss. In our dataset, allowing for multiple events results in only modest corrections to the rates estimated using parsimony.

Our analysis demonstrates a significant role for intron gain over the past few hundred million years in the fungi analyzed. Previous analyses of specific gene families have provided evidence of specific instances of gained introns (Logsdon et al. 1998; Robertson 2000; Hartung et al. 2002; Qiu et al. 2004). However, the relative importance of intron gain versus loss is not well understood. Recent large-scale analyses have suggested that intron gain may play a predominant role in shaping gene structures (Qiu et al. 2004), although lineage-specific differences are apparent (Rogozin et al. 2003). In particular, intron gain appears to occur rarely if at all in mammalian genes (Roy et al. 2003). Our data suggest that intron gain is a significant driving force in the evolution of genes in fungi. In *F. graminearum* and *M. grisea* the number of introns gained was on par with the number lost and similar in magnitude to the number of introns gained in *N. crassa*.

The mechanisms underlying intron gain are not known. We analyzed the set of predicted intron gains for possible signatures that might shed light on this process. No statisti-

Table 1. Intron Phase Distribution for Filtering and Conservation Classes

Intron Class	Organism	Phase 0	Phase 1	Phase 2	Total
All unfiltered	Nc[a]	1,465 (38.9%)	1,293 (34.3%)	1,007 (26.7%)	3,765
	Mg[a]	1,707 (40.6%)	1,380 (32.8%)	1,118 (26.6%)	4,205
	Fg[a]	1,755 (39.9%)	1,483 (33.7%)	1,165 (26.5%)	4,403
All passing	Nc[a]	672 (39.9%)	547 (32.5%)	465 (27.6%)	1,684
	Mg[a]	765 (40.8%)	602 (32.2%)	505 (27.0%)	1,872
	Fg[a]	765 (40.6%)	614 (32.6%)	507 (26.9%)	1,886
Passing conserved	Nc[a,b]	420 (42.9%)	302 (30.9%)	255 (26.1%)	977
	Mg[a,b]	420 (42.9%)	302 (30.9%)	255 (26.1%)	977
	Fg[a,b]	420 (42.9%)	302 (30.9%)	255 (26.1%)	977
Passing raw gains	Nc	60 (33.7%)	71 (40.0%)	47 (26.4%)	178
	Mg	106 (41.2%)	80 (31.1%)	71 (27.6%)	257
	Fg	69 (37.7%)	64 (35.0%)	50 (27.3%)	183
Passing raw losses	Nc	151 (37.1%)	130 (31.9%)	126 (31.0%)	407
	Mg	177 (38.7%)	154 (33.7%)	126 (27.6%)	457
	Fg	192 (38.3%)	168 (33.5%)	141 (28.1%)	501

[a] Significantly different from uniform distribution, at $p < 0.01$
[b] One intron removed because of phase discrepancy across *N. crassa, M. grisea,* and *F. graminearum.*
Fg, *F. graminearum;* Nc, *N. crassa*; Mg, *M. grisea.*
DOI: 10.1371/journal.pbio.0020422.t001

cally significant bias was detected in the positions of gained introns along the coding sequence (see Figure 3 data not shown). Similarly, no preferred insertion site sequence was detectable (Table 2), and no significant phase bias for gained introns was observed (see Table 1). The lack of an insertion site preference and absence of significant phase bias for gained introns in fungi is consistent with previous investigations and may set fungi apart from other organisms (Qiu et al. 2004).

Our data further indicate that intron gain can vary substantially between different gene families in a lineage-specific fashion. The PRPP synthetase gene is a particularly striking example, exhibiting significant increases in gained introns in two of the four lineages investigated. Moreover, the paucity of intron positions shared between *N. crassa* and *M. grisea* suggests the possibility of independent increases in gain rate in the two species. Alternatively, the apparent high intron gain rate exhibited by this gene may have arisen just prior to the last common ancestor of *N. crassa* and *M. grisea*. Although it is premature to speculate about possible mechanisms, one possibility is that a factor or factors responsible for intron insertion evolved to associate with the PRPP synthetase gene locus, transcript, or message at this point, leading to a higher rate of intron insertion in this gene.

Finally, our results do not support the mechanism commonly proposed to account for the 5′ positional bias of introns in intron-poor organisms (Mourier and Jeffares 2003). Contrary to what would be expected if intron loss primarily involved recombination of poly-adenosine-primed reverse transcripts, the rate of intron loss tends to be lower at the 3′ ends of genes. Instead, the highest rates of intron loss occur in the middles of genes in all three organisms. (This result is consistent with the results of Roy et al. (2003) in their analysis of intron evolution in mammals. Although their report describes only six instances of loss, in each case it was an internal intron.) The preference for internal introns may reflect a process of reverse transcription primed internally. Alternatively, there may be pressure to preferentially conserve introns near the 5′ and 3′ ends of genes. In particular, there is strong evidence for a functional role for the 5′-most intron in many genes. What remains clear is that the pattern of loss in these fungi over the last 330 million years cannot be explained solely by a mechanism involving 3′-end-primed reverse transcription of spliced messages. Instead, fungal intron dynamics appear to reflect a more complex interplay between intron gain and loss, an interplay that is likely to shape intron evolution in other eukaryotes.

Materials and Methods

Sequences and annotations. All sequences and annotations were taken from the Broad Institute Fungal Genome Initiative website (http://www.broad.mit.edu/annotation/fungi/fgi). The following datasets were used: *A. nidulans* (Assembly 1, 18 February 2003), *N.* crassa (Assembly 3, 1 February 2001), *F. graminearum* (Assembly 1, 11 March 2003), and *M. grisea* (Assembly 2, 18 July 2002).

Ortholog identification. A group of four proteins, one from each organism, was considered an ortholog set if each pair was a pairwise best bidirectional BLAST hit in the respective genomes, and all the BLAST hits overlapped by at least 60% of the length of the longest protein. This yielded 2,073 sets of orthologs (out of an average of 10,500 genes in the four organisms). We repeated our analysis, requiring that each best bidirectional hit also be the only BLAST hit in each genome (spanning 60% the length of the longest protein). This protocol yielded only 1,178 ortholog sets, but gave qualitatively similar results for intron gains and losses (Figure S1).

Ortholog alignment. The proteins in each ortholog set were aligned using ClustalW 1.82 (Chenna et al. 2003), and intron position characters were inserted into the alignments, using "0," "1," or "2" to indicate the intron phase. Phase 0 intron characters were inserted between the amino acids coded for by the codons adjacent to that intron, and phase 1 and 2 intron characters were inserted immediately following the amino acid coded for by the codon interrupted by the intron. If an intron was not present in all the sequences at a given position, special intron gap characters, were inserted in the other sequences in order to maintain the downstream

Table 2. Exonic Nucleotide Composition near Introns

Position Class	Organism	Nucleotide Composition	Upstream Position				Downstream Position			
			−4	−3	−2	−1	+1	+2	+3	+4
All intron positions[a]	Nc	% A	27.2	31.5	37.4	16.2	26.4	22.9	26.1	23.7
		% C	27.0	25.5	20.7	13.0	22.8	29.5	30.9	30.1
		% G	23.5	22.8	19.6	52.5	32.0	18.1	18.9	22.0
		% T	22.4	20.2	22.3	18.4	18.9	29.5	24.1	24.3
		Bits	0.01	0.02	0.06	0.26	0.03	0.03	0.02	0.01
	Mg	% A	27.7	33.8	38.0	16.6	26.0	22.0	25.1	24.2
		% C	27.6	25.8	20.8	12.8	24.4	29.7	29.1	29.3
		% G	23.5	21.6	18.8	51.2	30.7	18.6	19.6	21.2
		% T	21.3	18.8	22.3	19.3	19.0	29.8	26.2	25.3
		Bits	0.01	0.04	0.06	0.24	0.02	0.03	0.01	0.01
	Fg	% A	28.8	34.0	37.5	18.0	27.9	22.6	25.6	24.5
		% C	27.1	25.2	20.0	12.0	20.2	26.6	27.1	27.7
		% G	20.7	20.7	17.8	48.7	30.6	17.4	18.3	21.6
		% T	23.4	20.2	24.7	21.4	21.3	33.4	29.0	26.3
		Bits	0.01	0.03	0.06	0.21	0.02	0.04	0.02	0.01
Gain positions[b]	Nc	% A	21.4	33.7	28.4	18.0	26.1	17.4	22.5	22.8
		% C	30.1	25.3	28.7	23.3	20.5	29.5	27.8	33.2
		% G	23.6	20.2	21.1	43.8	31.2	22.8	26.1	23.0
		% T	25.0	20.8	21.9	14.9	22.2	30.3	23.6	21.1
		Bits	0.01	0.03	0.01	0.13	0.02	0.03	0.01	0.02
	Mg	% A	20.4	31.1	28.6	17.7	20.2	17.5	17.1	22.2
		% C	35.2	28.8	24.9	22.0	26.3	29.4	37.2	32.9
		% G	16.5	22.4	18.7	44.2	33.3	24.9	21.6	25.7
		% T	27.8	17.7	27.8	16.2	20.2	28.2	24.1	19.3
		Bits	0.06	0.03	0.02	0.13	0.03	0.03	0.06	0.03
	Fg	% A	25.4	23.5	31.4	13.9	19.4	16.1	20.5	24.3
		% C	29.2	35.8	28.7	19.7	23.2	35.0	33.3	30.9
		% G	23.5	23.8	18.6	48.1	41.8	16.4	20.5	25.1
		% T	21.9	16.9	21.3	18.3	15.6	32.5	25.7	19.7
		Bits	0.01	0.05	0.03	0.19	0.11	0.09	0.03	0.02

[a] Four nucleotides were extracted upstream and downstream of each intron in the specified organism.
[b] Four nucleotides were extracted upstream and downstream of the orthologous site in the other two organisms, consistent with method of Qiu et al. 2004.
Fg, *F. graminearum;* Nc, *N. crassa;* Mg, *M. grisea.*
DOI: 10.1371/journal.pbio.0020422.t002

amino acid alignment. A total of 9,352 intron positions were aligned. At only 28 (0.3%) of these positions were introns of different phases aligned, making it reasonable to ignore "phase shifting" in our analysis.

Alignment filtering. Regions of low alignment quality were eliminated with a filter that required at least 30% identity and 50% similarity in a window of ten residues on each side of the intron position. These parameters were determined following manual classification of a set of 181 randomly selected intron positions as "clearly homologous," "ambiguous/possibly homologous," or "non-homologous" (see Figure 2B). Using the parameters above, 92% of the homologous positions, 29% of the ambiguous positions and only 2% of the non-homologous positions passed the filter.

To further exclude likely annotation and alignment errors, intron positions were also filtered by eliminating positions adjacent to gaps in the amino acid alignment and by eliminating positions with nearby introns but low evidence of homology in the intervening sequence. It is possible that some of these positions may in fact reflect intron gain or loss events that occurred simultaneously with coding sequence insertion or deletion. However, removing these positions did not significantly impact our results, as the number of positions adjacent to gaps was only about one-tenth of the number of positions that passed the quality filter, and the introns removed did not have an apparent positional bias (see Figure 3A)

Statistical significance of high gain rate in PRPP synthetase. We modeled the number of gains in a particular organism as a Poisson distribution under two different null hypotheses. One null hypothesis was that the gains were spread uniformly across all genes. The other was that the number of gains in each gene was proportional to the length of the gene. In the first case the Poisson parameter λ is given by the total number of raw gains observed in that organism divided by the total number of ortholog sets ($p < 3 \times 10^{-22}$ for *M. grisea*, $p < 4 \times 10^{-9}$ for *N. crassa*, and $p < 0.007$ for *F. graminearum*). In the second case λ is given by the total number of raw gains observed in that organism multiplied by the length of that gene in amino acids and divided by the total number of amino acids in all genes in that organism ($p < 7 \times 10^{-25}$ for *M. grisea*, $p < 3 \times 10^{-10}$ for *N. crassa*, and $p < 0.003$ for *F. graminearum*). We reported the less significant of the two p-values in the results.

