# Targeting and Function of Mammalian MicroRNAs

by

Kyle Kai-How Farh

B.S., Computer Science (2001)
Rice University

Submitted to the Department of Biology in Partial Fulfillment of the Requirement for the Degree of Doctor of Philosophy in Biology

at the

Massachusetts Insitute of Technology

Jan 2009
[FEBRUARY 2009]

Signature of Author............................................................................................
Department of Biology
January 22, 2009

Certified by............................................................................................
David P. Bartel
Professor of Biology
Thesis Supervisor

Accepted by............................................................................................
Professor of Biology
Graduate Student Committee
Department of Biology

Targeting and Function of Mammalian MicroRNAs

Kyle Kai-How Farh

Submitted to the Department of Biology on January 22, 2009 in
Partial Fulfillment of the Requirement for the Degree of
Doctor of Philosophy in Biology

## ABSTRACT

In the span of a few short years, animal microRNAs have become recognized as broad regulators of gene expression, largely in part due to our improved understanding of how animal microRNAs recognize their targets.

Crucial to microRNA targeting are the ~7-nt seed sites complementary to nucleotides 2-8 at the 5' end of the microRNA. We show that protein-coding genes preferentially expressed at the same time and place as a highly expressed microRNA have evolved their 3' UTR sequence to specifically avoid seed sites matching that microRNA. In contrast, conserved sites appear to be preferentially expressed in developmental states prior to microRNA expression, and are downregulated upon induction of that microRNA. Combined with the result that both conserved and nonconserved seed sites are generally functional, our findings extend the direct and indirect influence of mammalian microRNAs to the majority of protein-coding genes.

Although seed sites account for much of the specificity of microRNA regulation, they are not always sufficient for repression, suggesting the contribution of additional specificity determinants. Combining independent computational and experimental approaches, we found five general features associated with site efficacy: AU-rich nucleotide composition near the site, proximity to sites for co-expressed microRNAs, pairing outside of the seed region at microRNA nucleotides 13-16, and positioning within the 3' UTR at least 15nt from the stop codon and away from the center of long UTRs. By incorporating these five features, we are able to explain much of the differences in site efficacy for both exogenously added microRNAs and for endogenous microRNA-message interactions.

We further refined the seed site motif involved in microRNA repression, by demonstrating experimentally an Adenosine preference across from the unpaired first nucleotide of the microRNA and ranking the relative effectiveness of different classes of seed sites. Although sites lacking perfect seed pairing were generally ineffective, a fraction of these sites were supplemented by detectable compensatory 3' pairing. In addition, by extending our conservation analysis to 11 genomes, we show that the confidence with which conserved target sites can be predicted is a function of the conservation of the seed site itself relative to the conservation of surrounding sequence. This allows individual conserved sites to be assigned a confidence score reflecting the likelihood that the site is being conserved due to selection rather than by chance.

Thesis Supervisor: David P. Bartel
Title: Professor of Biology

# Acknowledgements

I'd like to thank Professor David Bartel, my advisor, who I am also happy to identify as my coach, my scientific collaborator, and my friend; and to my committee, for their constant enthusiasm, guidance, and patience. I would also like to extend a big thank you to all members of the Bartel lab, past and present, for their friendship, community, and willingness to discuss both good and bad ideas. In particular, I want to thank postdoctoral fellow Dr. Andrew Grimson, my steadfast scientific co-worker *par excellence*, who not only complemented me in our collaborations with his experimental prowess and his remarkable insights both experimental and computational; but also his being awake at all hours when I would be asleep, ensuring that progress would be made around the clock. I'd like to thank my family, who could sympathize and encourage throughout the course of the PhD, particularly from the perspective of having experienced the process themselves. Finally, I'd like to thank Hannah Chang, my best friend, for her kindness, support, and good humor throughout the whole graduate school process, and her understanding of many a late evening at work.

# Table of Contents

# Introduction

## The Discovery of MicroRNAs and Their Distribution on The Tree of Life

The first microRNA target to be discovered was the *C. elegans lin-14* gene, although it would be some time before its significance as the founding example of a new, widespread mode of gene regulation would be fully appreciated. Mutations in *lin-14* were found to affect the timing of larval development, with gain-of-function mutations resulting in reiteration of early larval stages, and loss-of-function mutations resulting in precocious adult molting (Ambros and Horvitz, 1984). Loss-of-function mutations in the *lin-4* gene also resulted in reiteration of early larval stages (Chalfie et al., 1981), similar to the *lin-14* gain-of-function mutant. When the two loss-of-function mutants were crossed, animals carrying both mutations displayed the precocious phenotype, implying that *lin-4* acted upstream of *lin-14* and was a negative regulator of *lin-14* (Ambros, 1989). Further investigation showed that *lin-4* did not code for a protein, but rather a ~22-nucleotide non-coding RNA (Lee et al., 1993). Moreover, this small RNA had imperfect complementarity to sequences in the *lin-14* 3' UTR (Lee et al., 1993; Wightman et al., 1993) that a previous study had shown were lost in the *lin-14* gain-of-function mutants (Wightman et al., 1991). These experiments established *lin-4* as the first microRNA to be discovered, and strongly implied that *lin-4* negatively regulated *lin-14* by binding to complementary sites in the *lin-14* 3' UTR (Lee et al., 1993; Wightman et al., 1993). Wightman and colleagues also noted that *lin-4* appeared to decrease the protein levels of *lin-14* without impacting its mRNA levels, as measured by an RNA protection assay (Wightman et al., 1993). These results implied that the mechanism of negative regulation by *lin-4* was translational repression. For several years, the *lin-4::lin-14* interaction, and a second interaction, *lin-4::lin-28* (Moss et al., 1997), were the only isolated examples of microRNA regulation, until the identification of *let-7* and its target *lin-41* (Reinhart et al., 2000). Like *lin-4*, *let-7* was temporally expressed at a specific stage of development in *C. elegans*, and loss of *let-7* activity resulted in reiteration of larval cell fates during adulthood, while overexpression of *let-7* resulted in precocious adult cell fates during larval development. *let-7* orthologs were subsequently identified in vertebrates, ascidians, hemichordates, mollusks, annelids, and arthropods (Pasquinelli et al., 2000).

The discovery that *let-7* was deeply conserved among both vertebrates and invertebrates led to the notion that other similar small regulatory RNAs were likely to exist. This was further supported by the incidental observation of endogenous ~21-nt small RNAs in a *Drosophila* cell-free lysate by researchers investigating the mechanism of RNA interference (Elbashir et al., 2001). The following year, about a hundred microRNAs were cloned from *C. elegans*, *Drosophila*, and man (Lagos-

Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Many of the *C. elegans* microRNAs discovered in the initial cloning experiments were deeply evolutionarily conserved, sharing mammalian and *Drosophila* homologs. Plant microRNAs were also cloned out of *Arabidopsis* (Reinhart et al., 2002), although the lack of a common ancestral microRNA suggested that contemporary microRNAs evolved independently in the plant and animal kingdoms. Plant species were also missing the critical proteins Drosha and Pasha/DGCR8, later known to be responsible for the biogenesis of most animal microRNAs (Han et al., 2006; Lee et al., 2003). Deep cloning of basal metazoan species has confirmed that microRNAs also exist in the cnidarian *Nematostella vectensis* and the demosponge *Amphimedon queenslandica* (Grimson et al., 2008). Of the cloned *Nematostella* microRNAs, only miR-100 appears to have homologs in fly, worms, and mammals, suggesting that it may be the most deeply conserved animal microRNA. *Amphimedon* is the furthest diverged extant lineage on the metazoan phylogenic tree, has no known microRNAs in common with other species, and has pre-microRNA lengths that are widely diverged from those observed in worm, fly, and human, although it has Drosha and Pasha/DGCR8 homologs. Together, these results are beginning to frame our understanding of how metazoan microRNA regulation may have evolved. The basic protein machinery for elaborating microRNAs and siRNAs existed prior to the divergence of plants and animals, in the form of Dicer and Ago proteins. From *Nematostella* onwards, the metazoan microRNA machinery appears to have become more evolutionarily fixed, as evidenced by the closer resemblance of *Nematostella* pre-microRNA lengths to those in worm, fly, and human and the greater uniformity of *Nematostella* pre-microRNA lengths, while individual microRNA sequences continued to be largely fluid. Progress in genome sequencing has been invaluable for the discovery of new microRNAs, as this has allowed newly sequenced reads to be immediately confirmed against the genome (The *C. elegans* sequencing consortium, 1998). Reads which do not correspond to genomic sequence, or which do not appear in hairpin precursors, are unlikely to be microRNAs.

### Structure and Maturation of MicroRNA Genes

The availability of the genomes for *C. elegans*, *Drosophila*, mouse, and human also provided important clues about the structure and maturation of microRNA genes. Instead of originating from long stretches of double-stranded RNA like siRNAs, the newly discovered microRNA genes tended to be present at loci unrelated to their known targets and in regions of the genome not annotated to have bidirectional transcription (Bartel, 2004). The original *lin-4* experiments in *C. elegans* showed that *lin-4* appeared to derive from a ~61-nt RNA stem-loop precursor containing the full sequence of the ~22-nt mature microRNA along its 5' fold-back arm (Lee et al., 1993). Many of the newly cloned microRNAs also had ~65-nt long precursors identified during northern blotting, and these were termed pre-microRNAs. Pre-microRNAs had 5' phosphates and 2-nt 3' overhangs typical of cleavage products of an RNase III-type endonuclease, suggesting that they were once part of a longer transcript, which was termed the pri-microRNA (Lee et al., 2003; Lee et al., 2002). Cellular localization experiments established that microRNA maturation was a two step process, requiring processing of pri-microRNAs into pre-microRNAs in the nucleus, and processing of pre-microRNAs into mature microRNAs in the cytoplasm (Lee et al., 2002).

Further investigation into the cleavage of pri-microRNAs into pre-microRNAs revealed that Drosha, an RNase III-type endonuclease, performs the first cut ~11-nt from the base of the pri-microRNA stem loop, and that a partner protein, Pasha/DGCR8, is necessary for the recognition of the stem-loop and its flanking single-stranded RNA segments (Han et al., 2006; Lee et al., 2003). The Drosha cut occurs in the nucleus, and the pre-microRNA product is exported to the cytoplasm by Exportin5 (Lund et al., 2004; Yi et al., 2003). The second cut at the top of the stem of the pre-microRNA is accomplished in the cytoplasm by Dicer, another RNase III-type endonuclease, known for its role in cleaving double-stranded RNA at 21-nt to 22-nt intervals during the production of siRNAs (Zamore et al., 2000). Dicer is required for both microRNA and siRNA biogenesis, and depletion of Dicer results in the accumulation of the *let-7* pre-microRNA precursor in *Drosophila* (Hutvagner et al., 2001). Additionally, a small number of ~22-nt small RNAs were cloned from the opposite arm of pre-microRNA precursor stem-loops, complimentary to the microRNA with 2-nt 3' overhangs at both ends of the duplex, and were termed microRNA* sequences (Lau et al., 2001). These microRNA* sequences were recognized to originate from the opposite arm of the pre-microRNA, following RNase III-type cleavage by Drosha and Dicer. Although microRNAs and microRNA* sequences are indistinguishable in their structure and biogenesis, both exhibiting the 5' phosphates and 2-nt overhangs typical of cleavage by RNase III endonucleases, microRNAs are typically cloned at frequencies ~100 times that of microRNA* sequences due to their greater incorporation into the mature RNA-induced silencing complex (RISC) (Lim et al., 2003b). This asymmetry favors the strand of a microRNA-microRNA* duplex that has weaker base pairing at its 5' end, possibly due to differential activity of a helicase that unwinds the microRNA-microRNA* duplex (Schwarz et al., 2003). Following maturation, the microRNA strand is loaded into the RISC, while the complementary passenger strand is discarded and presumably degraded rapidly. The cloned microRNA* sequences may represent either passenger strands that have not yet been degraded or minor incorporation of the less favorable strand into the RISC. Two lines of evidence suggest that some microRNA / microRNA* duplexes may have dual functional strands, rather than these cloned microRNA* sequences merely being transient intermediates. First of all, experiments with siRNAs have shown that both strands of a transfected microRNA / microRNA* duplex can downregulate their targets simultaneously (Jackson et al., 2003). Second, conservation analysis shows that some microRNA-microRNA* pairs, such as miR-10 and miR-10* in Drosophila, may have conserved targets above background along both functional strands (Ruby et al., 2007b; Stark et al., 2007a).

Analysis of microRNA and pre-microRNA characteristics, such as nucleotide composition, stem-loop length, and loop size, and taking into account conservation to other species, has enabled computational prediction of new microRNAs from sequence (Lai et al., 2003; Lim et al., 2003a; Lim et al., 2003b). Based on the sensitivity of the MirScan algorithm at detecting known microRNA genes, an upper bound of 255 human microRNAs was predicted, assuming that Mirscan score and cloning frequency were independent variables. (Lim et al., 2003a). However, mouse and human microRNAs that were discovered later tended to be both less deeply conserved and were cloned at lower frequencies (Berezikov et al., 2006; Houbaviy et al., 2003; Seitz et al., 2004). This implied that more recently evolved microRNAs are expressed at lower levels or in a smaller number of cells or cell types than the more deeply conserved microRNA families. More recent high-throughput cloning has revealed a large number of species-specific microRNAs, most of which are cloned at low frequencies (Ruby et al., 2006). Most of these species-specific microRNAs are presumably recently evolved, suggesting that the sequence characteristics recognized by Drosha and DGCR8 are common enough that new microRNAs can readily be derived from random sequences over the course of neutral evolution, and that the birth of a new microRNA is frequently not detrimental to the fitness of the organism, at least when expressed at low levels or in limited cell types. A better appreciation of where and when these more newly evolved microRNAs are expressed, and how they recognize their targets, will greatly enhance our understanding of how new microRNAs and their targets evolve.

### Organization and Expression of MicroRNA Genes

The initial cloning of large numbers of microRNA genes in *C. elegans, Drosophila,* and human revealed that microRNA genes are often located in introns (Lagos-Quintana et al., 2003; Lau et al., 2001). The presence of microRNAs in the introns of protein coding genes suggested that the mechanism of pri-microRNA transcription was PolII for at least some of these transcripts. Chromatin immunoprecipitation experiments confirmed that PolII was indeed bound to the promoters of pri-microRNAs, and that pri-microRNA transcripts were 5' capped with a 3' polyA tail (Lee et al., 2004). Furthermore, intronic microRNAs and their host genes tended to follow similar patterns of expression (Baskerville and Bartel, 2005). In some cases, the identity of the protein gene has provided valuable clues to the function and localization of the intronic microRNA. For example, the blood-vessel specific miR-126 microRNA is found in the seventh intron of the EGFL7 gene, an endothelial cell-derived growth factor with roles in angiogenesis (Parker et al., 2004; Rodriguez et al., 2004). EGFL7 knockout mice were documented with embryonic lethal angiogenesis defects attributed to loss of the EGFL7 protein product, but newer work with mouse models with smaller deletions now suggests that these phenotypes were largely due to the loss of miR-126 (Kuhnert et al., 2008). In addition, microRNAs are sometimes found in clusters and have similar patterns of expression, as would be expected if they were processed by Drosha from the same primary transcript (Lau et al., 2001). Some of these clusters of microRNAs contain microRNA genes that have similar sequences, such as the miR-15~16 cluster in the intron of the DBCL gene (Cimmino et al., 2005), and may have evolved via duplication, while other microRNA gene clusters, such as the miR-1~133 cluster, harbor microRNAs that appear to have few sequence characteristics in common. Investigation of clusters of microRNAs has led to the prediction and validation of additional microRNAs that are members of known cluster but have been difficult to find by cloning, either because of their low expression in the cell types that have being cloned thus far, or because their sequence presents technical difficulties for the cloning procedure (Berezikov et al., 2005; Lau et al., 2001; Ohler et al., 2004). Understanding the function of clustered microRNAs is particularly interesting in the context of combinatorial regulation, because clustered microRNAs would be expected to share co-expression and could act in concert on the same mRNA.

MicroRNAs have shown great variability in the timing and location of their expression. Prior to the cloning of additional microRNAs, *lin-4* and *let-7* were initially termed small temporal RNAs (stRNAs) because their expression was timed to coincide with specific developmental stages (Pasquinelli et al., 2000). The expression of lin-4 and its pre-microRNA coincided with the downregulation of the protein product of the *lin-14* gene during the transition from the *C. elegans* larva to adult stage, while *let-7* was expressed later, coinciding with downregulation of its target *lin-41* (Lee et al., 1993; Reinhart et al., 2000). Both *lin-4* and *let-7* were expressed in a broad range of tissues, with tightly regulated temporal control. When new microRNAs were cloned from mammalian tissues, different microRNAs had much higher cloning frequencies in different tissue types (Lagos-Quintana et al., 2002). The differential expression of microRNAs in various tissue types was confirmed by microarray analysis and northern blotting, with microRNAs such as miR-1 showing exclusive expression in heart and skeletal muscle, miR-122 in the liver, and miR-223 in the hematopoetic system (Baskerville and Bartel, 2005; Lee and Ambros, 2001; Sempere et al., 2004). An extensive series of *in situ* hybridization experiments in zebrafish demonstrated the complexity of microRNA expression, with different brain microRNAs being expressed in different regions of the brain or in neurons at different stages of proliferation and differentiation, and other microRNAs distributed in specific cell types throughout the body (Wienholds et al., 2005). The evidence from the *in situ* experiments agreed with previous northern blotting and microarray data in showing that the largest concentration of microRNAs detected were preferentially expressed in the brain (~30% of known microRNAs). Approximately two-thirds of the deeply conserved vertebrate microRNAs, of which around ~70 distinct families exist, showed tissue-specific expression in the zebrafish *in situ* experiments. Nearly every tissue had a microRNA specifically expressed in it, with some tissue types having multiple microRNAs, such as miR-1 and miR-133 in skeletal muscle and miR-192 and miR-194 in the gastrointestinal tract. In many of these cases where multiple microRNAs were specifically expressed in a single tissue type, the microRNAs were part of a cluster. For most of these microRNAs, determining whether the microRNA is tissue specific or cell type specific will require higher resolution experiments.

In contrast to the tissue-specific expression that characterizes the majority of deeply conserved vertebrate microRNAs, other vertebrate microRNAs have well-documented stage-specific expression, and were typically ubiquitously expressed in the zebrafish *in situ* experiments. Several microRNAs have been cloned only from mouse ES cells (Houbaviy et al., 2003). Expression of these genes appears to be driven by the ES cell-specific transcription factors such as *Nanog* and *Oct4*, which are bound to the promoter that transcribes this microRNA cluster according to ChIPseq experiments (Marson et al., 2008). The fact that these microRNAs have not been cloned from later developmental stages suggests that the role of these microRNAs is limited to ES cells and these microRNAs are downregulated along with the transcription factors driving their expression during differentiation. Interestingly, the ES cell-specific microRNAs in mammals share the nucleotides at positions 2-8 in common with the miR-430 microRNA family in zebrafish. MicroRNAs of the miR-430 family in zebrafish are strongly expressed post-fertilization at the initiation of zygotic transcription, and play a role in suppressing the translation of maternal transcripts and accelerating their degradation (Giraldez et al., 2005; Giraldez et al., 2006). Loss of miR-430 expression in Dicer-null zebrafish embryos results in a phenotype with multiple abnormalities, most strikingly in brain development.

Although PolII transcription followed by Drosha processing appears to the dominant mode of microRNA transcription and maturation, there appear to be at least two other pathways in which microRNAs are generated. Transcription of Alu repeats is performed by PolIII, and a subset of pri-microRNA transcripts interspersed with Alu repeats appear to be transcribed by PolIII (Borchert et al., 2006). Introns may also be excised to form pre-microRNA precursors directly via splicing, skipping the requirement for Drosha altogether (Okamura et al., 2007; Ruby et al., 2007a). The lariat intermediates are then debranched and fold into pre-microRNA stem loops for export to the cytoplasm and Dicer cleavage. Termed mirtrons, these microRNAs tend to be present in species such as *Drosophila* and *C. elegans* that have average intron sizes close to the ~65-nt length of a typical pre-microRNA stem-loop.

**Early Computational Attempts at Target Prediction**

In contrast to the relative ease with which large numbers of new microRNAs were cloned, assigning targets to these newly discovered microRNAs proved difficult. However, just as the newly sequenced genomes greatly assisted in the validation of new microRNAs, they also provided the opportunity to use computational approaches to pick new candidate microRNA targets out of the vast genomic sequence. For computational target prediction to be successful, this would require that the rules describing how microRNA targeting worked would be simple enough to be discovered, and accurate and consistent enough that new target sites could be described primarily based on sequence. Transcription factors had long presented a similar problem, and computational approaches alone had proven challenging for predicting transcription factor binding sites and activities (MacIsaac and Fraenkel, 2006). Part of the difficulty was that rather than being a clearly defined pattern of nucleotides, transcription factor binding motifs were "fuzzy", with low information content, partly due to the complexity of DNA-protein interactions. On the other hand, the binding interactions between two strands of RNA were well understood and predictable (Hofacker, 1994). Furthermore, microRNA silencing complexes, despite containing different microRNA sequences, could be expected to recognize their targets and behave in an analogous manner due to their shared core protein machinery, whereas each transcription factor family was its own special case. MicroRNAs also appeared to act primarily in one direction: dowregulation, whereas the action of transcription factors could be unpredictable based on the transcriptional context.

Aside from these theoretical concerns, several practical factors made deriving the rules for microRNA targeting a challenge: first, given only the three known *C. elegans* targets, it was unclear if any insights gleaned from these interactions would apply to other organisms or other targets in general; second, the incomplete complementarity between the microRNA and its target in these three examples increased the complexity of finding target recognition rules that would be flexible enough to admit new candidates with similar patterns of base pairing while being stringent enough to avoid the huge numbers of potential matches over the large genome sequence space.

Although the discovery of plant microRNAs lagged behind their discovery in animals, the first progress towards predicting the targets of microRNAs was made in plants. Despite sharing many characteristics in common with animal microRNAs, such as similar lengths and similar hairpin precursors (Ambros et al., 2003), in contrast to the microRNA targets described in *C. elegans*, plant target sites had extensive Watson-Crick base pairing throughout the microRNA-target duplex (Rhoades et al., 2002), and were subject to cleavage in a manner analogous to sites of perfectly complementary siRNAs (Llave et al., 2002; Tang et al., 2003). A plant microRNA and its target would be complementary throughout their sequence, generally with no more than a couple mismatches or GU-wobbles. When present, asymmetric bulges tended to fall on the mRNA strand, rather than the microRNA strand. Because of the stringency required to find a near-perfect match to a ~22-nucleotide sequence, each microRNA only had one site, or a couple such sites, whereas shuffled control sequences had, on average, virtually no sites with such extensive base pairing. Many of these microRNA::target pairs were conserved to other flowering plants, with conservation of both the microRNA and its target site. Plant predictions could also be specifically validated using a 5' RACE assay to look for mRNA cleavage fragments at the expected cleavage position across from positions 10 and 11 of the microRNA (Llave et al., 2002). However, when the same prediction methods used in plants were applied to animal genomes, few such targets were identified. This was perhaps not unexpected, as the known targets of *lin-4* and *let-7* did not have sufficient pairing to qualify as plant microRNA sites. Furthermore, the mechanism of action of animal and plant microRNAs appeared to be different; while plant microRNAs cleaved their targets akin to perfectly complementary siRNAs, animal microRNAs appeared to downregulate their targets primarily through translational repression (Reinhart et al., 2000; Wightman et al., 1993). Although it was shown that animal microRNAs could also guide cleavage of extensively paired complementary sites (Hutvagner and Zamore, 2002; Yekta et al., 2004), it was recognized that animal microRNAs primarily downregulated their targets through a mechanism quite different from that observed for plant microRNAs and siRNAs.

Over the next couple of years, multiple algorithms were devised for predicting the targets of animal microRNAs. In general, the approaches could be lumped into two major strategies. One strategy was to screen for new animal microRNA sites based on similar patterns of pairing to the known microRNA sites from *C. elegans* – *lin-4::lin-14*, *lin-4::lin-28*, and *let-7::lin-41* (John et al., 2004; Kiriakidou et al., 2004; Rajewsky and Socci, 2004; Stark et al., 2003). Each of these original three targets displayed partial complementarity between the microRNA and the target site, with the greatest complementarity at the 5' and 3' ends of the microRNA::target duplex, and little or no complementarity in the middle of the duplex at the nucleotide positions necessary for siRNA cleavage

(Doench et al., 2003). For the case of the *let-7::lin-41* interaction, it also appeared that GU-wobble base pairs could also be tolerated alongside Watson-Crick base pairing (Reinhart et al., 2000). An alternate strategy was to systematically examine motifs complementary to the microRNAs to find preferential conservation of microRNA target site motifs compared to chance (Lewis et al., 2003). This approach had the benefit of being less reliant on characteristics that might be idiosyncratic to the small number of known microRNA sites.

The first published algorithms for finding microRNA targets all used the folding free energy for a hypothetical double stranded RNA helix formed between the microRNA sequence and the target sequence to predict the strength of a potential microRNA-target (Hofacker, 1994; John et al., 2004; Kiriakidou et al., 2004; Lewis et al., 2003; Rajewsky and Socci, 2004; Stark et al., 2003). Several features of the known sites in *C. elegans* suggested that the rules for predicting the free energy of helix formation between two free single strands of RNA were applicable to microRNA-target interactions – first, the presence of non-canonical GU-wobble base pairs which are typically well tolerated in RNA-RNA duplexes, and secondly, the large extent of pairing found in known targets suggested that a free energy threshold needed to be met for the target sequence to be recognized by the microRNA. With the exception of the TargetScan (Lewis et al., 2003), the algorithms for finding targets were heavily biased by the known *lin-4::lin-14*, *lin-4::lin-28*, and *let-7::lin-41* sites, and in each of these studies, these three founding targets comfortably passed under the criteria (TargetScan did not recognize the *let-7::lin-41* due to the GU-wobble in one site and the bulged nucleotide in the other). For each prediction algorithm, additional criteria were included based on these known targets. For instance, the study by Rajewsky *et al.* noted that the known microRNA::target duplexes had a "nucleus" of several continuous G-C base pairs, which were hypothesized to be significant thermodynamically; consequentially, the algorithm was based on finding such nuclei, and then folding the microRNA and the sequence flanking the "nucleus" to determine if the hypothetical duplex passed a free energy cutoff (Rajewsky and Socci, 2004). Based on the tendency of the three known *C. elegans* targets to have more base pairing at the 5' and 3' ends but less in the center, the Diana-MicroT algorithm required that potential sites pass a pairing energy cutoff at both the 5' and 3' ends of the microRNA::target duplex, as well as requiring a symmetric or asymmetric bulge of a certain number of nucleotides in the middle separating the two end duplexes (Kiriakidou et al., 2004). Somewhat unsurprisingly, the three known *C. elegans* sites were among the highest scoring targets identified in the computational screen. Other groups perceived that pairing at the 5' end of the microRNA appeared to be more consequential. The algorithm by Stark *et al.* searched the *Drosophila* genome for UTRs with at least two sites matching the first eight nucleotides of the microRNA while allowing a single GU-wobble to accommodate the *let-7::lin-41* target (Stark et al., 2003). The MiRanda algorithm by Enright *et al.* weighted pairing at positions 1-11 of the 5' end of the microRNA (Enright et al., 2003). The importance of pairing to the 5' end of the microRNA would later be borne out by both computational and experimental evidence; however these early studies failed to demonstrate that the microRNA target sites found using these algorithms were significantly more conserved compared to chance (i.e., when the algorithm was repeated substituting the microRNA sequences with rigorously selected random controls.) Although the MiRanda study attempted such an analysis, their controls were not corrected for dinucleotide frequency and contained dinulceotides with significantly lower conservation rates than the actual dinucleotide composition of microRNA sequences, resulting in an inaccurately inflated signal-to-background ratio (John et al., 2004). In contrast, the study by Stark et al, despite not showing signal for conserved microRNA target sites over background, proved valuable by detecting many new *Drosophila* microRNA sites which were then experimentally validated, effectively doubling the number of known animal microRNA targets (Stark et al., 2003).