Analysis of intron gain phase and sequence preference. For each of the three in-group lineages, the frequency of phase 0, 1, and 2 introns was determined for five different datasets: for each class of conservation (conserved, raw gains, and raw losses), for all introns passing our filter, and for all introns in the ortholog set. The p-value for the significance of phase 0 bias was determined by the χ^2 test with two degrees of freedom using equal expected phase frequencies. To detect sequence bias at intron insertion sites, we examined gained introns separately in *F. graminearum*, *M. grisea*, and *N. crassa*. For each

gained intron, we extracted four bases upstream and downstream of orthologous sites in the other two sequences, consistent with Qiu et al. (2004). The results are shown in Table 2.

Supporting Information

Figure S1. Intron Gains and Losses Inferred from Best-Only BLAST Hit Orthologs

Positional biases in intron gain, loss, and current distribution in three fungal genomes determined using orthologs predicted by a "bidirectional only hit" method. (A), (B), and (C) are roughly analogous to (D), (E), and (A), respectively, in Figure 3.

Found at DOI: 10.1371/journal.pbio.0020422.sg001 (78 KB DOC).

Table S1. Database of Alignments of All 1,447 Ortholog Sets with at Least One Passing Intron Position

Also available at http://genes.mit.edu/NielsenEtAl/.

Found at DOI: 10.1371/journal.pbio.0020422.st001 (4.3 MB ZIP).

Acknowledgments

We thank S. Calvo, L.-J. Ma, and E. S. Lander for comments. This work was supported by grants from the National Institutes of Health, National Science Foundation, and United States Department of Agriculture (CBN, BB, and JEG) and the Burroughs Wellcome Fund (CBB).

Conflicts of interest. The authors have declared that no conflicts of interest exist.

Author contributions. CBN, BB, CBB, and JEG conceived and designed the experiments. CBN, BF, and JEG analyzed the data. CBN, BF, and JEG contributed materials/analysis tools. CBN, BF, CBB, and JEG wrote the paper. ■

References

Berbee ML, Taylor JW (2000) Funal molecular evolution: Gene trees and geologic time. In: McLaughlin DJ, McLaughlin EG, Lemke PA, editors. The Mycota, Volume VII: Systematics and evolution, part B. New York: Springer-Verlag. pp. 229–246.

Boeke JD, Garfinkel DJ, Styles CA, Fink GR (1985) Ty elements transpose through an RNA intermediate. Cell 40: 491–500.

Chenna R, Sugawara H, Koike T, Lopez R, Gibson TJ, et al. (2003) Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res 31: 3497–3500.

Coghlan A, Wolfe KH (2004) Origins of recently gained introns in *Caenorhabditis*. Proc Natl Acad Sci U S A 101: 11362–11367.

Dollo L (1893) Les lois de l'evolution. Bull Soc Belge Geol Pal Hydr 7: 164–166.

Farris JS (1977) Phylogenetic analysis under Dollo's law. Syst Zool 26: 77–88.

Fedorov A, Roy S, Fedorova L, Gilbert W (2003) Mystery of intron gain. Genome Res 13: 2236–2241.

Fink GR (1987) Pseudogenes in yeast? Cell 49: 5–6.

Hartung F, Blattner FR, Puchta H (2002) Intron gain and loss in the evolution of the conserved eukaryotic recombination machinery. Nucleic Acids Res 30: 5175–5181.

Heckman DS, Geiser DM, Eidell BR, Stauffer RL, Kardos NL, et al. (2001) Molecular evidence for the early colonization of land by fungi and plants. Science 293: 1129–1133.

Llopart A, Comeron JM, Brunet FG, Lachaise D, Long M (2002) Intron presence-absence polymorphism in *Drosophila* driven by positive Darwinian selection. Proc Natl Acad Sci U S A 99: 8121–8126.

Logsdon JM Jr, Tyshenko MG, Dixon C, D-Jafari J, Walker VK, et al. (1995) Seven newly discovered intron positions in the triose-phosphate isomerase gene: Evidence for the introns-late theory. Proc Natl Acad Sci U S A 92: 8507–8511.

Logsdon JM Jr, Stoltzfus A, Doolittle WF (1998) Molecular evolution: Recent cases of spliceosomal intron gain? Curr Biol 8: R560–R563.

Mourier T, Jeffares DC (2003) Eukaryotic intron loss. Science 300: 1393

Nei M, Kumar S (2000) Molecular evolution and phylogenetics. New York: Oxford University Press. 333 p.

O'Neill RJ, Brennan FE, Delbridge ML, Crozier RH, Graves JA (1998) De novo insertion of an intron into the mammalian sex determining gene, SRY. Proc Natl Acad Sci U S A 95: 1653–1657.

Qiu WG, Schisler N, Stoltzfus A (2004) The evolutionary gain of spliceosomal introns: Sequence and phase preferences. Mol Biol Evol 21: 1252–1263.

Robertson HM (2000) The large srh family of chemoreceptor genes in *Caenorhabditis* nematodes reveals processes of genome evolution involving large duplications and deletions and intron gains and losses. Genome Res 10: 192–203.

Rogozin IB, Wolf YI, Sorokin AV, Mirkin BG, Koonin EV (2003) Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. Curr Biol 13: 1512–1517.

Roy SW, Fedorov A, Gilbert W (2003) Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. Proc Natl Acad Sci U S A 100: 7158–7162.

Taylor TN, Hass H, Kerp H (1999) The oldest fossil ascomycetes. Nature 399: 648.

Appendix 2

—

Supplementary Material for Chapter 2

Supporting Information

MEF Viability and LacZ Analysis

To assess cell viability, untreated and treated MEFs were viewed and images were acquired using a Zeiss microscope (Axiovert 200). For lacZ stained cells MEFs were viewed and images were acquired under a Nikon microscope (Eclipse TE300). Over 100 cells and at least three screen shots were counted at each time point.

Western Analysis

Total proteins were extracted using Lysis Buffer (20 mM Tris-HCl pH 7.5, 150 mM NaCl, 1 mM EDTA, 1 mM EGTA, 1% Triton, 2.5 mM sodium pyrophosphate, 1 mM beta-glycerophosphate) following cell rinsing in cold PBS. For Western analysis, 20 ug of total proteins were separated in 7.5% Tris-HCl gels (Bio-Rad), transferred to a Protran membrane (Schleicher and Schuell) using an electroblot in 1xTBST (10 mM Tris-Cl, pH 8.0; 150 mM NaCl, 0.05% Tween 20) at 250 mA for 3 hr. Blocking was carried out using 2% Blocking agent (GE/Amersham) for 1 hr at RT, followed by 3 short washes in 1xTBST, then antibody incubations, anti-Dicer (Abcam, 1:500 dilution) and GAPDH (Santa Cruz Biotech; 1:2,000 dilution), were carried out at 4^{o} C for 16 hours in 2% Block solution. Secondary anti-rabbit IgG-HRP conjugate (Sigma; 1:10,000 dilution) was incubated for 1 hr at RT in 2% Blocking solution. Washes were carried out in 1xTBST for 15 min (3 times). SuperSignal West Femto Maximum Sensitivity Substrate (Pierce) was used for chemiluminescence. As a positive control recombinant Dicer (Stratagene) was also analyzed.

LacZ Staining

Cells were washed in cold PBS+2mM $MgCl_2$, fixed in 0.2% Glutaraldehyde (Sigma; diluted in PBS+2mM $MgCl_2$) for 10 minutes on ice, then washed 3 times in PBS and rinsed with Rinse Buffer (0.1M Na phosphate, 0.1% NaDeoxycholate, 2mM $MgCl_2$, 0.2% NP-40). X-gal staining (5mM $K_3Fe(CN)_6$, 5mM $K_4Fe(CN)_6$, 1 mg/ml Xgal, in Rinse Buffer) was added for 3 hr at 37^oC before microscopy.

Apoptosis Assay

Cells were harvested, washed twice in cold PBS and resuspended at a concentration of 1 x 10^6 cells/ml in Binding Buffer (10mM HEPES, 140 mM NaCl, 2.5 mM $CaCl_2$, pH7.4). Annexin V-FITC (Sigma), which detects annexin V bound to apoptotic cells, and propidium iodide (2 ug/ml), which labels cellular DNA in necrotic cells, were added and incubated for 10 min at room temperature in the dark. Samples were then immediately analyzed by flow cytometer (Becton Dickinson, FACScan). Staining with both Dyes allowed differentiation among early apoptotic cells (annexin V positive, PI negative), necrotic cells (annexin V positive, PI positive), and viable cells (annexin V negative, PI negative).

Northern Analysis and Probes

Probes were generated by incubating a total of 20 ul of 20 uM oligo, 2 ul Polynucleotide Kinase (PNK; New England Biolabs), 2 ul 10x PNK Buffer (New England Biolabs) and ^{32}P gamma-ATP (6000Ci/mmol) for 1 hr at 37^o C. Non incorporated nucleotides were

removed using MicroSpin G-25 Columns (GE/Amersham). Sequences of oligonucleotide probes used were as follows: miR21-5'-tcaacatcagtctgataagcta; miR22-5'-acagttcttcaactggcagctt; miR23b-5'-ggtaatccctggcaatgtgat; miR34a-5'-aacaaccagctaagacactgcca; miR92-5'-acaggccgggacaagtgcaata; miR191-5'-agctgcttttgggattccgttg; miR199a2-5'-gaacaggtagtctgaacactggg; miR200b-5'-catcattaccaggcagtatta; U6-5'-ttgcgtgtcatccttgcgcagg.

MEF miRNA Microarray Preparation, Hybridization and Scanning

MicroRNA microarrays were printed using a Cartesian PixSys 5500 Arrayer on Epoxy slides (Corning) using Ambion's miRvana amine-modified DNA oligonucleotide probe set (version 1564V1). Probes were printed at 50 uM in Printing Buffer (0.25M Sodium Phosphate buffer pH8.5, 250 uM Sorkosyl) in quadruplicate. 30 ug of total RNA was separated in 15% TBE-UREA gels. The 15 to 25 nt gel region, identified using siRNA Marker (New England Biolabs) and Ethidium Bromide staining, was excised and RNA was extracted by overnight incubation at 4^{o}C in 1M NaCl followed by ethanol precipitation. Labeling of small RNA was carried out using the miRvana miRNA labeling kit (Ambion) and Cy3/5 (GE/Amersham). 10 pmol of each labeled Dye was added onto an array in 1x Hybridization Buffer (Ambion), covered by a LifterSlip (Erie Scientific), and hybridization was carried out in Corning Hybridization chambers II for 16 hr in a water bath set to 42^{o} C. Washes were performed at room temperature according to manufacturer's protocol and solutions (Ambion; salt and detergent reagents). Arrays were spun down at 500 g for 5 min and scanned immediately using an Axon Scanner GenePix 4000.

MEF miRNA Microarray Analysis

The raw GenePix (.gpr) data were imported into R (www.r-project.org) for subsequent analysis. The median background of each array was estimated by analyzing the distribution of intensities obtained for negative controls (spotted RNA for which no complementary RNA was spiked-in). The median intensity (of 4 duplicated spots) for each miRNA and array was compared to the array-background to identify miRNAs with intensities more than 2 standard deviations above the median background. To estimate changes in miRNA expression levels we compared the experiment/reference RNA (cy3/cy5; reference RNA for all arrays was size selected HeLa cell line RNA) from the control arrays with the two MEF CDKO+OHT samples across all miRNAs. Spiked-in control RNA was to verify that no systematic bias was introduced during sample processing and hybridization. This array platform probably cannot distinguish between closely related miRNA species (e.g., closely related miRNA family members) due to cross-hybridization. However, this does not affect our conclusions because all analyses using the array data were done at the level of miRNA families (defined by unique seeds) rather than individual miRNAs.

Identification of the 'Effective' siRNA Subset

For each of the 4096 possible 6mers, Refseq genes in the siRNA dataset were divided into two classes based on presence/absence of the 6mer in their annotated 3'UTR. For each 6mer, the significance of the difference in LFC distributions between these classes was tested by two-sided rank sum test. siRNAs were considered to be 'effective' if a

6mer complementary to the extended seed region (positions 1 through 8) on either the sense or antisense strand gave a P-value $<10^{-6}$ or was among the top ten 6mers having the lowest P-values. If 6mers corresponding to the sense and the antisense strands both passed these criteria, then the one with the lowest P-value was kept and UTRs having seed matches to the other strand were removed from the analysis. Of the 74 siRNAs available for analysis, 52 met these criteria and reduced to 44 unique seed regions (MAPK14-1as, -2, -3as, -4as, -5as, -6, -7, -8as, -193, -M1, -M2as, -M4as, -M5as, -M6as, -M15, -M18, IGF1R-1as, -2, -3, -4as, -5as, -6, -10as, -11as, -12, -13, MPHOSPHQ-202as, -2692, PIK3CA-2629, PIK3CB-6338as, -6340as, PLK1-1319as, -772as, PRKCE-1295, SOD1-SNPp13as, -SNPp15as, -SNPp18as, -SNPp19as, -SNPp2as, -SNPp8as, -SNPp9as, -1582as, VHL-2651as, and -2652 where 'as' indicates the strand antisense to the targeted mRNA). Data from these 44 siRNAs was pooled for analysis. For the analyses shown in Fig. 1, the subset of 33 of these sequences that began with non-U bases was used.