**The Seed Hypothesis and Supporting Conservation Analysis**

Among the newly cloned fly microRNAs, Eric Lai reported that several had 5' ends that were the Watson-Crick complement to known regulatory motifs found in *Drosophila* 3' UTRs, such as the K-Box (Lai, 2002; Lai et al., 1998; Lai and Posakony, 1997). The target sites for the *bantam* microRNA in the *hid* 3' UTR also showed extensive complementarity at the 5' end of the microRNA, and represented the first time a metazoan microRNA target interaction was initially identified through computational screening (Brennecke et al., 2003). Later that year, Lewis *et al.* demonstrated on a genome-wide scale that sites with 7-nt motifs corresponding to the seeds (nucleotides 2-8 at the 5' end) of mammalian microRNAs were strongly conserved above chance using carefully matched control sequences (Lewis et al., 2003). The controls were matched to microRNAs on the basis of dinucleotide content and similar motif frequency in one genome. The number of conserved instances of microRNA sites was termed the signal, and the number of conserved control sequences was termed the noise, also referred to as the background. Thus, the controls served as an estimate for the extent of background conservation expected for motifs not under selection. The number of conserved sites minus the number of expected control sequences (signal-above-background) gave the total number of microRNA sites conserved by selection. The number of conserved sites divided by the number of expected control sequences (signal-to-background ratio) gave the likelihood that a conserved site was being conserved due to selection rather than chance, and indicated the confidence in the prediction. The controls were carefully selected to match both the motif frequency in the genome and the dinucleotide composition, ensuring that the controls would be conserved by chance at approximately the same frequency as the real microRNA motifs. Although earlier studies had hinted at the importance of pairing to the 5' end of the microRNA (Stark et al., 2003), the TargetScan analysis provided the first empirical evidence that a computational algorithm was successfully identifying microRNA targets compared to chance.

In the original Targetscan analysis, genes were only counted as targets if they contained two or more sites conserved to mouse and rat, and sites were required to not only have a perfect seed match, but also to meet a free energy cutoff for pairing between the message and the additional 3' region of the microRNA outside of the seed site. This resulted in ~400 predictions with an estimated false positive rate of 31%, giving a signal-to-background ratio of 3.2 (Lewis et al., 2003). The specificity could be improved at the cost of sensitivity by requiring conservation to the pufferfish *Fugu Rubripes*. In a revised analysis performed a little over a year later, the 3' pairing constraint was lifted as well as the requirement for multiple sites to one gene (Lewis et al., 2005). In addition, the analysis was performed for aligned 3' UTRs of human / mouse / rat / dog / chicken, a major advantage since the pufferfish used in the earlier study was in many cases too divergent to prove informative. For microRNAs not beginning with a uridine, the algorithm asked for an adenosine across from the nucleotide at position 1 of the microRNA, instead of the Watson-Crick complement; an adenosine across from position 1 of the microRNA was shown to be superior on signal-to-noise analysis. Overall, these changes slightly improved signal-to-background to 3.5 for sites with either the 7mer-m8 (nucleotides 2-8) or the 7mer-A1 (nucleotides 2-7 + A1 anchor), while boosting sensitivity. The signal-above-background was estimated to be 7453 targets under selection when requiring at least a conserved 6mer site (nucleotides 2-7), an improvement of nearly 30-fold over the previous predictions. Most of this improvement was due to allowing single seed sites, since requiring two or more seed matches as in the original TargetScan analysis reduced the number of target sites above background by 90% while increasing signal-to-background only modestly. These results indicated that over a third of mammalian protein-coding genes were under selection to maintain pairing to one or more microRNAs, and that the vast majority of these target sites were single 7mers or 8mers. The study also showed that sites containing even one mismatch or GU-wobble in the seed had little or no signal-above-background, indicating that they are much less likely to be under selection, and are therefore mostly nonfunctional.

The updated 2005 TargetScan analysis also investigated the patterns of 3' UTR conservation and their impact on microRNA target prediction. Mammalian 3' UTRs varied widely in their conservation, with more highly conserved 3' UTRs containing both a higher density of conserved target sites and a higher density of conservation not associated with microRNA seeds. While the increase in conserved microRNA seeds exceeded the increase in non-microRNA-related conservation, the signal-to-background ratio was actually lower for these highly conserved 3' UTRs, reflecting the difficulty in confidently predicting whether the conserved microRNAs seeds were a result of selection as opposed to chance in the context of such a high background level of conservation. One approach was to require that microRNA conservation fell in islands of conservation, flanked by non-conserved sequence. For a 2-7 nt seed match, this approach increased signal-to-background from 2.4 to 2.9, at the cost of decreased sensitivity (loss of ~20% of target sites.) The wide variation in the rate of 3' UTR divergence presented a challenge to this type of conservation analysis, because highly diverged 3' UTRs would be most informative when compared against closely related species, while less diverged 3' UTRs would be most informative when compared against distantly related species. Furthermore, adding more species to the conservation analysis would lead to significant losses in sensitivity due to gaps in sequencing, annotation, and alignment. These concerns would be addressed using a more flexible phylogenetic branch-length algorithm, discussed in chapter 4 (Friedman et al., 2008; Stark et al., 2007b).

Similar computational results were reported in a study by Brennecke *et al.* in *Drosophila* and Krek *et al.* in mammals (Brennecke et al., 2005; Krek et al., 2005). The study by Brennecke *et al.* included experimental evidence demonstrating that Watson-Crick base pairing in the seed was critical for target recognition, and that a single mismatch or GU-wobble in the seed sequence was enough to de-repress the reporter gene (Brennecke et al., 2005). This study also addressed the issue of sites with imperfect seed matches but extensive 3' pairing, such as the *let-7::lin-41* interaction. Using a *Drosophila* wing imaginal disc assay, artificial sites with mismatches in the seed region but extensive compensatory 3' pairing (contiguous Watson Crick pairing of 15+ nucleotides in length) were successfully downregulated. Experimental investigation of 3' pairing was particularly useful since the extent of 3' pairing required to compensate for a mismatch to the seed would require such specificity that very few examples existed naturally in the genome, making computational analysis more difficult. This study also showed computationally that beyond a very high cutoff for 3' pairing energy, genuine microRNA sequences had many more occurrences of supplemental conserved 3' pairing than control microRNA sequences consisting of genuine seed sites but with scrambled 3' ends. Given the large number of microRNAs that share the same seed sequence but have different 3' sequences (the same microRNA family), one potential role for 3' pairing would be to distinguish between different microRNAs from the same family. For instance, several microRNAs in *C. elegans* share the *let-7* seed sequence but have different 3' ends (Lim et al., 2003b). Compensatory 3' pairing to *let-7* could explain why the *lin-41* 3' UTR has two conserved sites with extensive 3' pairing to *let-7* but a mismatch and a GU-wobble in the seed sites themselves (Brennecke et al., 2005; Lewis et al., 2005).

Additional support for the seed hypothesis came from a study by Xie *et al.*, which systematically catalogued conserved sequence elements in mammalian promoters and 3' UTRs (Xie et al., 2005). Because the authors were not limiting their investigation to known microRNAs, they employed an alphabet consisting of A, C, G, T, N, and the six two-fold degenerate characters to find all 6mers (possibly with a central gap) with preferential conservation above background. Motif conservation score was calculated by comparing the conservation rate of the motif to its expected conservation, determined by sampling 1000 motifs of the same length and redundancy at 1000 genomic loci. The motifs found in 3' UTRs were characterized by having strong directionality, with much higher conservation in the sense orientation than in the antisense orientation, as would be expected for recognition of single-stranded RNA elements as opposed to the symmetric conservation observed for double-stranded DNA elements. Of the 106 highly conserved motifs discovered in 3' UTRs, 72 corresponded to 8mers, often with an A at position 8. Because the algorithm used by Xie *et al.* clustered similar motifs together, it is not straightforward to assign these motifs clusters to specific microRNAs. However, of the top 50 8mer motifs in 3' UTRs, ~31 can be confidently assigned to known microRNAs, most of which are deeply conserved to all vertebrates and some to invertebrates as well. Several new microRNAs were also proposed, based on 8mer conservation and the presence of the motif in a conserved stem loop elsewhere in the genome. Twelve of these predictions were tested and six of these predicted microRNAs were experimentally supported, although the PCR assay used to support these candidates was later shown to produce some false positives (Ruby et al., 2006). It is unclear how many of these predictions represent real microRNAs, as they have not been confirmed by northern blot and it is unknown what their tissue of expression might be, because the PCR products were obtained from pooled samples of 10 different tissue types. Of the 60 remaining 3' UTR motifs not associated with known microRNAs, 11 correspond to the PUF family of RNA-binding proteins, one corresponds to the polyadenylation signal, and the remaining motifs are largely unknown (Spassov and Jurecic, 2003). Approximately 45% of the most highly conserved motifs and 18 of the top 20 8mer motifs in 3' UTRs correspond to known microRNA genes.

The study by Krek *et al.* introduced the PicTar algorithm that calculated the maximum likelihood score for individual 3' UTRs based on the probability that a given 3' UTR was regulated by a combination of one or more microRNAs (Krek et al., 2005). One of the microRNA targets that was experimentally verified was *MTPN*, a gene with a highly conserved 3' UTR and conserved seed sites corresponding to a dozen microRNAs. Co-expression of either miR-375, *let-7*, or miR-124 using siRNA transfection into N2A neuroblastoma cells produced ~20% repression in each case, whereas repression up to ~40% was observed when all three microRNAs were simultaneously transfected. *MTPN* was selected in this study because it was believed to be expressed in pancreatic islets along with miR-375, *let-7*, and miR-124. However, more recent cloning has shown that human pancreatic islets do not in fact express miR-124, calling into question the biological plausibility of these three microRNAs co-regulating *MTPN* expression (Landgraf et al., 2007). Nevertheless, the study highlighted important questions for further study, particularly for genes with highly conserved 3' UTRs and many conserved target sites. First, the presence of multiple conserved seed sites in a 3' UTR does not imply that these microRNAs exert combinatorial action, as many of these target sites are associated with microRNAs that are known to be expressed in non-overlapping tissue types. Some of these sites may be explained by incidental conservation, but the high signal-to-background ratio of many of these microRNAs implies that the majority are functional. Understanding how microRNA sites have cooperatively evolved relies on a greater understanding of how microRNA expression and target gene expression are linked, since repression of the target gene can only occur when the microRNA(s) targeting it are present in the same cell type at the same developmental stage. A second question is: when multiple microRNAs are present with the target gene in the proper spatial and temporal context, how large is their combined effect? If the RISC machinery is indeed common to all microRNAs, the effect of multiple sites to different microRNAs would be expected to be similar to the binding of multiple sites to the same microRNA, which had been shown to be geometrically additive for an artificially constructed tandem arrangement of microRNA sites (Doench and Sharp, 2004). Attempts to answer both of these questions are topics in chapters I and II.

### Mechanisms of MicroRNA Downregulation

At the time of the conservation studies, it was understood that animal microRNAs could downregulate their targets either through direct cleavage, or translational repression through incomplete pairing or seed-type pairing (Bartel, 2004). Despite the differences in their origin, microRNAs and siRNAs were indistinguishable in their modes of action, and the choice of cleavage versus translational repression was determined largely by the extent of base-pairing to the target (Hammond et al., 2000; Hutvagner and Zamore, 2002). Direct cleavage of the target strand occurred in the context of perfect or near-perfect base pairing, with hydrolysis of the phosphodiester bond across from positions 10 and 11 of the microRNA guide strand, counting from the 5' end (Elbashir et al., 2001; Martinez and Tuschl, 2004). This cleavage was catalyzed by Ago2 and resulted in two fragments, with a 3' hydroxyl on one end and a 5' phosphate on the other (Liu et al., 2004). Following cleavage, the 5' and 3' cleavage fragments could be degraded by the cellular exonuclease machinery (Orban and Izaurralde, 2005; Valencia-Sanchez et al., 2006). Artificially designed siRNAs that silence their target genes through this mechanism have become widespread in their experimental use, and some are in clinical development (Fire et al., 1998; Kleinman et al., 2008; Yang et al., 2008), while endogenous animal microRNA cleavage appears rare, with *miR-196::HOX-B8* and miR-127/miR-136 and Rtl/Peg11 being the only confirmed examples (Davis et al., 2005; Yekta et al., 2004).

In contrast, the vast majority of animal microRNA interactions consist of incomplete base pairing to the target site, typically 2-7 nucleotide seed sites supplemented by an invariant adenosine at position 1 or a Watson-Crick match at position 8, resulting in translational repression. Due to the initial experiments demonstrating translational repression without RNA downregulation using *lin-4* in *C. elegans*, it was assumed that this mode of regulation would remain invisible to expression microarrays (Wightman et al., 1993). However, based on microarray studies of multiple different siRNAs to IGF and MAPK, Jackson *et al.* found significant off-target effects from siRNA transfections (Jackson et al., 2003). These off-targets differed for different siRNAs targeted at MAPK, suggesting they were not secondary effects of knocking down the target gene, and several of the off-targets corresponded to the Watson-Crick complement of the 5' end of transfected siRNAs. However, it was not known at the time that the active siRNA strand of the siRNA::siRNA* duplex could be specified by the strand having weaker pairing at its 5' end (Schwarz et al., 2003). Consequently, approximately half of the transfected siRNAs favored the wrong strand, and thus the observed off-targets did not consistently correspond to the complement of the intended siRNA sequence, and the significance of these findings was not fully appreciated.

In 2005, Lim *et al.* reported that upon transfection of miR-1 and miR-124 in HeLa in a manner analogous to siRNA experiments, hundreds of genes were observed to be downregulated, and that the majority of these downregulated genes contained motifs corresponding to the seed sites to miR-1 and miR-124 (Lim et al., 2005). This provided further support for the seed site hypothesis, and indicated that previous conclusions about translational repression occurring independently of mRNA destabilization should be reconsidered, particularly in the context of a complementary microarray study where an inhibitor of miR-122, was shown to upregulate the targets of miR-122 in mouse (Krutzfeldt et al., 2005). These studies demonstrated the value of expression microarrays for measuring the functional effects of microRNAs on a genome wide scale, adding a valuable experimental method that could be used independently of the conservation analysis for further refining microRNA target prediction.

There were two additional insights gleaned from the Lim *et al.* study that would serve as a catalyst for studies examining the impact of microRNAs on their targets on a genome-wide scale. First, the introduction of miR-1 (a muscle-specific microRNA) and miR-124 (a brain-specific microRNA) resulted in a shift of the expression patterns towards patterns of muscle and brain, respectively. The genes downregulated by transfection of miR-124 in HeLa were among those that were expressed at the lowest levels in brain compared to other tissues. Since genes are typically expressed only in a subset of tissues, this would indicate that many of the genes downregulated in this assay by miR-124 (a brain-specific microRNA), are actually not expressed in brain and do not share co-expression with the microRNA. A second major insight was that the great majority of downregulated sites were nonconserved, and would have been missed by computational predictions, since these rely on conservation. Both of these results were challenging to explain at the time and would have to await a more thorough analysis of the expression patterns of microRNAs and their targets on a genome-wide scale.

Following the publication of the Lim *et al.* study, the original experiments demonstrating translational repression of *lin-14* by *lin-4* were repeated, this time using northern blots rather than an RNA protection assay, and the mRNA levels of *lin-14* were shown to decrease in response to *lin-4* regulation (Bagga et al., 2005). Nevertheless, the decrease in *lin-14* protein levels exceeded the decrease in *lin-14* mRNA levels. While mRNA degradation appears to be a widespread outcome of seed-type recognition, in some cases the extent of miRNA repression on translational output may be greater than what can be explained by the decrease in mRNA levels. Although there is a general correlation between amount of downregulation on the mRNA and the protein level, the nature their relationship has not been conclusively established (Baek et al., 2008; Selbach et al., 2008). Different groups of researchers have employed IRES reporters to bypass the various steps of translation initiation in order to understand whether translational repression operates at the initiation step or during elongation, without arriving at a consensus (Humphreys et al., 2005; Petersen et al., 2006; Pillai et al., 2005). Other studies have pointed to a pathway beginning with sequestration of targeted messages into P-bodies, followed by accelerated shortening of the poly-A tail, and concluding with eventual degradation by a mechanism distinct from Ago2 cleavage, most likely through the action of

3' to 5' and 5' to 3' exonucleases following deadenylation and decapping (Bagga et al., 2005; Giraldez et al., 2006; Liu et al., 2005; Sen and Blau, 2005).

### The Widespread Impact of Metazoan MicroRNAs on 3' UTR Sequence Evolution

*The full study discussed in this section is featured in Chapter 1 (Farh et al., 2005).*

The tissue-specific shifts observed during transfection of miR-1 and miR-124 were initially counterintuitive, since this implied that many of the targets of microRNAs would never be in the same cell as the microRNA. For instance, many of the genes downregulated by miR-124 (a brain specific microRNA) appeared to have intensities on the chip that were so low that they were most likely not expressed in the brain at all. When we divided microRNA targets into conserved and nonconserved, however, we found that these two sets behaved markedly differently. Nonconserved targets tended to absent in the tissue of microRNA expression, whereas conserved targets were expressed, but at a low relative level compared to their expression in other tissues. This suggested that conserved targets were being expressed in the presence of the microRNA, and that their expression was being dampened. This was further supported by the finding that when analyzing expression data from differentiation time courses (myoblast differentiation or egg to embryo), genes with conserved sites tended to be expressed at high levels prior to expression of the microRNA, and were then swiftly downregulated following induction of that microRNA.

Because the microRNA seed site is relatively short, it can be expected to occur at a certain frequency in 3' UTR sequence by chance. Based on the length and nucleotide content of 3' UTRs, the probability for each microRNA to have a site present in each UTR was calculated using a Markov model. Nonconserved sites occurred at roughly the frequency expected by chance in genes not expressing the microRNA, suggesting that the sites in these genes, although potentially functional in the presence of the microRNA, typically do not see the microRNA and therefore evolve neutrally and are primarily occurring by chance. On the other hand, genes that were preferentially expressed in the same tissue as the microRNA were markedly depleted in sites to that microRNA, with their 3' UTR sequence containing approximately 50% of the sites expected by chance. This finding could be explained in two different ways – either mRNAs containing sites to the microRNA were being degraded by the action of the microRNA directly, or genes that are preferentially expressed in the tissue (and presumably have important roles in the tissue) have evolved their sequences to avoid inhibition by the microRNA. In order to distinguish the two, we eliminated direct effects due to microRNA-mediated downregulation by repeating the analysis after excluding genes with seed sites in mouse and considering only human sequence data. The depletion signature was observed using mouse expression data but sites in human 3' UTR sequences, indicating that the depletion signature was a consequence of genes evolving to avoid targeting rather than merely direct effects of the microRNA on mRNA levels. Site depletion has been estimated to affect hundreds of genes per microRNA for highly expressed tissue-specific microRNAs, and may be even more pervasive for more globally expressed microRNAs.

Stark *et al.* published similar results for the expression of microRNAs and their targets in *Drosophila* (Stark et al., 2005). However, instead of using microarrays to measure mRNA levels, they used *in situ* data, which resulted in more discrete expression patterns and cell-specific resolution. One difference between the studies is that, based on the *in situ* data, microRNAs appear to have virtually no overlap with their conserved targets, whereas the microarray is able to detect the target transcript at a lower level. Because of potential contamination from a minority population of cells that do not express the microRNA, it is unclear whether the signal on the microarray represents a genuine lower level of expression that the *in situ* is not sensitive to. Based on these findings, Stark and Brennecke have posited that one role of microRNAs is to prevent the accidental leaky expression of genes that need to be turned off in that tissue. While there may be instances where microRNAs play such a role, microarray studies using sorted cell populations (thus avoiding the problem of contamination with minority cell populations) indicate that a large number of conserved microRNA targets appear to be co-expressed with the microRNA (Farh et al., 2005; Shkumatava et al., submitted). Despite the co-expression of microRNAs and their targets, it is unclear how many of these represent "tuning"-type targets that are being adjusted to the correct level of expression, versus "switch"-type targets that are rendered effectively functionally absent by microRNA despite their low level of expression in the tissue (Bartel and Chen, 2004). Stark *et al.* also adapted their analysis to show that the selective avoidance of microRNA targets applied to entire categories of genes – for instance, ubiquitously expressed ribosomal proteins had shorter 3' UTRs than transcription factors, and also had fewer conserved sites per unit length, presumably because they are under selection to avoid targeting by all microRNAs (Stark et al., 2005). These distinctions were also evident in other metazoan species, but absent in plants and yeast, which lack widespread miRNA targeting (plants) or miRNAs altogether (yeast).

### MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing

*The full study discussed in this section is featured in Chapter 2 (Grimson et al., 2007)*

While an overwhelming body of evidence supports the seed hypothesis of microRNA targeting, not all seed sites appear to be functional, and when tested in a reporter assay, the extent of repression for sites with identical motifs varies greatly (Didiano and Hobert, 2006; Farh et al., 2005). These findings strongly suggest that other features, perhaps present in the target message but outside of the seed itself, also contribute heavily to site efficacy. Even among genes that are preferentially expressed in the same tissue as a microRNA, the extent of site depletion is around ~50%, suggesting that approximately half of newly evolved sites appear in a context that is optimal for their efficacy and are selected against, while the other half emerges in unfavorable contexts and therefore have little or no effect and are allowed to persist (Farh et al., 2005).

To identify features associated with greater efficacy, we employed three independent methods that had been used previously to show the effects of microRNA targeting on a genome-wide scale: conservation analysis (Lewis et al., 2005), downregulation on the microarray (Lim et al., 2005), and site depletion analysis (Farh et al., 2005). Features associated with increased efficacy would presumably be associated with conserved sites rather than nonconserved sites; with sites that cause greater downregulation on the array; and with genes that were more heavily depleted in the site depletion analysis. The importance of using all three methods is that individually each method has potential drawbacks. For instance, conservation can be confounded by other, non-microRNA related effects that produce indirect statistical correlations between conserved sites and features of conserved sequence. Microarray data is free of such concerns, but it is not clear to what extent repression on the mRNA level correlates with repression on the protein level for individual genes, whereas the conservation and depletion analyses have the advantage of speaking to the ultimate effects on the fitness of the animal. A fourth independent method, the luciferase assay, was saved for the validation of these determinants and to provide experimental support that these determinants described behavior at the protein output level, rather than just the level of mRNA expression.

An important difference in the conservation analysis used in this paper that sets it apart from previous work is that when conservation was used to find associated specificity determinants, we were typically guided by signal-above-background as a metric instead of signal-to-background ratio (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Lewis et al., 2003). The distinction is particularly important for cases where the extent of sequence conservation is significantly different. For instance, classes of genes such as transcription factors have particularly well-conserved 3' UTRs, and the two ends of the 3' UTR tend to be better conserved relative to the sequence in the middle. In highly conserved sequences, the high level of background conservation adds to the count of both conserved microRNA motifs and conserved control motifs; hence these sequences generally have a net lower signal-to-noise ratio, even though they tend to have greater signal-above-background per unit length. Put differently, such sequences tend to be more heavily favored by microRNAs and have a greater density of sites under selection, but because they are also being conserved for other reasons not directly related to microRNA seed sites, we have less confidence that a conserved microRNA site in the sequence is due to selective maintenance of microRNA targeting rather than by chance. These findings suggest that the dense conservation of multiple microRNA sites either brings in a good deal of incidental "collateral" conservation, or that the same specificity determinants that favor microRNA targeting also favor other modes of posttranscriptional regulation.

Using these approaches, five specificity determinants were discovered. They are briefly summarized below:

1. Multiple sites generally act independently, but two sites to the same or different microRNAs that are between ~8 and ~40 nucleotides apart (end of the first site to the start of the next site) can be cooperative, increasing repression. While our work was in review, consistent results were published, reporting cooperativity from 13 to 35 nucleotides, as measured from the starting positions of both sites (Saetrom et al., 2007).
2. Additional pairing outside of the seed region improves site efficacy if it is centered at positions 13-16 of the microRNA and is Watson-Crick in nature.
3. Local A+U content immediately adjacent to the site is strongly correlated with site efficacy. This is believed to be, in whole or in part, a consequence of local secondary structure.
4. Sites are most effective in the 3' UTR and only marginally effective in the ORF. The transition between 3' UTR targeting and ORF targeting occurs ~14 nt into the start of the 3' UTR, suggesting that actively translating ribosomes may preclude productive microRNA binding.
5. Sites falling in the middle of long 3' UTRs are less effective.

In addition to these determinants, the strongest ($0^{th}$) determinant is the extent of the motif. Motifs with 7mer-m8 sites (corresponding to microRNA nucleotides 2-8) generally outperformed those with 7mer-A1 sites (corresponding to microRNA nucleotides 2-7 + an A at position 1), while both were stronger than the 6mer site (nts 2-7) and weaker than the 8mer site (nts 2-8 + A1). This is more fully described in Chapter 3.

Using the data from 11 microRNA transfections in HeLa (miR-1, miR-7, miR-9, miR-122, miR-124, miR-128, miR-132, miR-133, miR-142, miR-148, miR-181), we constructed a quantitative model that allowed prediction of site efficacy from sequence alone, without considering site conservation. To validate these predictions, we selected 25 genes with single miR-25 7mer-m8 sites, and tested these in a luciferase reporter assay in 293T cells. The validation was conducted using a different microRNA (outside of the transfected dataset) transfected into a different cell type (293T instead of HeLa) to ensure that the results could be maximally generalized. Of the 13 genes with high predicted efficacy, 12/13 were significantly downregulated, while of the 12 genes with low predicted efficacy, 0/12 were significantly downregulated. Overall, these results were more striking than what would have been expected from the array data, and indicated not only that the features described in this study apply to the final protein output, but also that their magnitude may have been underestimated due to the measurement noise of the microarray. The predictions were also applied to in vivo examples, using published microarray datasets – the knockout and rescue of miR-430 in zebrafish embryos, the antisense inhibition of miR-122 in mouse liver, and the knockout of miR-155 in mouse T-cells (Giraldez et al., 2006; Krutzfeldt et al., 2005; Rodriguez et al., 2007). In each of these experiments, the model successfully predicted the relative repression for different microRNA sites in the animal.

One outstanding question is whether the five features described here represent the full range of specificity determinants affecting microRNA targeting. Conserved microRNA sites tend to perform better than nonconserved microRNA sites, and although approximately 40-50% of this difference is accounted for by controlling for the specificity determinants described in this study, the remainder remains to be explained. This discrepancy could be due to other unknown determinants, mis-scoring or incomplete understanding of the determinants described here, or the ability of conserved sites to arrange their determinants in a more synergistic manner (our model assumes independence of the specificity determinants.) Other studies have proposed additional specificity determinants, particularly those relating to secondary structure (Long et al., 2007; Robins et al., 2005; Zhao et al., 2005). However, when implementing these algorithms, we found that while some were informative, they were generally inferior to the method of simply counting the local A+U content adjacent to the site. This may be because of the difficulty of accurately predicting secondary structure in an mRNA, with large stretches of neutrally evolving sequence, or because the silencing complex has a primary sequence preference for A's and U's adjacent to the site. A more detailed discussion on using mRNA structure to improve target predictions follows near the end of this chapter.