Identification of 'Strongly Detected' and 'Responsive' MEF miRNAs

For the analysis shown in Fig. 3A, miRNAs with hybridization intensities above a threshold of 2 standard deviations above the median background level in seven or more of the eight miRNA microarrays were considered to be 'strongly detected'. Those below this threshold on all eight microarrays were considered 'not detected'. For the analysis shown in Fig. 3C, 'responsive' miRNAs were defined as follows. From the set of conserved miRNAs (seed region m1-m8 common to a miRNA in both mouse and human miRBase 8.2 (microrna.sanger.ac.uk)), we compared the LFC CDFs for a set of

mRNAs containing an extended seed match to the miRNA with the set of all mRNAs that lacked a seed match to the miRNA. This generated a list of 31 miRNAs (listed below) where seed match containing mRNAs were significantly upregulated relative to the non-seed match containing mRNAs ($P < 0.001$ by two-sided rank sum test). For this analysis, the sets of mRNAs containing an extended seed match were selected so as to include equal numbers of conserved and non-conserved seed matches, so as to avoid introducing any biases related to conservation. This was accomplished by sampling from the (invariably larger) set of mRNAs containing non-conserved seed matches a subset of the same size as the conserved mRNA set. Such sampling was performed at least 10 times and median P-values used. Responsive miRNAs: let-7d, let-7g, miR-9*, miR-15b, miR-19b, miR-26a, miR-30a-5p, miR-101a, miR-106a, miR-106b, miR-130a, miR-135a, miR-142-5p, miR-154, miR-155, miR-181a, miR-182, miR-186, miR-200b, miR-214, miR-291a-3p, miR-291b-3p, miR-302b*, miR-302c*, miR-320, miR-367, miR-381, miR-410, miR-424, miR-448, miR-495.

Controlling for Seed Match Type, Expression, Conservation and CG Content

Analyses of miRNA effects on mRNA levels were corrected for the effects of potentially confounding variables not under investigation. In the conservation analyses (Figs. 1C,E,G and Fig. 3C), for each mRNA in the conserved set we sampled at random and without replacement an mRNA from the non-conserved set that had the same number and type of seed matches, and roughly the same (within 10%) hybridization intensity value and fraction of conserved 7mers in its 3'UTR. Details of these controls are shown in Fig. S2. Variables that did not differ significantly between the sets (UTR length and

CG content shown in Fig. S2C) were not explicitly controlled. This same policy was applied to all analyses (e.g. mRNA expression levels across different seed match type mRNA sets). A similar approach was used to control for overall 3'UTR CG content in the nucleotide composition analyses (Fig. 5). Analyses of the effects of UTR length on targeting found either no difference (miR-1) or moderately increased downregulation for mRNAs with shorter UTRs (miR-124) (not shown). No significant effect of 3' UTR CG content was observed in the miRNA transfection data. For the analyses of seed match count shown in Fig. 4, there was not sufficient data to permit analysis of target downregulation as a function of seed match count for each seed match type separately. For each of the plots shown in Fig. 4, the proportion of seed match types for different seed match counts remained fairly constant.

No increased mRNA repression was associated with conserved versus non-conserved siRNA seed matches, when controlling for seed match type, expression and overall UTR conservation (Fig. 1G). However, slightly increased downregulation was associated with conserved siRNA seed matches when the control for overall UTR conservation was relaxed; this affect appears to be a consequence the increased local conservation that is associated with seed match conservation (not shown).

Calculation of signal:noise

Signal: noise ratios were calculated as in Lewis et al. (2005), but considering conservation only across human, mouse, rat, and dog genomes (HMRD) using cohorts of control oligonucleotides matched for both count and exact CG content. Ratios were pooled for the set of conserved human miRNAs used for target prediction by (Lewis et

al., 2005) after removal of miRNAs with common m2-m8 seed regions but different m9 nucleotides and pairs of miRNAs in the same super-family.

Orthologous 3' UTRs for zebrafish and *Tetraodon* were collected as described in Methods. Using an approach similar to Lewis et al. (2005), the number of occurrences of each 7mer was enumerated in each zebrafish 3' UTR and 7mers which also occurred in the corresponding *Tetraodon* UTR were recorded as conserved. In cases of multiple occurrences of the same 7mer in a zebrafish UTR and fewer occurrences in the *Tetraodon* UTR, only the common counts were recorded as conserved. For both the miR-430 A1 7mer and M8 7mer, sets of control 7mers with roughly equal total occurrences (within 10 counts) were collected and the mean fraction conserved for control and miR-430 7mers calculated. The ratio of miR-430 7mer to control 7mer fraction conserved is reported as the signal:noise.

Repression for non-conserved 8mers versus conserved 7mers

In the Lim miRNA transfection data, we observed stronger downregulation associated with non-conserved 8mer seed matches, especially those with a t9W, than for conserved 7mer seed matches (Fig. S8A), using conservation criteria identical to those used by the TargetScanS algorithm. This observation suggests that presence of a non-conserved 8mer seed match is at least as reliable a predictor of miRNA targeting – given co-expression with the corresponding miRNA – as is a TargetScanS prediction based on 7mer conservation (Lewis et al., 2005). Consistently, in the MZdicer knockout

system, presence of a non-conserved 8mer seed match was associated with stronger repression than for 7mer seed matches conserved to other fish (Fig. S8B).

TargetRank Scoring

TargetRank scores the seed matches in a UTR relative to a given siRNA or miRNA, and then calculates an overall score for the mRNA as a whole by summing the scores for all seed matches present in the 3' UTR. The score for each seed match, m, is calculated according to $S(m) = S_{SeedMatchType}(m) + R_{5'conservation}(m) + R_{3'AU}(m)$, where $S_{SeedMatchType}(m)$ is the mean nLFC for the seed match type represented by m, and $R_{5'conservation}(m)$ and $R_{3'AU}(m)$ represent the residual contribution to nLFC associated with the level of sequence conservation immediately 5' of the seed match and the AU content immediately 3' of the seed match, respectively. For A1 7mer and 6mer seed matches, the $S_{SeedMatchType}(m)$ value is determined as in Fig. 1F (parameters used: 6mer: 0.04; A1 7mer: 0.11). For M8 7mer and 8mer seed matches, the t9W effect is also incorporated by assigning $S_{SeedMatchType}(m)$ dependent on the seed match type and t9 base, as in Fig. 5A (parameters used: M8 7mer with t9W: 0.15; M8 7mer with t9S: 0.08; 8mer with t9W: 0.25; 8mer with t9S: 0.17). $R_{3'AU}(m)$ is determined by first assigning the seed match to one of three bins based on the %AU content in the 50 bases immediately 3' of the seed match (as in Fig. 6D). $R_{3'AU}(m)$ is then set equal to the mean nLFC for this bin in the training set of siRNA data, less the average mean nLFC across the 3 bins. For example, if the mean nLFC values of the 3 bins are 0.10, 0.12, and 0.17 (average: 0.13), then the residual values for the three bins would be -0.03, -0.01 and 0.04, respectively. Unlike in Fig. 6D, binned mRNA sets were only controlled for seed match type and t9 composition, which are

variables already accounted for in the first term of the model (parameters used: bin 1: 3' AU < 53%, mean nLFC = 0.083, bin 2: 3' AU between 53% and 66%, mean nLFC = 0.126; bin 3: 3' AU > 66%, mean nLFC = 0.182). $R_{5'conservation}(m)$ is determined by assigning the seed match to one of three bins based on the %conservation in the 50 bases 5' of the seed match (as in Fig. 6C), and then calculating a residual score for this bin as described for $R_{3'AU}(m)$. Unlike in Fig. 6C, binned mRNA sets are controlled only for seed match type, t9 composition, and %AU in the 50 bp 5' of the seed match (parameters used: bin 1: 5' conservation < 33%, mean nLFC = 0.085, bin 2: 5' conservation between 33% and 56%, mean nLFC = 0.135; bin 3: 5' conservation > 56%, mean nLFC = 0.163). For a 3' UTR containing n seed matches $m_1, m_2, \ldots$, the TargetRank score is calculated simply as the sum $S(UTR) = \sum_{k=1}^{n} S(m_k)$, using the log-additivity of seed matches derived from Fig. 4. For Figs. 7A and 7C, a random subset of 8 siRNAs were held out from the Jackson/Schwarz datasets and parameters were estimated based the remaining 36 siRNA transfections. The same parameters were used for Figs. 7B and 7D.

Supplemental References

Jackson, A. L., Bartz, S. R., Schelter, J., Kobayashi, S. V., Burchard, J., Mao, M., Li, B., Cavet, G., and Linsley, P. S. (2003). Expression profiling reveals off-target gene regulation by RNAi. Nat Biotechnol *21*, 635-637.

Lewis, B. P., Burge, C. B., and Bartel, D. P. (2005). Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell *120*, 15-20.

Legends to Supplemental Figures

Fig. S1. Seed match type effects on mRNA repression for conserved and non-conserved seed matches. (A) Cumulative distribution functions (CDFs) of LFCs for mRNAs containing the indicated non-conserved miR-124 seed match types. Plots are based on mRNAs containing exactly one non-conserved miR-124 seed match of each seed type (and no conserved seed matches). Set sizes are shown in parentheses. (B) LFC CDFs for mRNAs containing a single conserved miR-124 seed match (and no non-conserved seed matches). (C) LFC CDFs for mRNAs containing a single non-conserved miR-1 seed match. (D) LFC CDFs for mRNAs containing a single conserved miR-1 seed match.

Fig. S2. Effects of seed match conservation on mRNA repression following miRNA transfection for controlled and uncontrolled datasets. (A) CDF of LFCs for mRNAs containing conserved (red) or non-conserved (blue) extended seed matches to miR-124, or no seed matches (gray). mRNAs with conserved seed matches may also contain non-conserved seed matches, though the non-conserved class is strict. Set sizes are shown in parentheses. (B) mRNAs from the non-conserved set were sampled without replacement to generate a set having the same extended seed match count and distribution across 3'UTR conservation and expression as the conserved set. 3'UTR conservation is measured as the fraction of all 7mers (not just miRNA seed matches) in the 3'UTR that are perfectly conserved in human, mouse, rat and dog aligned genomes. Expression is measured as $\log_2$ of the hybridization intensity in mock transfected cells. 3'UTR conservation and expression CDFs are shown for the sampled set having the

median rank sum statistic ($P \geq 0.05$). (C) $\log_2$ 3' UTR length and fraction CG content are shown for these same sets. Although these variables were not explicitly controlled, there is no significant difference between the sets ($P \geq 0.05$). (D) CDFs of LFCs for the controlled sets (same as in Figure 1C).

Fig. S3. Genomic organization of MEF CDKO mouse

Schematic representation of the genomic organization of the MEFcdko mouse and of Dicer inactivation. (A) MEFcdko mice were generated from mice bearing three unique genomic regions: (i) CAG-ERT promoter with the Prx1-Cre allele, inducible by Orthohydroxy Tamoxifen addition (4-OHT); (ii) R26 Promoter with a LacZ gene downstream of a floxed stop codon; (iii) Dicer1 gene with a floxed exon 24. (B) The CAG-ERT promoter is activated upon addition of 4-OHT to the medium, driving expression of Cre protein. (C) The floxed stop codon upstream of the LacZ gene and (D) Dicer1 exon 24 is excised, producing beta-galactosidase and a non-functional Dicer allele.

Fig. S4. Kinetics of Dicer knockout monitored by LacZ staining

(A) MEFwt and MEFcdko cells in the absence and presence of 0.5 uM OHT were stained daily to monitor LacZ expression. At 24 hour intervals the numbers of stained (blue) and unstained cells were counted. Images are presented for MEFwt and MEFcdko in the presence of OHT. (B) The percent of stained/blue cells counted on each day is plotted (1D to 4D). The mean and standard deviation of the mean of three replicates are shown.

Fig. S5. Knockout of Dicer does not induce apoptosis in MEFs.

FACS analysis of Annexin V was performed in MEFwt and MEFcdko in the absence and presence of 0.5 uM OHT for 1 to 4 days. Shown here are data for MEFwt and MEFcdko with and without OHT after 4 days (4D). PI Staining (dead cells) is shown on the x-axis; Annexin V-FITC staining (apoptotic cells) is shown on the y-axis. Each quadruple, clockwise from bottom left, shows: (i) unstained live cells; (ii) PI stained dead cells; (iii) PI and FITC stained dead/apoptotic cells; and (iv) FITC stained apoptotic cells, respectively. The number at the bottom right hand side in each section denotes the percentage of cells in the particular state as a fraction of the total. Percent apoptotic cells measured along a four day time course (1D to 4D) is plotted below. The day 4 experiment was repeated twice.

Fig. S6. Northern Analysis of miRNA Expression

Northern analysis is shown for four of the miRNAs represented in Fig. 2D. Background-corrected hybridization intensities were calculated for each experimental sample (MEFcdko+OHT; right lane of each gel) and for the three control samples (MEFcdko/MEFwt+OHT/MEFwt; first three lanes from the left). All bands were then normalized to U6 snRNA and the fold-change was calculated by dividing the normalized average of the control samples by the normalized experimental sample. In (A) membrane was stripped by incubating in 1% SDS solution for 10 min at 60°C and then re-hybridyzed several times.

Fig. S7. mRNA derepression following Dicer knockout in zebrafish varies with conservation status. CDFs of LCFs for mRNAs containing conserved (red) or non-conserved (blue) miR-430 extended seed matches, or no seed matches (gray). mRNAs containing conserved seed matches (see Methods) may also contain non-conserved seed matches, though the non-conserved set is strict. Set sizes are shown in parentheses.

Fig. S8. Prediction of Non-conserved miRNA Targets Containing 8mer Seed Matches. (A) Mean LFC for mRNAs containing non-conserved 7mers (blue), conserved 7mers (red), and non-conserved M8-A1 8mers (purple) to miR-1 and miR-124 following transfection of the corresponding miRNA. Dashed box indicates mean LFC for miR-1 and miR-124 W9-M8-A1 9mers. (B) Same as (A) for miR-430 following Dicer knockout in zebrafish embryos.

Fig S9. Effects of local conservation and AU content following miRNA transfection for controlled and uncontrolled datasets. (A) CDF of LFCs for equal sized sets of mRNAs containing a single siRNA extended seed match and grouped by conservation level in the 50 nt region immediately upstream of the siRNA seed match. Mean percent conservation values for the sets are as follows (most conserved (red) = 72%, moderately conserved (gray) = 44%, least conserved (green) = 11%). (B) mRNAs from the three sets were sampled without replacement such that the distributions of UTR conservation, expression level, upstream AU content and UTR AU content were not significantly different (rank sum test, $P >= 0.05$). Seed match types were also

matched across each bin (not shown). UTR conservation is measured as the number of positions in the human 3' UTR that are perfectly conserved in alignments to mouse, rat, and dog. Expression is measured as the log2 of the hybridization intensity in mock transfected cells. (C) CDFs of LFCs for the controlled sets (same data as shown in Figure 6A (upstream).

Fig. S1.

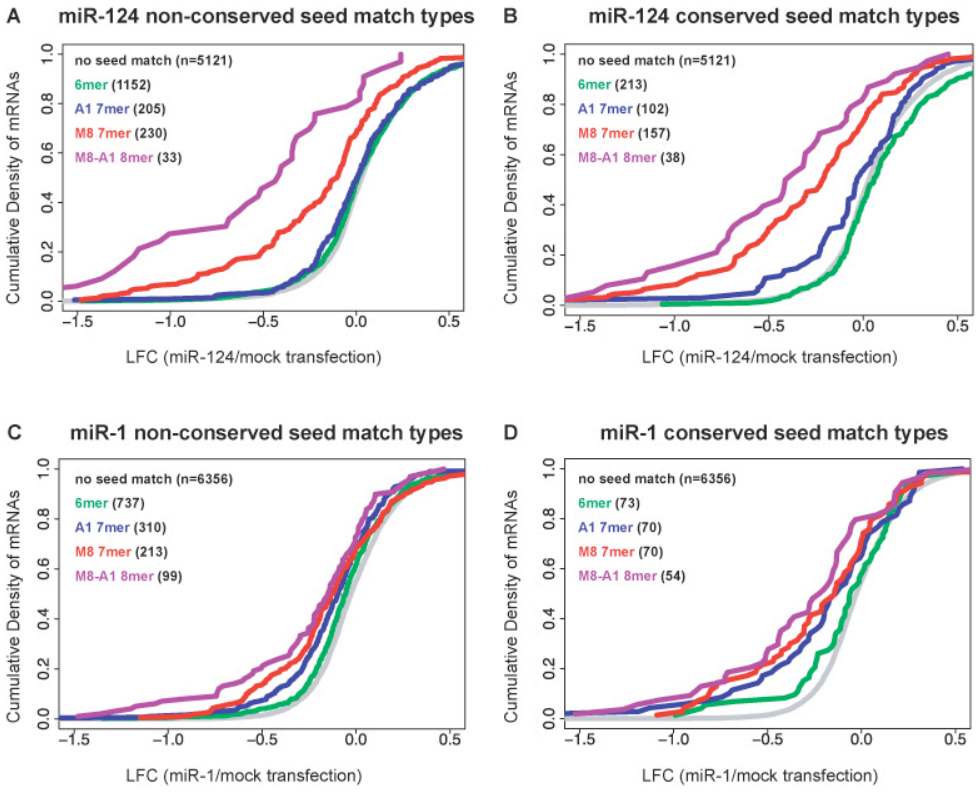
A
miR-124 non-conserved seed match types
no seed match (n=5121)
6mer (1152)
A1 7mer (205)
M8 7mer (230)
M8-A1 8mer (33)
Cumulative Density of mRNAs
LFC (miR-124/mock transfection)
B
miR-124 conserved seed match types
no seed match (n=5121)
6mer (213)
A1 7mer (102)
M8 7mer (157)
M8-A1 8mer (38)
Cumulative Density of mRNAs
LFC (miR-124/mock transfection)
C
miR-1 non-conserved seed match types
no seed match (n=6356)
6mer (737)
A1 7mer (310)
M8 7mer (213)
M8-A1 8mer (99)
Cumulative Density of mRNAs
LFC (miR-1/mock transfection)
D
miR-1 conserved seed match types
no seed match (n=6356)
6mer (73)
A1 7mer (70)
M8 7mer (70)
M8-A1 8mer (54)
Cumulative Density of mRNAs
LFC (miR-1/mock transfection)

Fig. S2.

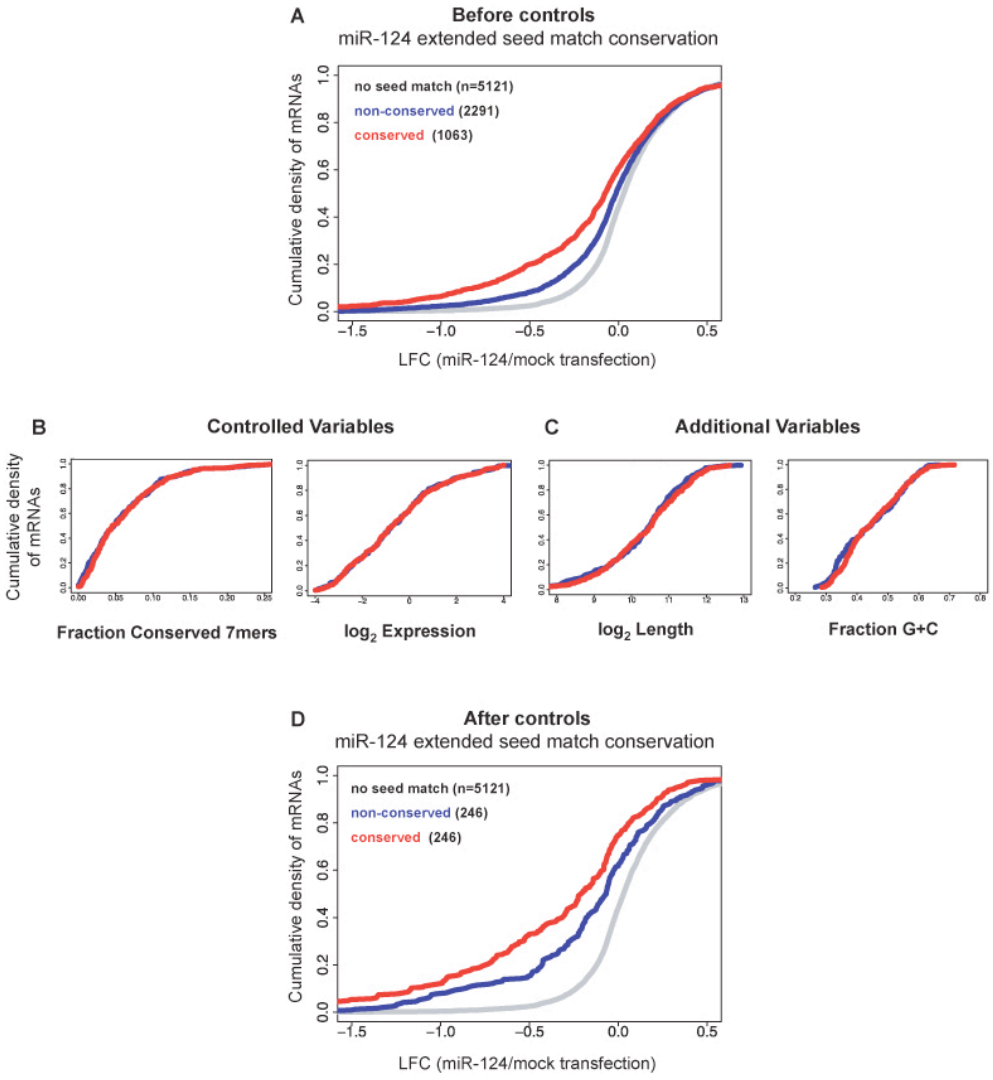
A
Before controls
miR-124 extended seed match conservation
no seed match (n=5121)
non-conserved (2291)
conserved (1063)
Cumulative density of mRNAs
LFC (miR-124/mock transfection)
B
Controlled Variables
C
Additional Variables
Cumulative density of mRNAs
Fraction Conserved 7mers
log2 Expression
log2 Length
Fraction G+C
D
After controls
miR-124 extended seed match conservation
no seed match (n=5121)
non-conserved (246)
conserved (246)
Cumulative density of mRNAs
LFC (miR-124/mock transfection)

Fig. S3.