### Prediction of Mammalian MicroRNA Target Sites Under Selection

*The full study discussed in this section is featured in Chapter 3.*

The large number of potential microRNA target sites presents a challenge for experimental investigation, given the time and resources required to pursue individual leads. Two features that can aid in distinguishing high priority targets are the probability that the site is under selection and the degree of expected repression. From the standpoint of biological relevance, the most interesting sites are those that are deeply conserved and likely to be maintained by selection. However, evidence for selection does not directly speak to the magnitude of the repression. For instance, millions of years of selection may have settled upon an optimal interaction that does not repress the target microRNA completely, but rather mildly tunes it to the proper level (Bartel and Chen, 2004). Therefore, sites should be chosen based on the expected degree of downregulation as well as the likelihood that the site is under selection. Such a strategy would be the most likely to identify a target whose in vivo inhibition would manifest as an observable phenotype.

The availability of newly sequenced mammalian genomes (Rabbit, Cow, Tenrec, Elephant, Armadillo, Opossum) presented an opportunity to perform a more sensitive conserved site analysis, including the assignment of a confidence score to individual sites indicating the likelihood that a conserved site is being conserved due to microRNA-related selection rather than occurring due to chance. Combining microarray expression data and conservation data, we show that the confidence with which conserved target sites can be predicted is a function of the conservation of the seed site itself relative to the conservation of surrounding 3' UTR sequence. This method successfully identifies microRNA target sites while adjusting for both quickly and slowly evolving genes. While the TargetScanS study had previously shown that signal-to-noise could be improved by searching within shorter "islands" of conservation (Lewis et al., 2005), this approach assigns a quantitative confidence score to each individual site, reflecting the likelihood that the site is being conserved due to selection rather than by chance.

Along with the improved canonical seed sites analysis, we also revisited sites with incomplete or mismatched base pairing from the TargetScanS study (Lewis et al., 2005). While these sites were marginally effective both in conservation analyses and on the array, the high frequency with which mismatched seed sites occurred suggested that a large number of such sites are under selection, even though their effects may be subtle and we generally cannot reliably distinguish the sites under selection from background. One exception to this rule is the small fraction of mismatched seed sites supplemented by extensive, conserved compensatory 3' pairing, including the miR-196::HOXB8 interaction. While the number of such sites is small, it significantly exceeds background expectation, suggesting that in mammals, as with the *let-7::lin-41* site in *C. elegans*, a handful of these sites have been preferentially maintained during the course of evolution.

## Summary

Literature reporting the cloning of mammalian microRNAs did not appear until 2001, and progress towards identifying the principles by which they recognize their targets has been even more recent. Although the field is young, our understanding of this class of regulatory molecules has moved forward rapidly. In contrast to transcription factors, another class of molecules with whom microRNAs have often been compared, microRNA regulation has been in many ways a much more computationally tractable problem, because of the predictable nature of RNA-RNA interactions compared to DNA-protein interactions, and because microRNA silencing complexes containing different microRNAs appear to recognize and engage their targets in an analogous manner due to shared protein machinery. Over the course of four years, through the development of both computational and experimental tools, the microRNA target prediction field has evolved from early over-fitted efforts to its current state, where one can begin to imagine that predicting function might be possible on the basis of sequence alone. Recently, the results of microRNA knockouts in mice have underscored both the importance of being able to identify the targets of microRNAs and connect them to observed phenotypes and the challenges associated with making sense of hundreds of direct and secondary microRNA targets (Johnnidis et al., 2008; Rodriguez et al., 2007; Zhao et al., 2007). The efforts described in this thesis towards developing accurate prediction algorithms capable of ranking targets both by their predicted efficacy and their probability of being under selection will hopefully serve as a temporary harbor for exploring the vast sea ahead.

## References

Ambros, V. (1989). A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. Cell *57*, 49-57.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M., *et al.* (2003). A uniform system for microRNA annotation. Rna *9*, 277-279.

Ambros, V., and Horvitz, H. R. (1984). Heterochronic mutants of the nematode Caenorhabditis elegans. Science *226*, 409-416.

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

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.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartel, D. P., and Chen, C. Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet *5*, 396-400.

Baskerville, S., and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA *11*, 241-247.

Berezikov, E., Guryev, V., van de Belt, J., Wienholds, E., Plasterk, R. H., and Cuppen, E. (2005). Phylogenetic shadowing and computational identification of human microRNA genes. Cell *120*, 21-24.

Berezikov, E., Thuemmler, F., van Laake, L. W., Kondova, I., Bontrop, R., Cuppen, E., and Plasterk, R. H. (2006). Diversity of microRNAs in human and chimpanzee brain. Nat Genet *38*, 1375-1377.

Borchert, G. M., Lanier, W., and Davidson, B. L. (2006). RNA polymerase III transcribes human microRNAs. Nat Struct Mol Biol *13*, 1097-1101.

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

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

Chalfie, M., Horvitz, H. R., and Sulston, J. E. (1981). Mutations that lead to reiterations in the cell lineages of *C. elegans*. Cell *24*, 59-69.

Cimmino, A., Calin, G. A., Fabbri, M., Iorio, M. V., Ferracin, M., Shimizu, M., Wojcik, S. E., Aqeilan, R. I., Zupo, S., Dono, M., *et al.* (2005). miR-15 and miR-16 induce apoptosis by targeting *BCL2*. Proc Natl Acad Sci U S A *102*, 13944-13949.

consortium, C. e. s. (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012-2018.

Davis, E., Caiment, F., Tordoir, X., Cavaille, J., Ferguson-Smith, A., Cockett, N., Georges, M., and Charlier, C. (2005). RNAi-mediated allelic trans-interaction at the imprinted *Rtl1/Peg11* locus. Curr Biol *15*, 743-749.

Didiano, D., and Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol *13*, 849-851.

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

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

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

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

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.

Fire, A., Xu, S., Montgomery, M. K., Kostas, S. A., Driver, S. E., and Mello, C. C. (1998). Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. Nature *391*, 806-811.

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2008). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res.

Giraldez, A. J., Cinalli, R. M., Glasner, M. E., Enright, A. J., Thomson, J. M., Baskerville, S., Hammond, S. M., Bartel, D. P., and Schier, A. F. (2005). MicroRNAs regulate brain morphogenesis in zebrafish. Science *308*, 833-838.

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.

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

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S., and Bartel, D. P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193-1197.

Hammond, S. M., Bernstein, E., Beach, D., and Hannon, G. J. (2000). An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. Nature *404*, 293-296.

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

Hofacker, I., Fontana, W, Stadler, PF, Bonhoeffer, S, Tacker, M, Shuster P (1994). Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie, 167-168.

Houbaviy, H. B., Murray, M. F., and Sharp, P. A. (2003). Embryonic stem cell-specific MicroRNAs. Dev Cell *5*, 351-358.

Humphreys, D. T., Westman, B. J., Martin, D. I., and Preiss, T. (2005). MicroRNAs control translation initiation by inhibiting eukaryotic initiation factor 4E/cap and poly(A) tail function. Proc Natl Acad Sci U S A *102*, 16961-16966.

Hutvagner, G., McLachlan, J., Pasquinelli, A. E., Balint, E., Tuschl, T., and Zamore, P. D. (2001). A cellular function for the RNA-interference enzyme Dicer in the maturation of the *let-7* small temporal RNA. Science *293*, 834-838.

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

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.

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

Johnnidis, J. B., Harris, M. H., Wheeler, R. T., Stehling-Sun, S., Lam, M. H., Kirak, O., Brummelkamp, T. R., Fleming, M. D., and Camargo, F. D. (2008). Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. Nature *451*, 1125-1129.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. Genes Dev *18*, 1165-1178.

Kleinman, M. E., Yamada, K., Takeda, A., Chandrasekaran, V., Nozaki, M., Baffi, J. Z., Albuquerque, R. J., Yamasaki, S., Itaya, M., Pan, Y., *et al.* (2008). Sequence- and target-independent angiogenesis suppression by siRNA via TLR3. Nature *452*, 591-597.

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

Krutzfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M., and Stoffel, M. (2005). Silencing of microRNAs in vivo with 'antagomirs'. Nature *438*, 685-689.

Kuhnert, F., Mancuso, M. R., Hampton, J., Stankunas, K., Asano, T., Chen, C. Z., and Kuo, C. J. (2008). Attribution of vascular phenotypes of the murine Egfl7 locus to the microRNA miR-126. Development *135*, 3989-3993.

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

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

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

Lai, E. C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet *30*, 363-364.

Lai, E. C., Burks, C., and Posakony, J. W. (1998). The K box, a conserved 3' UTR sequence motif, negatively regulates accumulation of *Enhancer of split* Complex transcripts. Development *125*, 4077-4088.

Lai, E. C., and Posakony, J. W. (1997). The Bearded box, a novel 3' UTR sequence motif, mediates negative post-transcriptional regulation of *Bearded* and *Enhancer of split* Complex gene expression. Development *124*, 4847-4856.

Lai, E. C., Tomancak, P., Williams, R. W., and Rubin, G. M. (2003). Computational identification of Drosophila microRNA genes. Genome Biol *4*, R42.

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

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science *294*, 858-862.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science *294*, 862-864.

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.

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., and Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. Nature *425*, 415-419.

Lee, Y., Jeon, K., Lee, J. T., Kim, S., and Kim, V. N. (2002). MicroRNA maturation: stepwise processing and subcellular localization. Embo J *21*, 4663-4670.

Lee, Y., Kim, M., Han, J., Yeom, K. H., Lee, S., Baek, S. H., and Kim, V. N. (2004). MicroRNA genes are transcribed by RNA polymerase II. Embo J *23*, 4051-4060.

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.

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.

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

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.

Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., and Bartel, D. P. (2003b). The microRNAs of *Caenorhabditis elegans*. Genes Dev *17*, 991-1008.

Liu, J., Carmell, M. A., Rivas, F. V., Marsden, C. G., Thomson, J. M., Song, J. J., Hammond, S. M., Joshua-Tor, L., and Hannon, G. J. (2004). Argonaute2 is the catalytic engine of mammalian RNAi. Science *305*, 1437-1441.

Liu, J., Valencia-Sanchez, M. A., Hannon, G. J., and Parker, R. (2005). MicroRNA-dependent localization of targeted mRNAs to mammalian P-bodies. Nat Cell Biol *7*, 719-723.

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.

Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. Nat Struct Mol Biol *14*, 287-294.

Lund, E., Guttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004). Nuclear export of microRNA precursors. Science *303*, 95-98.

MacIsaac, K. D., and Fraenkel, E. (2006). Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol *2*, e36.

Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J., *et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

Martinez, J., and Tuschl, T. (2004). RISC is a 5' phosphomonoester-producing RNA endonuclease. Genes Dev *18*, 975-980.

Moss, E. G., Lee, R. C., and Ambros, V. (1997). The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. Cell *88*, 637-646.

Ohler, U., Yekta, S., Lim, L. P., Bartel, D. P., and Burge, C. B. (2004). Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. Rna *10*, 1309-1322.

Okamura, K., Hagen, J. W., Duan, H., Tyler, D. M., and Lai, E. C. (2007). The mirtron pathway generates microRNA-class regulatory RNAs in *Drosophila.* Cell *130*, 89-100.

Orban, T. I., and Izaurralde, E. (2005). Decay of mRNAs targeted by RISC requires XRN1, the Ski complex, and the exosome. Rna *11*, 459-469.

Parker, L. H., Schmidt, M., Jin, S. W., Gray, A. M., Beis, D., Pham, T., Frantz, G., Palmieri, S., Hillan, K., Stainier, D. Y., *et al.* (2004). The endothelial-cell-derived secreted factor Egfl7 regulates vascular tube formation. Nature *428*, 754-758.

Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M. I., Maller, B., Hayward, D. C., Ball, E. E., Degnan, B., Muller, P., *et al.* (2000). Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. Nature *408*, 86-89.

Petersen, C. P., Bordeleau, M. E., Pelletier, J., and Sharp, P. A. (2006). Short RNAs repress translation after initiation in mammalian cells. Mol Cell *21*, 533-542.

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.

Rajewsky, N., and Socci, N. D. (2004). Computational identification of microRNA targets. Dev Biol *267*, 529-535.

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature *403*, 901-906.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. Genes Dev *16*, 1616-1626.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. Cell *110*, 513-520.

Robins, H., Li, Y., and Padgett, R. W. (2005). Incorporating structure to predict microRNA targets. Proc Natl Acad Sci U S A *102*, 4006-4009.

Rodriguez, A., Griffiths-Jones, S., Ashurst, J. L., and Bradley, A. (2004). Identification of mammalian microRNA host genes and transcription units. Genome Res *14*, 1902-1910.

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.

Ruby, J. G., Jan, C., Player, C., Axtell, M. J., Lee, W., Nusbaum, C., Ge, H., and Bartel, D. P. (2006). Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. Cell *127*, 1193-1207.

Ruby, J. G., Jan, C. H., and Bartel, D. P. (2007a). Intronic microRNA precursors that bypass Drosha processing. Nature *448*, 83-86.

Ruby, J. G., Stark, A., Johnston, W. K., Kellis, M., Bartel, D. P., and Lai, E. C. (2007b). Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. Genome Res *17*, 1850-1864.

Saetrom, P., Heale, B. S., Snove, O., Jr., 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.

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

Seitz, H., Royo, H., Bortolin, M. L., Lin, S. P., Ferguson-Smith, A. C., and Cavaille, J. (2004). A large imprinted microRNA gene cluster at the mouse Dlk1-Gtl2 domain. Genome Res *14*, 1741-1748.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.

Sempere, L. F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., and Ambros, V. (2004). Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol *5*, R13.

Sen, G. L., and Blau, H. M. (2005). Argonaute 2/RISC resides in sites of mammalian mRNA decay known as cytoplasmic bodies. Nat Cell Biol *7*, 633-636.

Shkumatava, A., Stark, A., and Bartel, D. P. (submitted).

Spassov, D. S., and Jurecic, R. (2003). The PUF family of RNA-binding proteins: does evolutionarily conserved structure equal conserved function? IUBMB Life *55*, 359-366.

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.

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

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

Stark, A., Lin, M. F., Kheradpour, P., Pedersen, J. S., Parts, L., Carlson, J. W., Crosby, M. A., Rasmussen, M. D., Roy, S., Deoras, A. N.*, et al.* (2007b). Discovery of functional elements in 12 *Drosophila* genomes using evolutionary signatures. Nature *450*, 219-232.

Tang, G., Reinhart, B. J., Bartel, D. P., and Zamore, P. D. (2003). A biochemical framework for RNA silencing in plants. Genes Dev *17*, 49-63.

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.

Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H. R., Kauppinen, S., and Plasterk, R. H. (2005). MicroRNA expression in zebrafish embryonic development. Science *309*, 310-311.

Wightman, B., Burglin, T. R., Gatto, J., Arasu, P., and Ruvkun, G. (1991). Negative regulatory sequences in the *lin-14* 3'-untranslated region are necessary to generate a temporal switch during *Caenorhabditis elegans* development. Genes Dev *5*, 1813-1824.

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

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

Yang, Z., Stratton, C., Francis, P. J., Kleinman, M. E., Tan, P. L., Gibbs, D., Tong, Z., Chen, H., Constantine, R., Yang, X.*, et al.* (2008). Toll-like receptor 3 and geographic atrophy in age-related macular degeneration. N Engl J Med *359*, 1456-1463.

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

Yi, R., Qin, Y., Macara, I. G., and Cullen, B. R. (2003). Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes Dev *17*, 3011-3016.

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

Zhao, Y., Ransom, J. F., Li, A., Vedantham, V., von Drehle, M., Muth, A. N., Tsuchihashi, T., McManus, M. T., Schwartz, R. J., and Srivastava, D. (2007). Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. Cell *129*, 303-317.

Zhao, Y., Samal, E., and Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets *Hand2* during cardiogenesis. Nature *436*, 214-220.

# Chapter I.

The following chapter includes the manuscript, figures, and supplementary data for the paper titled "The widespread impact of MicroRNAs on mRNA repression and evolution", published in *Science* 2005. I performed the bulk of the computational analyses in the paper, while Ben Lewis performed analyses on the question of whether microRNA seed sites were more conserved compared to control sequences in regions of open secondary structure. On the experimental side, co-author Andrew Grimson performed the luciferase assays with the assistance of Wendy Johnston, and and Calvin Jan performed the northern blot for mir-7 in pituitary tissue. Andrew Grimson, David Bartel, and I each helped write and revise the text. Lee Lim contributed some early computational observations and aided us in revising the text.

# The Widespread Impact of Mammalian MicroRNAs on mRNA Repression and Evolution

**Kyle Kai-How Farh,[1]* Andrew Grimson,[1]* Calvin Jan,[1] Benjamin P. Lewis,[1,3] Wendy K. Johnston,[1] Lee P. Lim,[2] Christopher B. Burge,[3] David P. Bartel[1]†**

[1]Whitehead Institute for Biomedical Research, Department of Biology, Massachusetts Institute of Technology, and Howard Hughes Medical Institute, Nine Cambridge Center, Cambridge, MA 02142 USA

[2]Rosetta Inpharmatics, a wholly owned subsidiary of Merck and Co., 401 Terry Avenue N, Seattle, Washington 98109 USA

[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139 USA

*These authors contributed equally this work.

†To whom correspondence and requests for materials should be addressed.

dbartel@wi.mit.edu, 617-258-5287, fax 617-258-6768

**Thousands of mammalian mRNAs are under selective pressure to maintain 7-nucleotide sites matching microRNAs (miRNAs). We find that these conserved targets are often highly expressed at developmental stages prior to miRNA expression, and that their levels fall as the miRNA that targets them begins to accumulate. Nonconserved sites, which outnumber the conserved ten-to-one, also mediate repression. As a consequence, genes preferentially expressed at the same time and place as a miRNA have evolved to selectively avoid sites matching the miRNA. This phenomenon of selective avoidance extends to thousands of genes and enables spatial and temporal specificities of miRNAs to be revealed by finding tissues and developmental stages in which messages with corresponding sites are expressed at lower levels.**

MicroRNAs are an abundant class of endogenous ~22-nucleotide (nt) RNAs that specify posttranscriptional gene repression by basepairing to the messages of protein-coding genes (Ambros, 2004; Bartel, 2004). Hundreds of miRNAs have been identified in humans (Bartel, 2004), and thousands of messages are under selection to maintain pairing to miRNA seeds (nucleotides 2-7 of the miRNA), enabling regulatory targets of miRNAs to be predicted by simply searching 3' UTRs for evolutionary conserved 7-nt matches to miRNA seed regions (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005).

We used the mouse expression atlas (Su et al., 2004) to examine the expression of the predicted targets of six tissue-specific miRNAs: miR-1 and miR-133 (skeletal muscle), miR-9 and miR-124 (brain), miR-122 (liver) and miR-142-3p (hematopoietic organs and blood cells) [(Baskerville and Bartel, 2005; Lagos-Quintana et al., 2002; Sempere et al., 2004; Wienholds et al., 2005), fig. S1]. The 250 messages with conserved miR-133 sites were generally expressed in muscle but at lower levels in muscle than in other tissues (Fig. 1A). Likewise, predicted targets of the other miRNAs were usually at lower levels in the tissue expressing the miRNA than in other tissues (Fig. 1A). Brain-specific miR-9 and miR-124 displayed more complex patterns, perhaps reflecting the heterogeneous cell types within the brain.

The low relative expression of predicted targets in differentiated tissues raised the question of whether they might be more highly expressed earlier in differentiation, prior to miRNA expression. To address this, we analyzed expression profiles of myotube differentiation (Tomczak et al., 2004), during which miR-1 and miR-133 accumulate following cell-cycle arrest (Rao et al., unpublished data). Predicted targets of these muscle-specific miRNAs were preferentially high prior to miRNA expression then dropped as the miRNAs accumulated (Fig. 1B; fig. S3). The observation that miRNAs induced during differentiation tend to target messages highly expressed in the previous developmental stage suggests a function analogous to that proposed for plants, whereby miRNAs dampen the output of pre-existing messages to facilitate a more rapid and robust transition to a new expression program (Rhoades et al., 2002). The tendency of predicted targets to be expressed at substantial levels on the absolute scale (Fig. 1A, x-axis) further suggested that metazoan miRNAs are often optimizing protein output without eliminating it entirely (Bartel and Chen, 2004).

Our results are consistent with the idea that miRNAs are destabilizing many target messages to further define tissue-specific transcript profiles (Lim et al., 2005) but also leave open the possibility that many targets are repressed translationally without mRNA destabilization. If miRNAs were usually working in concert with transcriptional and other regulatory processes to down-regulate the same genes, then a correlation between conserved targeting and lower mRNA levels would be observed even for messages that miRNAs translationally repress without destabilizing.

Mammalian miRNA families have an average of ~200 conserved targets above estimated background, a figure approximately one tenth the number of 3' UTRs with 7-nt sites in a single genome (Krek et al., 2005; Lewis et al., 2005). Computational algorithms rely on evolutionary conservation to distinguish functional miRNA targets from the thousands of messages that would pair equally well; in contrast, the cell must rely on specificity determinants intrinsic to a single genome. To determine whether these nonconserved sites might be functional, we used reporter assays to compare repression mediated by conserved and nonconserved sites. We selected two targets of miR-1, predicted by TargetScan based on conservation in human, mouse and rat (Lewis et al., 2003) and six human UTRs that had comparable TargetScan scores in human but low or nonexistent scores in mouse or rat. When UTR fragments of ~0.5 kilobases containing the sites were placed in reporters, specific repression was observed for all eight (Fig. 2A). Analogous experiments with eight predictions from our more sensitive analysis, TargetScanS, which searches for conserved 7- or 8-nt matches (Lewis et al., 2005), and 17 genes with nonconserved matches also detected little difference between UTR fragments containing conserved and nonconserved sites (Fig. 2B), even when the concentration of transfected miRNA was titrated to suboptimal levels (fig. S4). Apparently, most nonconserved sites fortuitously reside in local contexts suitable for mediating repression and therefore have the potential to function when exposed to the miRNA. These results generalize previous work showing that in certain contexts 7- or 8-nt matches appear sufficient for miRNA-like regulation (Brennecke et al., 2005; Doench and Sharp, 2004; Lai et al., 2005). We conclude that additional recognition features, such as pairing to the remainder of the miRNA, accessible mRNA structure, or protein-binding sites, are usually dispensable, or occur so frequently that they impart little overall specificity (supporting online text).

To explore the impact of this vast potential for nonconserved targeting, we examined the expression of messages with nonconserved 7-nt matches to tissue-specific miRNAs, focusing first on messages with sites present in mouse but not in the orthologous human UTRs (Fig. 3A). In contrast to the conserved sites, the nonconserved sites had a propensity to fall in the UTRs of genes that were not expressed in the same tissue as the miRNA. Also striking was the depletion of sites among those genes that were most highly and specifically expressed in the tissue. Such depletion could result primarily from direct miRNA-mediated destabilization of messages (Lim et al., 2005), or some depletion might be from selective avoidance of sites—evolutionary pressure for messages highly specific to a tissue to lose sites for coexpressed miRNAs.

To distinguish between these two possibilities, we plotted the expression, in mouse, of genes that lacked sites in the mouse UTR but contained a site in the human ortholog. Because such messages would not be subject to miRNA-mediated destabilization in mouse, the depletion signal would vanish if it reflected only direct destabilization. However, the signal persisted; mouse genes expressed highly and specifically in the tissue were less likely to harbor sites in their human orthologs (Fig. 3B), indicating that genes preferentially co-expressed with the miRNA have evolved to avoid targeting by that miRNA. The enrichment for genes expressed at low levels also explained some of the many potentially functional nonconserved sites; they accumulate by chance, without consequence, in messages not co-expressed with the miRNA. The reduction in signal in Figure 3B compared to 3A hints that species-specific mRNA destabilization might also be frequent, presumably as both neutral and consequential species-specific targeting.

Quantifying selective depletion of sites among messages preferentially expressed in muscle indicated that ~420 of the 8511 genes of the expression atlas are under selective pressure to avoid miR-133 sites. These are "antitargets," an anticipated class of genes not observed previously (Bartel and Chen, 2004). The estimated numbers of antitargets for miR-1, miR-122, miR-142, miR-9 and miR-124 were 300, 190, 170, 240, and 440, respectively—comparable to the numbers of their conserved targets. Extrapolating to include other miRNA families that are also highly expressed with specific spatial or temporal expression patterns, we estimate that selective avoidance of miRNA targeting extends to thousands of genes (supporting online text). A signal for messages avoiding targeting in all tissue types would be harder to detect in our analysis. For some messages, acquiring miRNA pairing might be so detrimental that they are under selective pressure to have short UTRs, perhaps helping to explain why highly expressed "house-keeping" genes have substantially shorter UTRs than do other messages (Eisenberg and Levanon, 2003).

In addition to revealing target avoidance, thesedata extendresults of our heterologous reporter system (Fig. 2) into the animal, showing that 7-nt sites are often sufficient to specify a biological effect. Messages expressed highly and specifically in muscle are ~59% less likely than controls to possess a 7-nt match to muscle-specific miR-133 (Fig. 3A). For the other five miRNAs, this depletion averaged 45% (range 31–57%). This extent of depletion implies that as sites for highly expressed miRNAs emerge during sequence drift of UTRs, about half emerge in a context suitable for miRNA targeting—causing either mRNA destabilization or a selective disadvantage sufficient for preferential loss of the site from the gene pool.

Site depletion due to miRNA activity should occur specifically in tissue types expressing the miRNA. To explore the specificity of depletion, we used a modified Kolmogorov-Smirnov (KS) test to determine whether the set of genes with sites in either human or mouse orthologs were expressed at lower levels than cohorts of genes with the same estimated expectation for having sites, controlling for UTR length and nucleotide composition. In muscle, but not in T cells, the set of transcripts with a miR-133 site was depleted compared to control cohorts (Fig. 4A). Repeating the miR-133 analysis for all 61 tissues in the mouse atlas showed that this effect was pronounced in skeletal muscle and heart, the two tissues in which miR-133 is preferentially expressed. Plotting color-coded *P* values for relative depletion of transcripts with miR-133 sites illustrated a signature reflecting the tissue-specific influence of miR-133 (Fig. 4B, top row).

Signatures for all 73 miRNA families (representing 169 human miRNA genes) conserved among the four sequenced mammals and zebrafish were derived (fig. S7). For many miRNA families prominently expressed in specific tissues (Baskerville and Bartel, 2005; Lagos-Quintana et al., 2002; Sempere et al., 2004; Wienholds et al., 2005), the signatures corresponded to tissues in which these miRNAs are expressed (Fig. 4B). These included the six families featured in Figure 3, as well as let-7, miR-99, miR-10, miR-29, and miR-153 (brain), miR-30 (kidney), miR-194 (liver, gut, kidney), miR-141 and miR-200b (olfactory epithelium, gut), miR-96 (olfactory epithelium), and miR-375 (pituitary). miR-7 had highest signal in the pituitary. This miRNA is known to be preferentially expressed in the brain (Baskerville and Bartel, 2005; Sempere et al., 2004; Wienholds et al., 2005), but preferential expression in pituitary had not been noted. An RNA blot confirmed that miR-7 expression is highest in the pituitary (Fig. 4D).