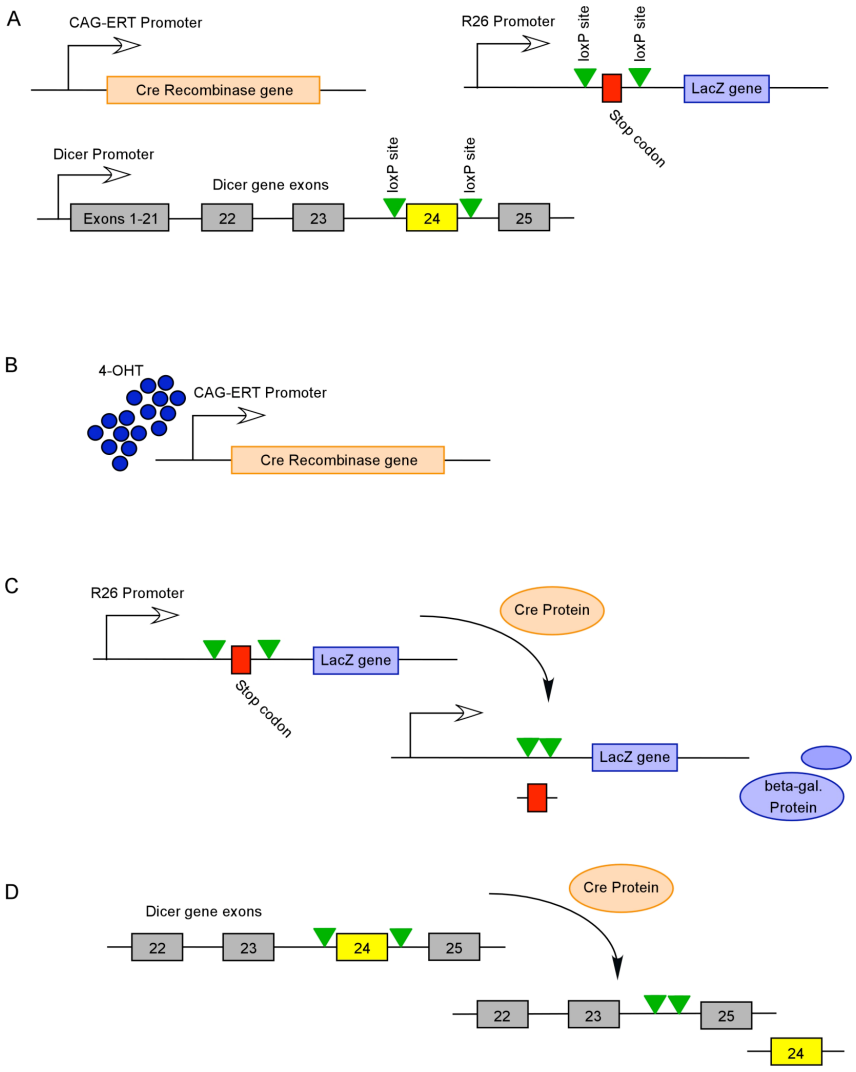
A
CAG-ERT Promoter
Cre Recombinase gene
R26 Promoter
loxP site
loxP site
LacZ gene
Stop codon
Dicer Promoter
Dicer gene exons
loxP site
loxP site
Exons 1-21
22
23
24
25
B
4-OHT
CAG-ERT Promoter
Cre Recombinase gene
C
R26 Promoter
Cre Protein
LacZ gene
Stop codon
LacZ gene
beta-gal.
Protein
D
Dicer gene exons
Cre Protein
22
23
24
25
22
23
25
24

Fig. S4.

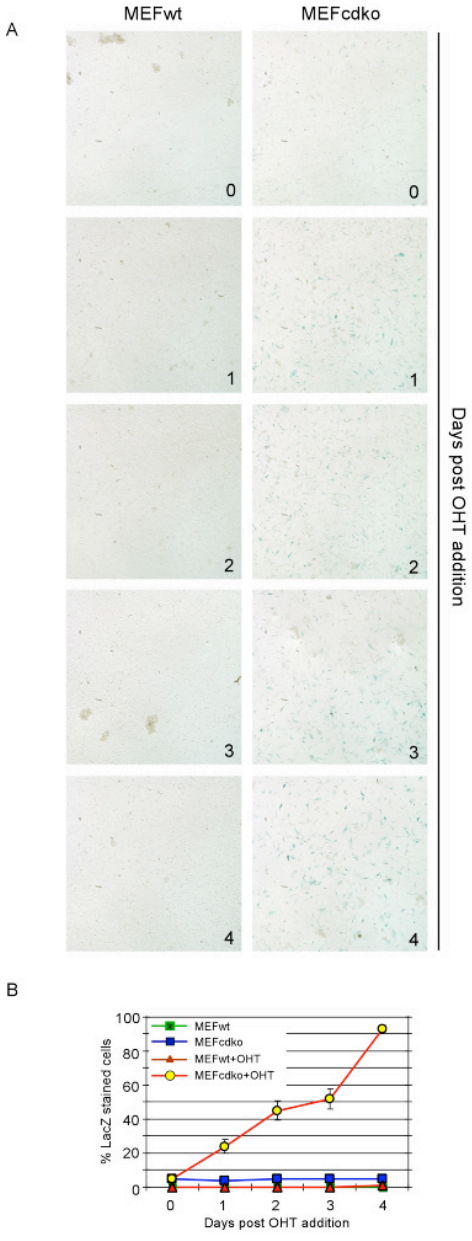
A
MEFwt
MEFcdko
0
1
2
3
4
Days post OHT addition
B
% LacZ stained cells
MEFwt
MEFcdko
MEFwt+OHT
MEFcdko+OHT
Days post OHT addition

Fig. S5.

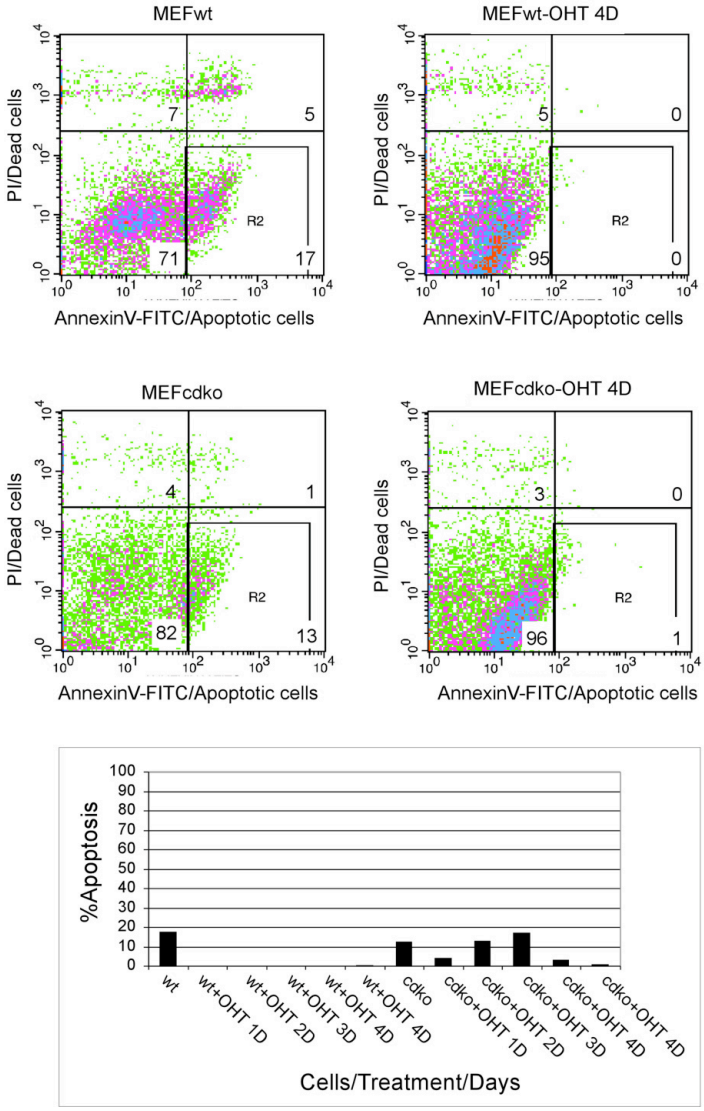
MEFwt
MEFwt-OHT 4D
MEFcdko
MEFcdko-OHT 4D
PI/Dead cells
AnnexinV-FITC/Apoptotic cells
R2
7
5
71
17
5
0
95
0
4
1
82
13
3
0
96
1
%Apoptosis
100
90
80
70
60
50
40
30
20
10
0
wt
wt+OHT 1D
wt+OHT 2D
wt+OHT 3D
wt+OHT 4D
wt+OHT 4D
cdko
cdko+OHT 1D
cdko+OHT 2D
cdko+OHT 3D
cdko+OHT 4D
cdko+OHT 4D
Cells/Treatment/Days

Fig. S6.

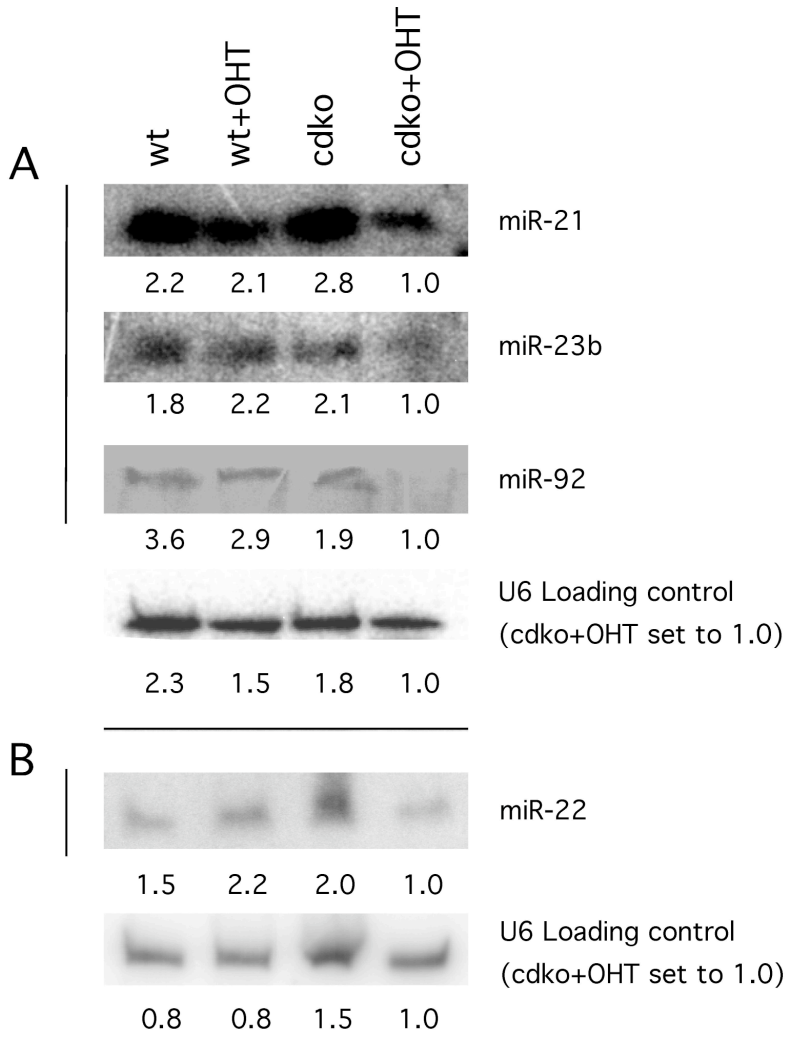
MicroRNA Northern blot relative band intensity
(loading control normalized)
wt
wt+OHT
cdko
cdko+OHT
A
miR-21
2.2 2.1 2.8 1.0
miR-23b
1.8 2.2 2.1 1.0
miR-92
3.6 2.9 1.9 1.0
U6 Loading control
(cdko+OHT set to 1.0)
2.3 1.5 1.8 1.0
B
miR-22
1.5 2.2 2.0 1.0
U6 Loading control
(cdko+OHT set to 1.0)
0.8 0.8 1.5 1.0

Fig. S7.

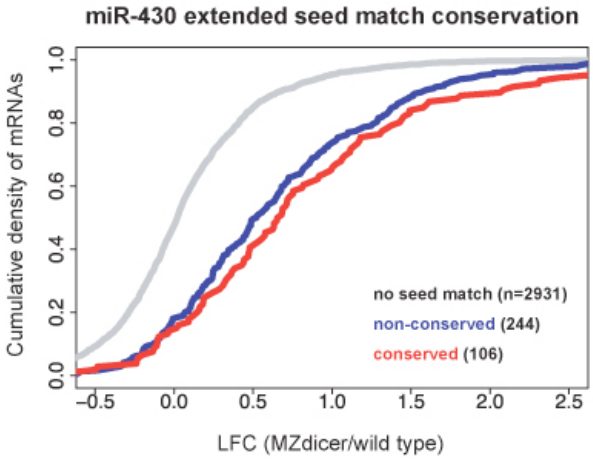
miR-430 extended seed match conservation
Cumulative density of mRNAs
0.0
0.2
0.4
0.6
0.8
1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5
LFC (MZdicer/wild type)
no seed match (n=2931)
non-conserved (244)
conserved (106)

Fig. S8.

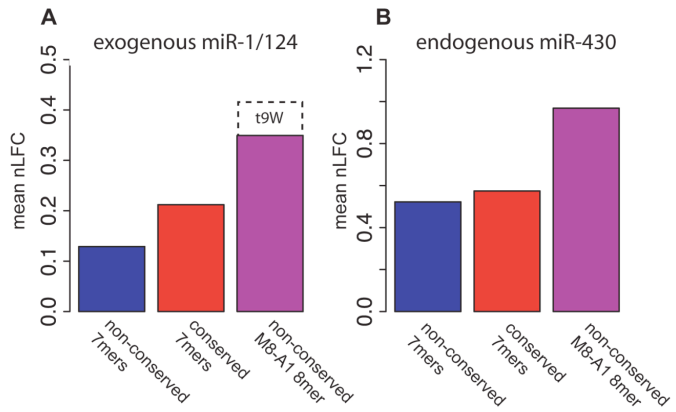
A
exogenous miR-1/124
mean nLFC
0.0
0.1
0.2
0.3
0.4
0.5
t9W
non-conserved 7mers
conserved 7mers
non-conserved M8-A1 8mer
B
endogenous miR-430
mean nLFC
0.0
0.4
0.8
1.2
non-conserved 7mers
conserved 7mers
non-conserved M8-A1 8mer

Fig. S9.

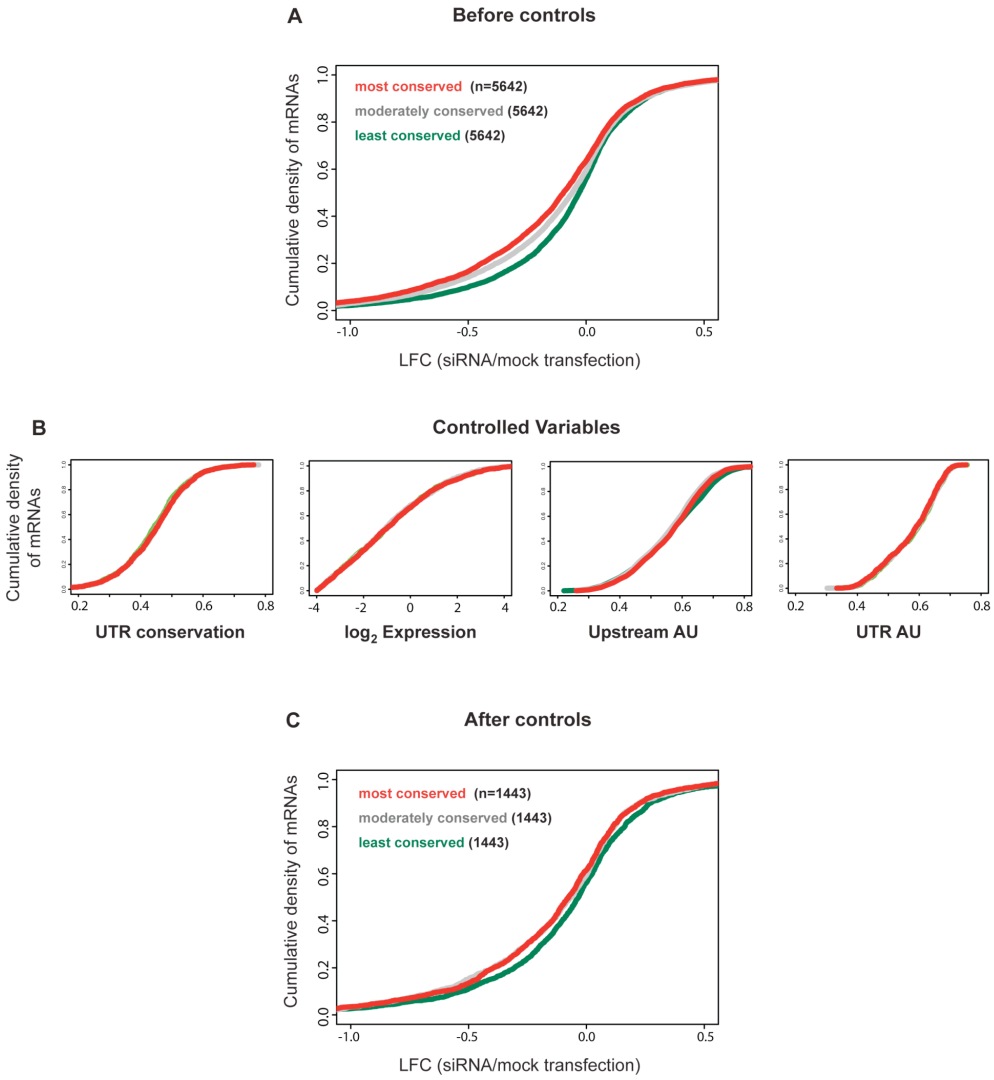
A
Before controls
most conserved (n=5642)
moderately conserved (5642)
least conserved (5642)
Cumulative density of mRNAs
LFC (siRNA/mock transfection)
B
Controlled Variables
Cumulative density of mRNAs
UTR conservation
log2 Expression
Upstream AU
UTR AU
C
After controls
most conserved (n=1443)
moderately conserved (1443)
least conserved (1443)
Cumulative density of mRNAs
LFC (siRNA/mock transfection)

Table S1. Statistics related to Fig. 1.

Figure 1B data – miR-124 seed match types								
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**frac. DR**	**Wilcoxon Rank Sum Test (2 sided) P-values**			
					no seed match	**M2-7 6mer**	**A1 7mer**	**M8 7mer**
no seed match	5121	0.05	0.00	0.025	-			
M2-7 6mer	1373	0.04	0.01	0.036	0.354	-		
A1 7mer	317	0	0.05	0.054	0.003	0.023	-	
M8 7mer	396	-0.25	0.25	0.250	$<10^{-43}$	$<10^{-33}$	$<10^{-13}$	-
M8-A1 8mer	77	-0.51	0.56	0.454	$<10^{-23}$	$<10^{-21}$	$<10^{-15}$	$<10^{-4}$

Figure 1C data – miR-124 extended seed match conservation					
con. type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**Wilcoxon Rank Sum Test (2 sided) P-values**	
				no seed match	**non-conserved**
no seed match	5121	0.05	0.00	-	
non-conserved	246	-0.15	0.20	$<10^{-12}$	-
conserved	246	-0.34	0.39	$<10^{-34}$	$<10^{-3}$

Figure 1D data – miR-1 seed match types								
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**frac. DR**	**Wilcoxon Rank Sum Test (2 sided) P-values**			
					no seed match	**M2-7 6mer**	**A1 7mer**	**M8 7mer**
no seed match	6356	-0.01	0.00	0.025	-			
M2-7	813	-0.05	0.04	0.052	$<10^{-4}$	-		
A1 7mer	400	-0.13	0.12	0.123	$<10^{-13}$	$<10^{-5}$	-	
M8 7mer	286	-0.15	0.14	0.196	$<10^{-13}$	$<10^{-5}$	0.202	-
M8-A1 8mer	170	-0.25	0.24	0.276	$<10^{-15}$	$<10^{-8}$	0.002	0.055

Figure 1E data – miR-1 extended seed match conservation					
con. type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**Wilcoxon Rank Sum Test (2 sided) P-values**	
				no seed match	**non-conserved**
no seed match	6356	-0.01	0.00	-	
non-conserved	178	-0.17	0.16	$<10^{-8}$	-
conserved	178	-0.29	0.28	$<10^{-20}$	0.007

Figure 1F data – non-m1U siRNA seed match types										
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**frac. DR**	**Wilcoxon Rank Sum Test (2 sided) P-values**					
					no seed	**M2-7 6mer**	**M1 7mer**	**A1 7mer**	**M8 7mer**	**M8-M1 8mer**
no seed match	155998	0.02	0.00	0.025	-					
M2-7	8102	-0.02	0.04	0.055	$<10^{-24}$	-				
M1 7mer	4296	-0.01	0.03	0.054	$<10^{-14}$	0.915	-			
A1 7mer	4673	-0.08	0.10	0.098	$<10^{-87}$	$<10^{-20}$	$<10^{-15}$	-		
M8 7mer	3350	-0.10	0.12	0.108	$<10^{-90}$	$<10^{-28}$	$<10^{-22}$	0.021	-	
M8-M1 8mer	1780	-0.10	0.12	0.117	$<10^{-54}$	$<10^{-20}$	$<10^{-19}$	0.014	0.561	-
M8-A1 8mer	1875	-0.21	0.23	0.186	$<10^{-149}$	$<10^{-79}$	$<10^{-67}$	$<10^{-29}$	$<10^{-19}$	$<10^{-13}$

Figure 1G data – siRNA extended seed match conservation					
con. type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**Wilcoxon Rank Sum Test (2 sided) P-values**	
				no seed match	**non-conserved**
no seed match	187980	0.02	0.00	-	
non-conserved	1643	-0.11	0.13	$<10^{-52}$	-
conserved	1643	-0.13	0.15	$<10^{-65}$	0.198

DR = down-regulated

Table S2. Statistics related to Fig. S1.

Figure S1A data – non-conserved miR-124 seed match types			
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**
no seed match	5121	0.05	0.00
M2-7	1152	0.03	0.02
A1 7mer	205	0.03	0.02
M8 7mer	230	-0.21	0.26
M8-A1 8mer	33	-0.55	0.60
Figure S1B data – conserved miR-124 seed match types			
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**
no seed match	5121	0.05	0.00
M2-7	213	0.1	-0.05
A1 7mer	102	-0.05	0.10
M8 7mer	157	-0.3	0.35
M8-A1 8mer	38	-0.44	0.49
Figure S1C data – non-conserved miR-1 seed match types			
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**
no seed match	6356	-0.01	0.00
M2-7	737	-0.05	0.04
A1 7mer	310	-0.12	0.11
M8 7mer	213	-0.13	0.12
M8-A1 8mer	99	-0.23	0.22
Figure S1D data – conserved miR-1 seed match types			
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**
no seed match	6356	-0.01	0.00
M2-7	73	-0.06	0.05
A1 7mer	70	-0.19	0.18
M8 7mer	70	-0.19	0.18
M8-A1 8mer	54	-0.3	0.29

Table S3. Downregulation of mRNA and Protein Levels in miR-1 and miR-124 Transfection Data

refseq ID	miRNA transfected	nLFC (Lim et al, 2005)	luciferase reporter LFC (Farh et al, 2005)
NM_170735	miR-1	0.39	0.99
NM_024652	miR-1	0.30	0.74
NM_031453	miR-1	0.34	0.37
NM_182692	miR-1	0.77	0.26
NM_181358	miR-1	0.28	0.13
NM_014325	miR-1	0.84	0.58
NM_000402	miR-1	1.14	1.77
NM_012395	miR-1	0.98	1.23
NM_015318	miR-1	0.99	0.87
NM_181358	miR-124	0.40	1.27
NM_139168	miR-124	0.24	0.21
NM_020639	miR-124	0.28	0.99
NM_014397	miR-124	1.07	1.31

Table S4. miRNAs with detectable expression in MEFs[1]

let-7a	miR-137	miR-195	miR-300
let-7b	miR-138	miR-196a	miR-301
let-7c	miR-140	miR-198	miR-302c
let-7d	miR-141	miR-199a*	miR-30a
let-7e	miR-142	miR-19a	miR-31
let-7f	miR-143	miR-200a	miR-320
let-7f	miR-144	miR-200b	miR-330
let-7g	miR-145	miR-202	miR-335
let-7i	miR-146	miR-203	miR-338
miR-101	miR-147	miR-206	miR-34c
miR-103	miR-153	miR-208	miR-361
miR-106a	miR-155	miR-214	miR-367
miR-106b	miR-15a	miR-215	miR-372
miR-107	miR-16	miR-217	miR-374
miR-10a	miR-17	miR-21	miR-376b
miR-122a	miR-181a	miR-221	miR-381
miR-124a	miR-182	miR-223	miR-382
miR-125a	miR-183	miR-22	miR-384
miR-125b	miR-184	miR-23b	miR-410
miR-126	miR-185	miR-24	miR-422a
miR-1	miR-18	miR-26a	miR-7
miR-130a	miR-190	miR-27a	miR-9
miR-130b	miR-191	miR-292	miR-93
miR-134	miR-192	miR-297	miR-98
miR-136	miR-194	miR-32	

[1]A set of 99 miRNAs were detected by microarray, defined as having median microarray intensity (of the four duplicate probes) greater than two standard deviations above background in at least 7 out of 8 samples, for those miRvana probes targeting miRNAs found in mouse (according to miRBase) that had at least five mRNA targets expressed in MEFs. The 99 different miRNAs represent 80 unique seeds.