Other miRNA families, including most described as having ubiquitous, complex, or undetectable expression patterns, were indistinguishable from controls (Fig. 4C, fig. S7). Nonetheless, some described as ubiquitous displayed stage-specific signatures. These included families in the miR-17~18~19a~20~19b~92 cluster, which had a strong embryo signature, consistent with their association with proliferation and cancer (He et al., 2005; Ota et al., 2004). The miR-302 family also had a strong early-embryo signature, consistent with its sequence similarity to the 17~92 proliferation cluster and its expression in embryonic stem cells (Houbaviy et al., 2003; Suh et al., 2004). The conserved targets of these embryonic miRNAs were preferentially at high levels in the oocyte and zygote then dropped to low levels in the blastocyst and embryo (Fig. 1C), as expected if these miRNAs help dampen expression of maternal transcripts.

A signal for motif conservation is a mainstay of bioinformatics and previously indicated the widespread scope of conserved miRNA targeting (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Xie et al., 2005), but a signal for absence of a motif is unusual. The ability to observe such a signal revealed an additional dimension to the impact of miRNAs on UTR evolution—a widespread potential for nonconserved targeting leading to the selective loss of many 7-nt sites. When considering conserved targeting, nonconserved targeting, and targeting avoidance, it is hard to escape the conclusion that miRNAs are influencing the expression or evolution of most mammalian mRNAs.

### References and Notes

1. D. P. Bartel, *Cell* **116**, 281 (2004).
2. V. Ambros, *Nature* **431**, 350 (2004).
3. B. P. Lewis, C. B. Burge, D. P. Bartel, *Cell* **120**, 15 (2005).
4. J. Brennecke, A. Stark, R. B. Russell, S. M. Cohen, *PLoS Biol* **3**, e85 (2005).
5. A. Krek *et al.*, *Nat Genet* **37**, 495 (2005).
6. A. I. Su *et al.*, *Proc Natl Acad Sci U S A* **101**, 6062 (2004).
7. M. Lagos-Quintana *et al.*, *Curr Biol* **12**, 735 (2002).
8. L. F. Sempere *et al.*, *Genome Biol* **5**, R13 (2004).
9. S. Baskerville, D. P. Bartel, *Rna* **11**, 241 (2005).
10. E. Wienholds *et al.*, *Science* **309**, 310 (2005).

11. K. K. Tomczak *et al.*, *Faseb J* **18**, 403 (2004).
12. P. K. Rao, M. Farkhondeh, S. Baskerville, H. F. Lodish, (unpublished data).
13. M. W. Rhoades *et al.*, *Cell* **110**, 513 (2002).
14. D. P. Bartel, C. Z. Chen, *Nat Rev Genet* **5**, 396 (2004).
15. L. P. Lim *et al.*, *Nature* **433**, 769 (2005).
16. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge, *Cell* **115**, 787 (2003).
17. J. G. Doench, P. A. Sharp, *Genes Dev* **18**, 504 (2004).
18. E. C. Lai, B. Tam, G. M. Rubin, *Genes Dev* **19**, 1067 (2005).
19. E. Eisenberg, E. Y. Levanon, *Trends Genet* **19**, 362 (2003).
20. A. Ota *et al.*, *Cancer Res* **64**, 3087 (2004).
21. L. He *et al.*, *Nature* **435**, 828 (2005).
22. H. B. Houbaviy, M. F. Murray, P. A. Sharp, *Dev Cell* **5**, 351 (2003).
23. M. R. Suh *et al.*, *Dev Biol* **270**, 488 (2004).
24. X. Xie *et al.*, *Nature* **434**, 338 (2005).

25. We thank Graham Ruby, Michael Axtell and Hannah Chang for helpful discussions. Supported by a predoctoral fellowships from the DOE (B.P.L.) and NSF (C.J.), and a postdoctoral fellowship and grants from the NIH (A.G, C.B.B., D.P.B.). D.P.B. is an HHMI Investigator.

**Figure Legends**

**Fig. 1.** Gene-density maps of conserved miRNA targets. (**A**) Predicted targets of miRNAs in tissues expressing the miRNAs. For muscle (large panel, left), the genes of the expression atlas were first placed in 61 equally populated bins along the x-axis and 61 equally populated bins along the y-axis. Along the x-axis genes were sorted based on whether they were expressed at low (left) or high (right) levels in muscle. Along the y-axis genes were sorted based on whether they were expressed higher (top) or lower (bottom) in muscle compared to other tissues. Predicted targets of miR-133 were then mapped onto this 61×61 grid. Local density (after background subtraction, fig S2, and smoothing) of miR-133 targets is color-coded, with regions of enrichment (red) or depletion (blue) shown (key at far right). Other miRNA–tissue pairs were analyzed analogously (smaller panels, right). (**B**) Time course of predicted targets during myoblast (C2C12) differentiation to myotubes, analyzed using a 24×24 grid. (**C**) Time course of predicted targets during mouse embryogenesis, analyzed as in (A). Predicted targets of let-7 are included for comparison in (B) and (C).

**Fig. 2.** MicroRNA-mediated repression of luciferase reporter genes containing 3' UTR fragments with conserved or nonconserved sites. (**A**) UTR fragments with TargetScan-like miR-1 sites. Luciferase activity from HeLa cells cotransfected with miRNA and wild-type reporters was normalized to that from cotransfection with mutant reporters with three point substitutions distrupting each seed match. The miR-124 transfections served as specificity controls. Error bars represent 3rd largest and smallest values among 12 replicates (one asterisk, $P < 0.01$; two asterisks, $P < 0.001$, Wilcoxon rank-sum test). (**B**) UTR fragments with TargetscanS-like miR-1 (top) and miR-124 (bottom) sites, analyzed as in (A).

**Fig. 3.** Density maps for genes with nonconserved sites. (**A**) Messages with site present in mouse UTR but absent in human ortholog. Panels are as in Figure 1, but enrichment is relative to matched cohorts (figs. S5 and S6), controlling for UTR length and nucleotide composition. (**B**) Messages with site present in human UTR but absent in orthologous mouse UTR, analyzed as in (A).

**Fig. 4.** Depletion of sites in genes preferentially co-expressed with the miRNA. (**A**) miR-133 sites in skeletal muscle and CD8+ T-cells. For each panel, genes were binned based on their expression in the indicated tissue compared to expression in the 60 other tissues, with bin 1 lowest and bin 61 highest. Top: difference between observed and expected number of messages with miR-133 sites at each expression rank. Bottom: modified KS test and estimate of significance, showing the running sum of the difference between the observed and expected distributions across expression ranks for messages with sites (red) compared to control cohorts (blue). (**B**) Summary map of KS tests for each

miRNA-tissue pair for 28 miRNAs; *P*-value key is shown above. Reported expression is from zebrafish *in situ* data (Wienholds et al., 2005), supplemented with notable mammalian data (Baskerville and Bartel, 2005; Sempere et al., 2004) (parentheses). **(C)** Tail of *P*-value distribution for all 73 miRNA families (left, fig. S7) and for a mock analysis using control sequences (right). *P*-values greater than $10^{-3}$, which are gray in (B), were only marginally less frequent for controls. **(D)** RNA-blot analysis of miR-7 in rat tissues, reprobed for miR-124 and U6 snRNA.

## Figure 1

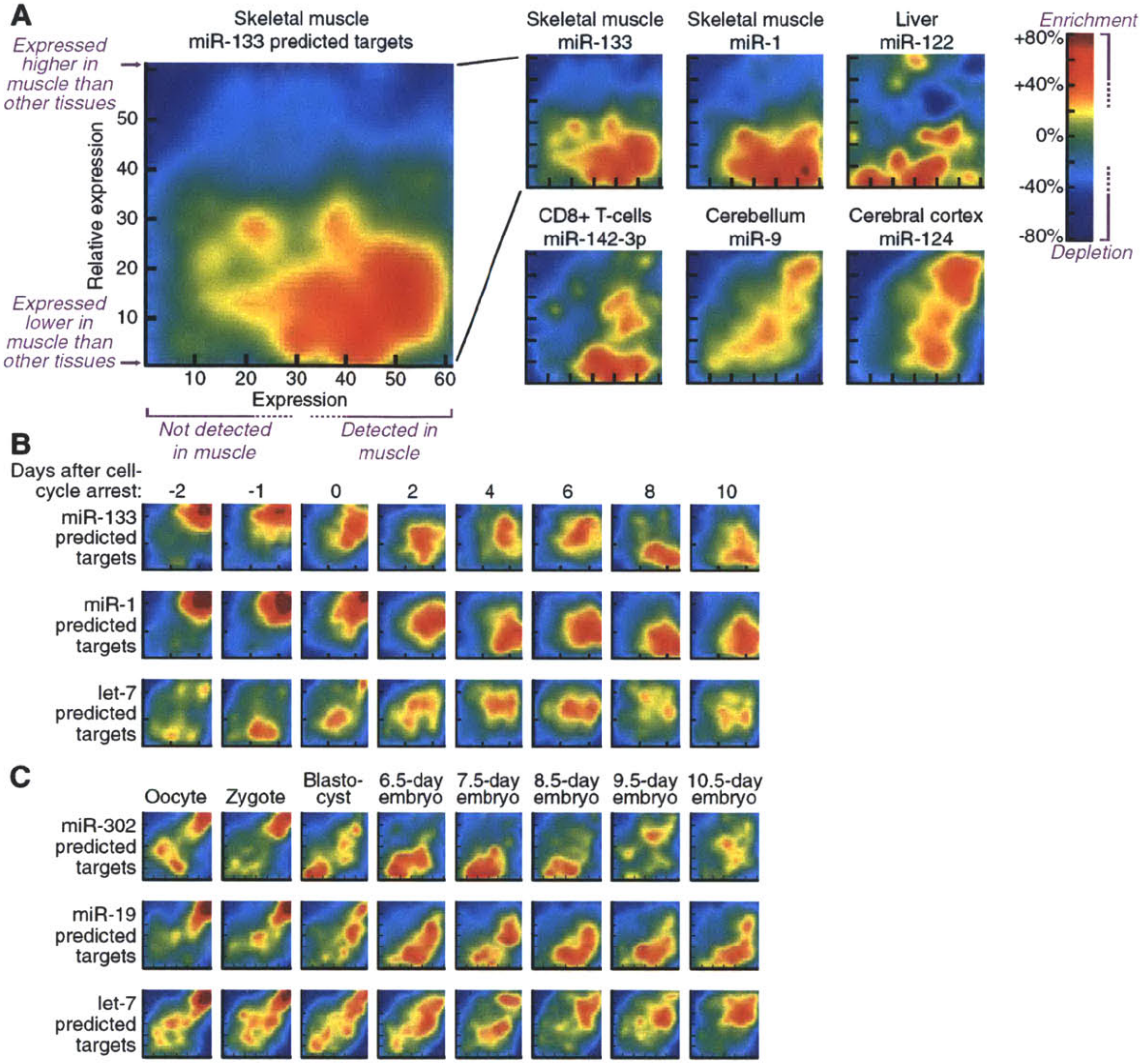

A
Skeletal muscle
miR-133 predicted targets
Expressed higher in muscle than other tissues
Expressed lower in muscle than other tissues
Relative expression
10
20
30
40
50
Expression
10
20
30
40
50
60
Not detected in muscle
Detected in muscle
Skeletal muscle miR-133
Skeletal muscle miR-1
Liver miR-122
CD8+ T-cells miR-142-3p
Cerebellum miR-9
Cerebral cortex miR-124
Enrichment
+80%
+40%
0%
-40%
-80%
Depletion
B
Days after cell-cycle arrest:
-2
-1
0
2
4
6
8
10
miR-133 predicted targets
miR-1 predicted targets
let-7 predicted targets
C
Oocyte
Zygote
Blasto-cyst
6.5-day embryo
7.5-day embryo
8.5-day embryo
9.5-day embryo
10.5-day embryo
miR-302 predicted targets
miR-19 predicted targets
let-7 predicted targets