Table S5. mRNAs with Significant Expression Change Following Dicer Knockout[2]

Refseq ID	score(d)	Fold Change	Gene Symbol
NM_028523	10.77	12.63	Dcbld2
NM_130861	4.64	8.28	Slco1a5
NM_011348	9.84	7.60	Sema3e
NM_011213	4.21	6.05	Ptprf
NM_007399	9.02	5.97	Adam10
XM_484932	3.99	5.97	NA
NM_009252	4.67	5.79	Serpina3n
NM_015762	4.30	5.02	Txnrd1
NM_029575	4.42	4.76	Tgfbr2
NM_009364	5.19	4.69	Tfpi2
NM_145390	4.45	4.69	Tnpo2
NM_013737	5.99	4.61	Pla2g7
XM_130125	6.53	4.59	NA
NM_020275	4.57	4.36	Tnfrsf10b
NM_008402	8.56	4.23	Itgav
NM_009684	8.71	4.03	Apaf1
NM_026735	4.44	3.97	Mobkl1a
NM_011052	4.46	3.82	Pdcd6ip
NM_010442	4.30	3.82	Hmox1
XM_194424	4.71	3.78	NA
NM_011198	4.65	3.70	Ptgs2
NM_011452	7.73	3.64	Serpinb9b
NM_028527	7.04	3.63	1700047I17Rik
NM_011563	4.21	3.59	Prdx2
NM_172891	4.54	3.58	Styk1
NM_030155	4.22	3.57	Sdccag3
NM_028744	6.48	3.50	Pi4k2b
NM_019819	7.84	3.48	Dusp14
NM_029438	4.63	3.43	Smurf1
NM_001004143	4.13	3.43	Usp22
NM_029000	4.15	3.37	Gvin1
NM_145413	4.17	3.35	C530043G21Rik
NM_019547	4.49	3.28	Rnpc1
XM_484088	4.34	3.25	NA
NM_011502	5.19	3.18	Stx3
NM_023785	3.73	3.16	Cxcl7
NM_153584	5.87	3.11	BC031353
NM_015806	4.84	3.11	Mapk6
NM_015760	4.91	3.09	Nox4
NM_175201	9.12	3.08	Rnf38
NM_172967	4.07	3.06	4930503L19Rik
NM_011179	4.86	3.06	Psap
NM_172507	3.72	3.03	Sh3bgrl2
NM_172787	8.84	2.96	L3mbtl3
NM_013601	6.61	2.94	Msx2
NM_026177	3.90	2.93	1200011I18Rik

NM_011267	4.71	2.93	Rgs16
NM_010913	4.47	2.91	Nfya
NM_010786	8.54	2.89	Mdm2
NM_008924	6.28	2.87	Prkar2a
XM_140740	3.93	2.83	NA
NM_017368,NM_198683	5.35	2.81	Cugbp1
NM_013609	4.40	2.80	Ngfb
NM_011951	4.59	2.79	Mapk14
NM_009648	5.80	2.78	Akap1
NM_021451	4.13	2.73	Pmaip1
NM_172513	4.68	2.70	BC049806
NM_007406	4.62	2.69	Adcy7
XM_358611,XM_359418	4.82	2.66	NA
NM_153103	3.81	2.65	Kif1c
NM_020012	3.84	2.64	Rnf14
NM_024269	5.97	2.64	Arl2bp
NM_029352	4.06	2.60	Dusp9
NM_010345	3.79	2.60	Grb10
NM_011026	5.17	2.59	P2rx4
NM_011595	7.10	2.59	Timp3
NM_025673	3.72	2.58	Golph3
NM_010923,NM_180960	3.94	2.58	Nnat
NM_009443	3.86	2.54	Tgoln1
NM_013862	4.04	2.54	Rabgap1l
NM_008338	3.76	2.54	Ifngr2
NM_053153	4.58	2.53	Klra18
NM_028932	4.46	2.52	Eaf1
NM_007690	5.80	2.51	Chd1
NM_172734	4.39	2.50	Stk38l
NM_176845	5.51	2.48	Ddhd1
NM_009831	5.12	2.47	Ccng1
NM_008442	4.90	2.46	Kif2a
NM_007453	5.35	2.45	Prdx6
XM_135842	3.92	2.41	NA
NM_178615	4.87	2.39	Rgmb
NM_007952	4.86	2.37	Pdia3
NM_030721	5.04	2.36	Acox3
NM_207239	7.63	2.34	Gtf3c1
NM_019661	7.22	2.33	0610042I15Rik
NM_019927	4.11	2.32	Arih1
XM_489703	6.15	2.30	NA
NM_026195	7.31	2.29	Atic
NM_026662	5.50	2.29	Prps2
XM_622555	6.39	2.28	NA
NM_029777	6.54	2.28	4930418P06Rik
NM_028243	3.87	2.28	Prcp
NM_028651	5.49	2.27	4930403J22Rik
NM_178907	5.63	2.26	Mapkapk3
NM_010324	4.56	2.25	Got1

NM_144543	3.94	2.25	Thy28
NM_011018	4.27	2.24	Sqstm1
NM_013882	4.34	2.24	Gtse1
NM_009516	5.10	2.22	Wee1
NM_011699	4.17	2.21	Lin7c
NM_026424	3.70	2.21	1500041J02Rik
NM_011299	6.41	2.21	Rps6ka2
NM_031256	3.95	2.20	Plekha3
NM_139154	4.35	2.19	Rab40c
NM_138681	13.21	2.19	Bcas3
NM_133349	4.69	2.18	Zfand2a
NM_019930	3.97	2.17	Ranbp9
NM_134013	4.87	2.17	Psme4
NM_007836	4.18	2.17	Gadd45a
NM_024226,NM_194052 ,NM_194053	4.98	2.17	Rtn4
NM_007614	3.68	2.16	Ctnnb1
NM_030690	3.77	2.16	Rai14
NM_172699	5.65	2.16	Foxj3
NM_026563	4.48	2.15	Sdccag3
NM_134133	5.07	2.14	2010002N04Rik
NM_023066	4.72	2.14	Asph
NM_010718	4.47	2.13	Limk2
NM_009798	4.28	2.12	Capzb
NM_008928	4.34	2.12	Map2k3
NM_030015	6.72	2.11	Peli1
NM_172863	5.29	2.11	Zfp697
NM_007922	8.75	2.10	Elk1
XM_128959	4.28	2.10	NA
NM_030246	4.42	2.10	Wdr21
NM_025762	3.69	2.08	4933434E20Rik
NM_008465	5.69	2.08	Kpna1
NM_008576	5.24	2.05	Abcc1
NM_019432	5.00	2.04	Tmem37
XM_127105	4.21	2.03	NA
NM_025951	4.70	2.03	Pi4k2b
NM_175245	3.79	2.02	2410129H14Rik
NM_177613	3.81	2.01	Cdc34
NM_019403	4.11	2.01	Rnf5
NM_010249	4.25	2.00	Gabpb1
NM_025716	5.60	2.00	4633402N23Rik
NM_026418	-2.86	-2.00	Rgs10
NM_008538	-2.79	-2.00	Marcks
NM_020276	-3.34	-2.00	Nelf
NM_175098	-3.29	-2.01	6330407D12Rik
NM_016778	-4.49	-2.01	Bok
NM_020026	-10.07	-2.01	B3galt3
NM_009685	-2.92	-2.02	Apbb1
NM_019869	-2.96	-2.03	Rbm14

NM_021605	-2.80	-2.03	Nek7
XM_622635	-3.62	-2.05	NA
NM_009878	-6.10	-2.05	Cdkn2d
NM_175130	-3.29	-2.06	Trpm4
NM_026530	-8.58	-2.06	E130307M08Rik
XM_143175	-2.92	-2.08	NA
NM_016765	-6.30	-2.08	Ddah2
NM_027309	-4.16	-2.08	Lysmd2
NM_172546	-2.97	-2.09	Cnksr3
NM_026447,NM_198931	-8.46	-2.09	Ppm1m
NM_027878	-3.94	-2.11	1200002N14Rik
NM_172711	-4.94	-2.11	AA407526
NM_011838	-3.29	-2.12	Lynx1
NM_009746	-2.84	-2.12	Bcl7c
NM_009166	-2.82	-2.12	Sorbs1
NM_025656	-2.94	-2.12	Sip1
NM_145524	-3.70	-2.14	BC004636
NM_152813	-3.00	-2.14	Plcd3
NM_030004	-4.13	-2.16	Cryl1
NM_001003946	-3.02	-2.18	Als2cr13
NM_207269	-2.93	-2.18	D330050I23Rik
NM_144862	-4.02	-2.19	Lims2
NM_011146	-3.49	-2.20	Pparg
NM_026298	-3.23	-2.23	4930553F24Rik
NM_009968	-3.31	-2.23	Cryz
NM_026122	-4.88	-2.24	Hmgn3
NM_013496	-2.91	-2.25	Crabp1
NM_017373	-3.71	-2.25	Nfil3
NM_175074	-4.84	-2.26	Hmgn3
NM_026024	-2.88	-2.27	Ube2t
NM_001024225	-4.46	-2.28	Defcr24
NM_007852	-4.46	-2.28	Defcr6
NM_029624	-2.99	-2.29	2400010G15Rik
NM_007760	-2.92	-2.29	Crat
NM_009004	-2.83	-2.34	Kif20a
NM_009672	-5.35	-2.34	Anp32a
NM_013543	-4.16	-2.35	H2-Ke6
NM_173752	-4.26	-2.35	1110067D22Rik
NM_025658	-4.15	-2.37	Ms4a4d
NM_009822	-2.85	-2.38	Cbfa2t1h
NM_133990	-3.53	-2.40	Il13ra1
XM_133813	-5.01	-2.40	NA
NM_016762	-4.39	-2.44	Matn2
NM_016764	-4.88	-2.44	Prdx4
NM_025522	-2.94	-2.45	Dhrs7
NM_010216	-3.22	-2.47	Figf
NM_178660	-2.81	-2.48	Rbms3
NM_010744	-4.28	-2.48	Tmed1
NM_133859	-3.40	-2.48	Olfml3

NM_175205	-2.80	-2.49	4632419I22Rik
NM_183254	-3.09	-2.49	1700025K23Rik
NM_029413	-3.07	-2.50	Morc4
NM_199195	-3.54	-2.50	Bckdhb
NM_134163	-2.99	-2.51	Mbnl3
NM_146162	-3.02	-2.53	BC025600
NM_010194	-3.24	-2.54	Fes
XM_134902	-3.02	-2.54	NA
NM_173011	-3.65	-2.56	Idh2
NM_148928	-2.91	-2.57	Gtf3c5
XM_619217	-2.96	-2.58	NA
NM_146040	-2.78	-2.60	Cdca7l
NM_178884	-2.95	-2.61	AW822216
NM_173426	-2.77	-2.66	1700012H17Rik
NM_009155	-3.73	-2.67	Sepp1
NM_021342	-3.11	-2.68	Kcne4
NM_010726	-3.99	-2.71	Phyh
NM_009472	-3.16	-2.76	Unc5c
XM_194370	-2.91	-2.76	NA
NM_028915	-2.86	-2.79	Lrrcc1
NM_026303	-2.80	-2.81	4930562C03Rik
XM_620727	-2.80	-2.81	NA
NM_010826,NM_194464	-2.85	-2.90	Mrvi1
NM_146249	-3.59	-2.91	BC031441
NM_026772	-5.04	-2.92	Cdc42ep2
XM_355247	-3.06	-2.99	NA
XM_283635	-3.22	-3.00	NA
NM_013665	-3.03	-3.04	Shox2
NM_016873	-4.56	-3.05	Wisp2
NM_144794	-4.38	-3.05	Tmem63a
NM_008046	-3.58	-3.07	Fst
NM_134147	-3.45	-3.10	D930010J01Rik
NM_011129	-3.15	-3.10	4-Sep
NM_177135	-3.20	-3.21	D830030K20Rik
XM_489067	-7.18	-3.28	NA
NM_028724	-2.91	-3.31	Rin2
NM_138315	-3.80	-3.34	Mical1
NM_007630	-3.14	-3.41	Ccnb2
NM_012006	-3.23	-3.43	Cte1
NM_001012335,NM_001012336,NM_010784	-3.14	-3.52	Mdk
NM_009776	-4.62	-3.57	Serping1
NM_026125	-2.91	-3.61	1110035L05Rik
NM_010931	-2.92	-3.74	Uhrf1
NM_027954	-4.92	-3.88	Syce2
NM_026514	-3.61	-3.93	Cdc42ep3
NM_010226	-4.13	-3.98	Fkhl18
NR_001592	-3.48	-4.01	NA
NM_009141	-3.06	-4.04	Cxcl5

XM_358515	-4.71	-4.05	NA
NM_001004359,NM_001005385,NM_026081	-2.91	-4.07	Gprasp1
XM_354836	-3.09	-4.10	NA
XM_130991	-3.21	-4.14	NA
NM_026928	-5.71	-4.31	1810014F10Rik
NM_007825	-4.31	-4.42	Cyp7b1
NM_172604	-2.99	-4.48	Scara3
NM_008452	-3.10	-4.51	Klf2
NM_008987	-3.20	-4.74	Ptx3
NM_016847	-2.76	-5.17	Avpr1a
NM_198161	-3.74	-5.23	Bhlhb9
XM_181304	-3.02	-6.86	NA
NM_148948	-2.76	-8.23	Dicer1
NM_138304	-3.38	-12.70	Calml4

[2]Statistically significant differences in mRNA expression levels following Dicer knockout in MEFs were identified using Significance Analysis of Microarrays (SAM) with a False Discovery Rate cutoff of 2%, and then requiring a fold change of at least two (up or down). Refseq transcript identifiers, d-statistics (from SAM), fold change and gene symbols are listed for each significant mRNA and ordered according to fold change. In cases where multiple Refseq transcripts from the same gene were not distinguishable by the probes on the Mouse 430 2.0 array, all Refseq ids are listed. This list includes 135 mRNAs whose expression increased following Dicer knockout and 119 mRNAs whose expression decreased (including Dicer1).

Table S6. Statistics related to Fig. 3.

Figure 3B data – zebrafish miR-430 seed match types								
seed type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**fraction up-regulated**	**Wilcoxon Rank Sum Test (2 sided) P-values**			
					no seed match	**6mer**	**A1 7mer**	**M8-A1 8mer**
no seed match	2931	0.07	0.00	0.025	-			
6mer	269	0.45	0.38	0.082	$<10^{-25}$	-		
A1 7mer	71	0.65	0.58	0.155	$<10^{-14}$	0.017	-	
M8 7mer	170	0.59	0.52	0.159	$<10^{-20}$	0.070	0.364	-
M8-A1 8mer	23	1.04	0.97	0.348	$<10^{-7}$	0.001	0.066	0.020

Figure 3C data – MEF extended seed match conservation					
conservation type	**no. of mRNAs**	**mean LFC**	**mean nLFC**	**Wilcoxon Rank Sum Test (2 sided) P-values**	
				no seed match	**non-conserved**
no seed match	556	-0.24	0.00	-	
non-conserved	761	-0.08	0.16	$<10^{-6}$	-
conserved	761	-0.01	0.23	$<10^{-10}$	0.030

Appendix 3

Supplementary Material for Chapter 3

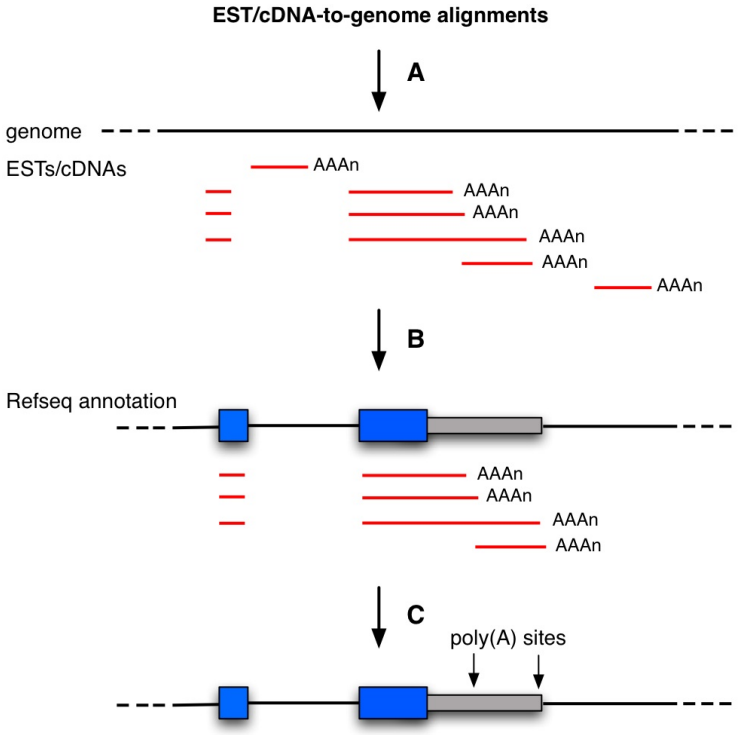

Figure S1: Outline of the approach used to build a database of poly(A) sites. **A**. First EST/cDNA-to-genome alignments from UCSC were filtered to keep only uniquely mapping ESTs/cDNAs with non-genomic poly(A) tails (minimum of 8 terminal *A* or *T* characters). **B**. Second, ESTs/cDNAs overlapping Refseq annotations were kept (blue boxes: exon; 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 nts of each other were clustered (based on Tian et al., 2005). The -1 to -40 region upstream of each poly(A) site was searched for a PAS or variant (Beaudoing et al., 2001). If a signal was found, the cluster was recorded as a poly(A) site (black arrow).

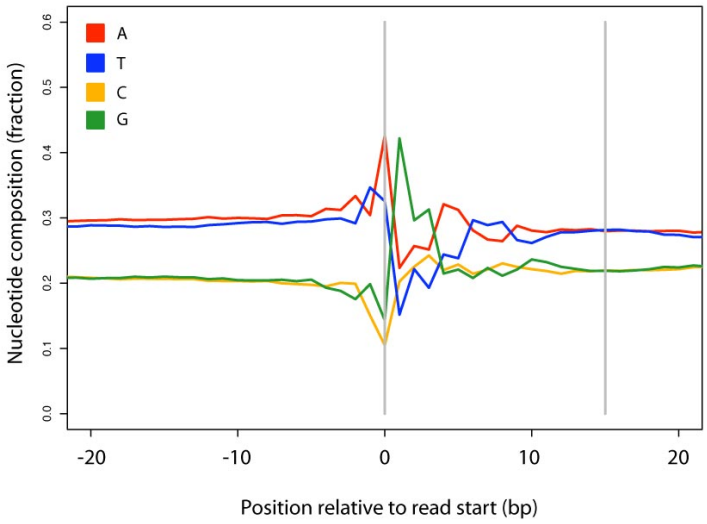

Figure S2: Nucleotide sequence composition around Illumina read starts from MNase digested chromatin (Barski et al., 2007). Gray vertical lines mark the 0 to +14 positions that were excluded from the nucleosome affinity model due to technical bias.

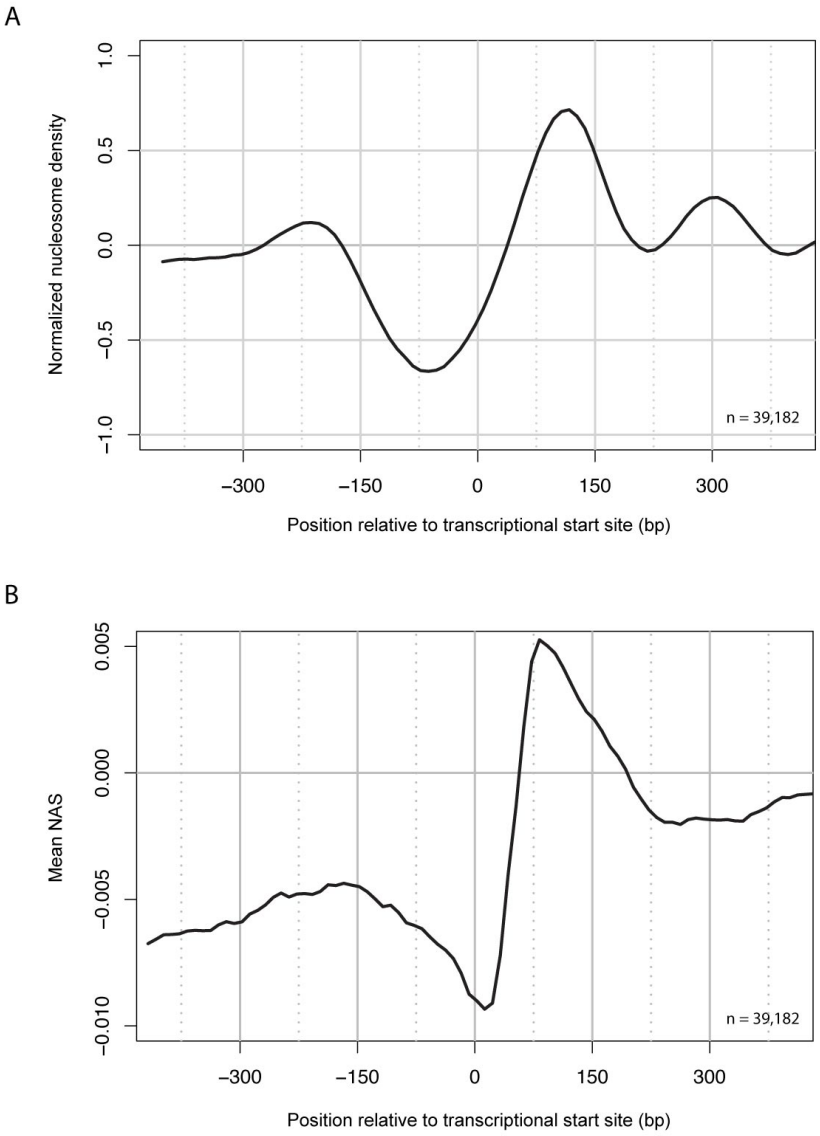

Figure S3: **A**. Nucleosome distribution around human transcriptional start sites for Refseq genes. Read density values (Barski et al., 2007 data) were normalized and smoothed as in Figure 1. **B**. Mean nucleosome affinity scores (NAS) around transcriptional start sites (same gene set as in A). Scores were smoothed for plotting using the average score from a 50 nt sliding window positioned every 10 nt (as in A).

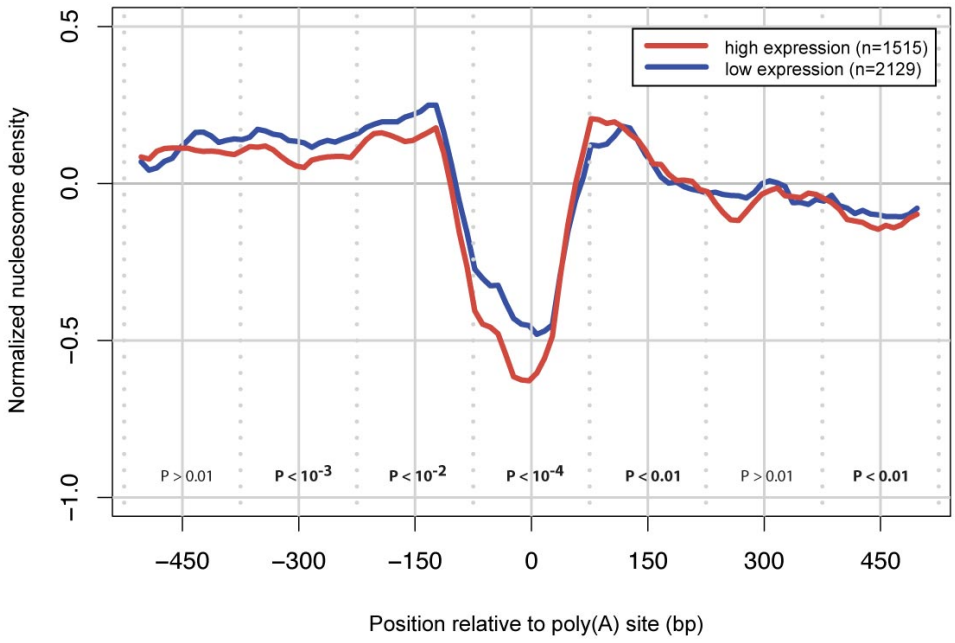

Figure S4: Nucleosome distribution around constitutive poly(A) sites from transcriptionally inactive (blue) and active (red) genes. Read density values (Schones et al., 2008) data) were normalized and smoothed as in Figure 1. Wilcoxon rank sum P-values shown for 150 bp windows centered on the indicated positions.