**Figure 2**

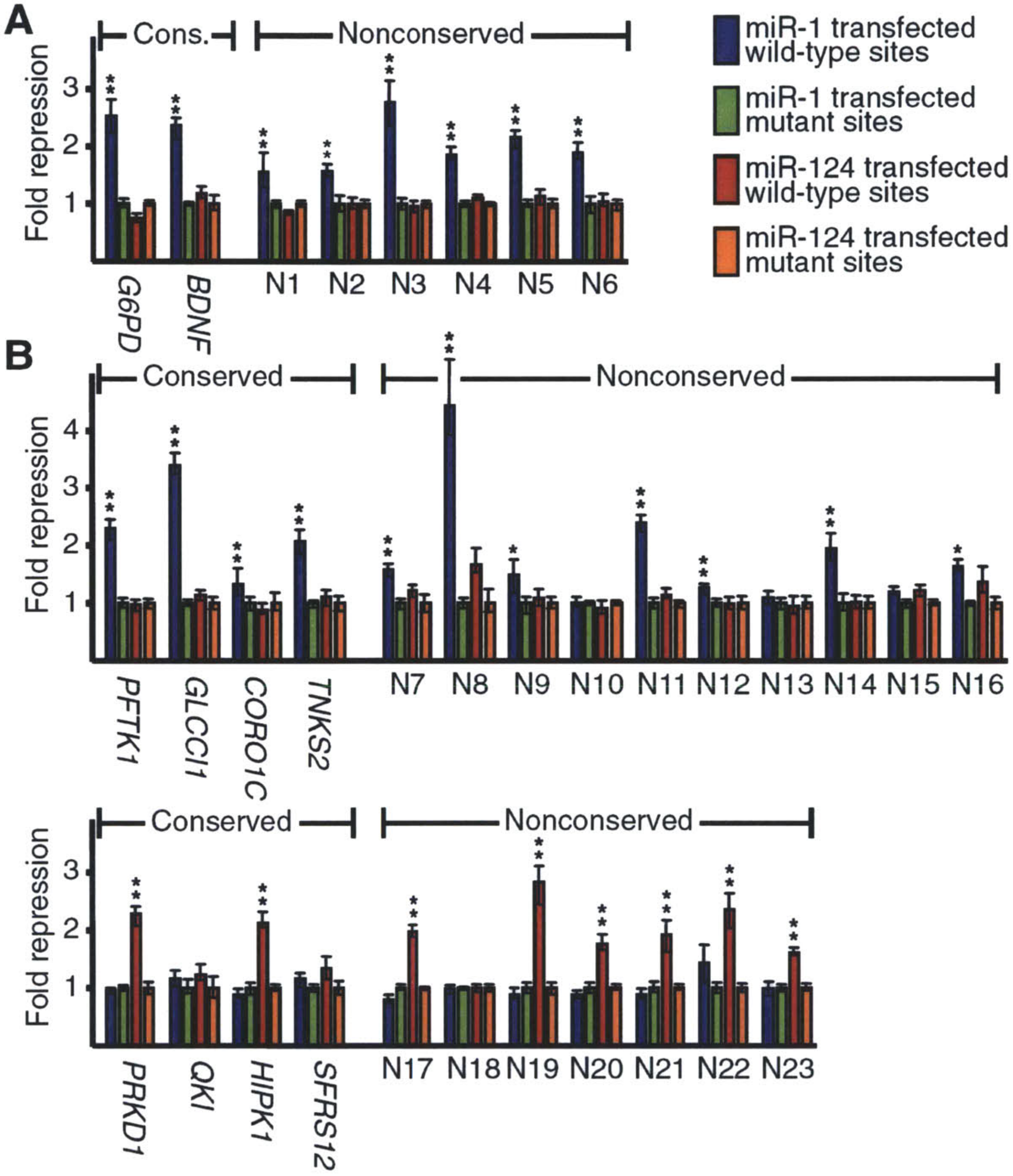
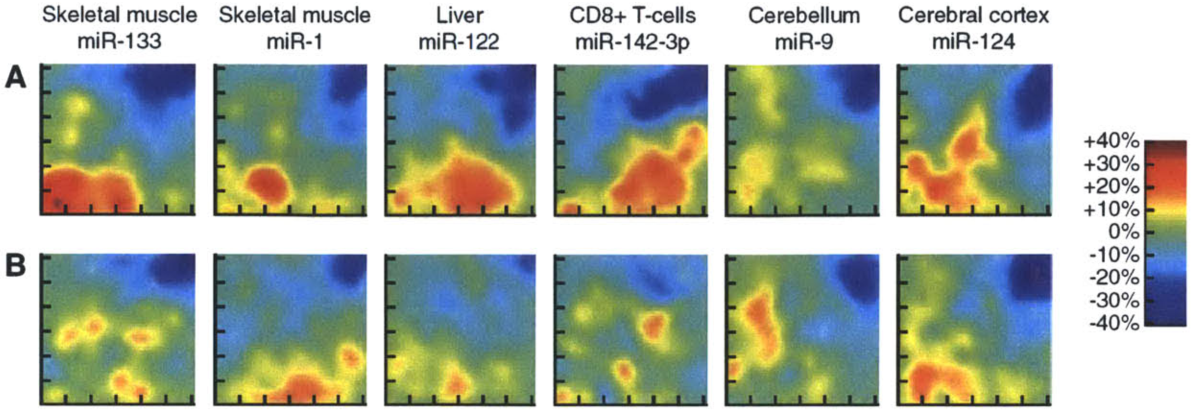
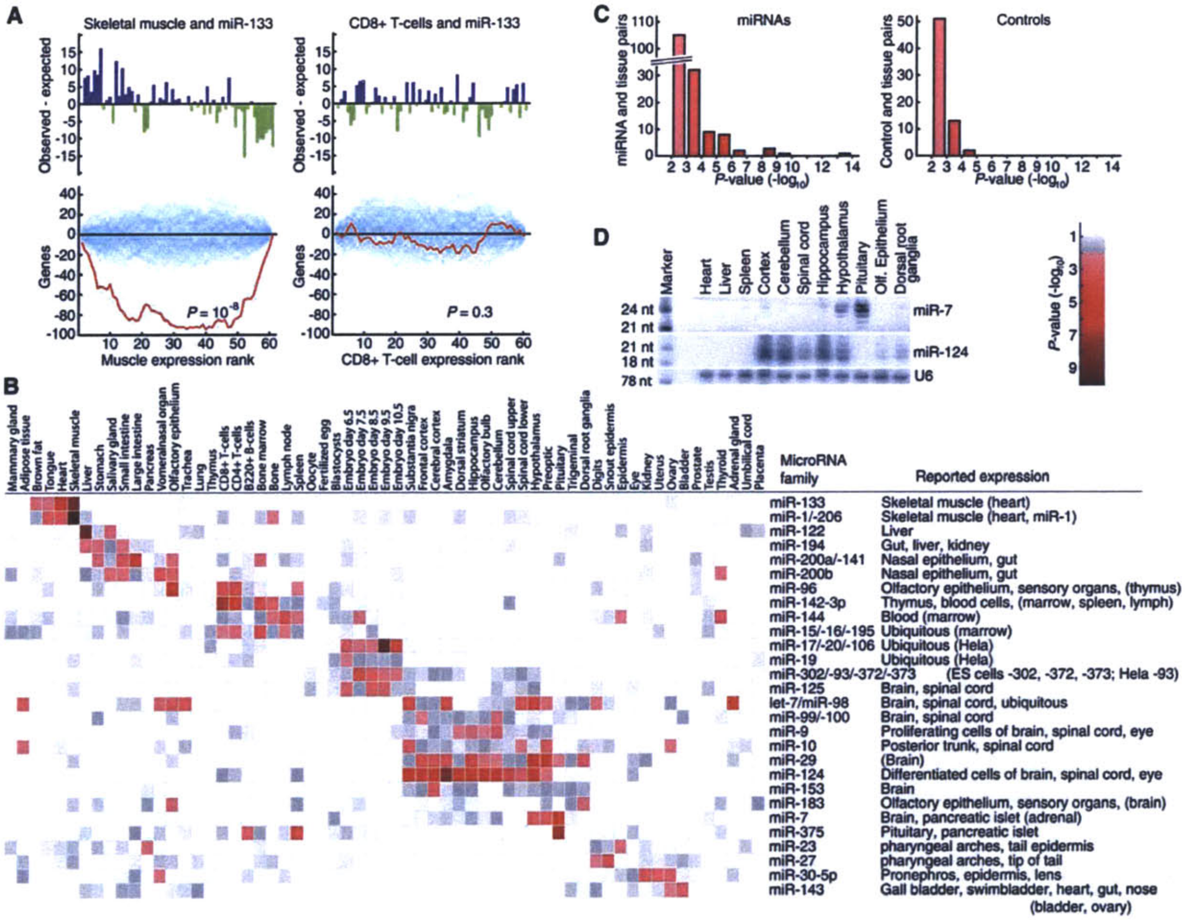

A
Cons.
Nonconserved
Fold repression
1
2
3
G6PD
BDNF
N1
N2
N3
N4
N5
N6
miR-1 transfected wild-type sites
miR-1 transfected mutant sites
miR-124 transfected wild-type sites
miR-124 transfected mutant sites
B
Conserved
Nonconserved
Fold repression
1
2
3
4
PFTK1
GLCCI1
CORO1C
TNKS2
N7
N8
N9
N10
N11
N12
N13
N14
N15
N16
Conserved
Nonconserved
Fold repression
1
2
3
PRKD1
QKI
HIPK1
SFRS12
N17
N18
N19
N20
N21
N22
N23

**Figure 3**


Skeletal muscle
miR-133
Skeletal muscle
miR-1
Liver
miR-122
CD8+ T-cells
miR-142-3p
Cerebellum
miR-9
Cerebral cortex
miR-124
A
B
+40%
+30%
+20%
+10%
0%
-10%
-20%
-30%
-40%

## Figure 4



| MicroRNA family | Reported expression |
|---|---|
| miR-133 | Skeletal muscle (heart) |
| miR-1/-206 | Skeletal muscle (heart, miR-1) |
| miR-122 | Liver |
| miR-194 | Gut, liver, kidney |
| miR-200a/-141 | Nasal epithelium, gut |
| miR-200b | Nasal epithelium, gut |
| miR-96 | Olfactory epithelium, sensory organs, (thymus) |
| miR-142-3p | Thymus, blood cells, (marrow, spleen, lymph) |
| miR-144 | Blood (marrow) |
| miR-15/-16/-195 | Ubiquitous (marrow) |
| miR-17/-20/-106 | Ubiquitous (Hela) |
| miR-19 | Ubiquitous (Hela) |
| miR-302/-93/-372/-373 | (ES cells -302, -372, -373; Hela -93) |
| miR-125 | Brain, spinal cord |
| let-7/miR-98 | Brain, spinal cord, ubiquitous |
| miR-99/-100 | Brain, spinal cord |
| miR-9 | Proliferating cells of brain, spinal cord, eye |
| miR-10 | Posterior trunk, spinal cord |
| miR-29 | (Brain) |
| miR-124 | Differentiated cells of brain, spinal cord, eye |
| miR-153 | Brain |
| miR-183 | Olfactory epithelium, sensory organs, (brain) |
| miR-7 | Brain, pancreatic islet (adrenal) |
| miR-375 | Pituitary, pancreatic islet |
| miR-23 | pharyngeal arches, tail epidermis |
| miR-27 | pharyngeal arches, tip of tail |
| miR-30-5p | Pronephros, epidermis, lens |
| miR-143 | Gall bladder, swimbladder, heart, gut, nose (bladder, ovary) |

## Supplementary Data

## Materials and Methods

### Expression and Sequence Data

Mouse expression data were obtained from the Novartis Research Foundation (Su et al., 2004). The data comprised two replicates of 61 different tissue samples hybridized to the Affymetrix GNF1M mouse chip, and were normalized using Affymetrix Microarray Suite 5.0 (MAS5) software. Human and mouse annotated 3'UTR sequence data were obtained from Refseq. Orthologous human, mouse, rat, and dog 3'UTR data were derived from the UCSC genome browser (Karolchik et al., 2003) multiZ multiple genome alignments (Blanchette et al., 2004).

We selected 8,551 genes for our analysis, using the following criteria: (1) each gene had unambiguous mouse and human reciprocal orthologs, (2) each ortholog had both mouse and human UTRs annotated by Refseq, and (3) the gene was represented on the GNF1M chip. Affymetrix probe IDs ending with *_x_at*, which do not uniquely complement the target sequence and hence are likely to cross-hybridize, were excluded. When multiple probes mapped to a single gene entry, we used the arithmetic mean of the probe intensities.

For each gene in each tissue sample, we assigned an absolute expression rank and a relative expression rank, thereby creating two 8,551 X 61 gene–by-tissue sample matrices. To calculate the absolute expression rank for each gene, the geometric mean of the two replicates was sorted with respect to the geometric means of the other 8550 genes in that sample; values were placed in 61 equal-sized bins, of increasing absolute expression. To assign the relative expression rank for each gene in a sample, we ranked the gene in that sample with respect to expression in all other samples, again using the geometric mean of the two replicate values. We then sorted the 8,551 genes for each tissue into 61 equal-sized bins according to their relative rank. Bin 1 contained genes which had the lowest expression compared with expression of those genes in other tissues, while bin 61 contained genes which were highest in that tissue compared with their expression in other tissues.

### MicroRNA Families and Site Identification

Our analysis included 73 miRNA families [listed in fig. S7; miRNA sequences can be found at miRBase (Griffiths-Jones, 2004)], which were defined as sets of miRNAs with identical nucleotides at positions 2-8. All families were required to have at least one member conserved across human, mouse, rat, dog, and zebrafish. Except for sites in Fig. 2A, potential miRNA regulatory sites were found by searching the 3'UTR sequences for 7-nt matches, which included a 6-nt match to the miRNA seed (nucleotides 2-7) and either a seventh Watson-Crick match to miRNA nucleotide 8 or an adenosine opposite nucleotide 1 (Lewis et al., 2005). Conserved sites were identified using UCSC MultiZ alignments of orthologous 3'UTR regions from human, mouse, rat and dog (Blanchette et al., 2004). Genes were considered to contain conserved sites if their 3'UTR contained an aligned 7-nt match in all four genomes. Mouse and human annotated UTRs often differed slightly in length; for such cases, the alignments were based on the species with the shorter annotated UTR. Mouse and human genes with nonconserved sites (Fig 2B and 2C, respectively) were identified using the Refseq annotated 3'UTR set for that genome. To simplify the description of our methods, we sometimes refer to genes with conserved sites as "conserved targets" and those with nonconserved sites as "nonconserved targets," recognizing that these are not necessarily biological targets.

For the purposes of evaluating our method (e.g., Fig. 4C), we generated a cohort of control sequences in which the miRNA seed regions were shuffled while maintaining dinucleotide and mononucleotide frequencies approximating actual miRNA seed regions, which were further required to differ in sequence from any known conserved mammalian miRNA.

**Background Calculation**

We calculated the likelihood of a 7-nt miRNA seed matching with a given 3'UTR as a function of trinucleotide frequencies and length. The second order Markov probability of a 7-nucleotide seed matching an arbitrary 7mer $x_1x_2x_3x_4x_5x_6x_7$ in the UTR was calculated as:

$$\begin{aligned}&P(\text{matching\_a\_given\_7mer})\\&\sim P(x_1x_2x_3) \text{ X } P(x_2x_3x_4|x_2x_3) \text{ X } P(x_3x_4x_5|x_3x_4) \ldots \text{ X } P(x_5x_6x_7|x_5x_6)\\&= P(x_1x_2x_3) \text{ X } P(x_2x_3x_4) / P(x_2x_3) \text{ X } P(x_3x_4x_5) / P(x_3x_4) \ldots \text{ X } P(x_5x_6x_7) / P(x_5x_6)\end{aligned}$$

where $P(x_m..x_n)$ is the probability of matching nucleotides *m* through *n* of the 7mer.

The probability of each trinucleotide and dinucleotide was estimated for each UTR as the observed frequency of the trinucleotide or dinucleotide in that UTR, without using pseudocounts. This approach assumes that the overall dinucleotide and trinucleotide frequencies in a UTR approximate those for smaller windows within the UTR, and that the sequences of the miRNA seeds and the UTR are sufficiently nondegenerate such that the probabilities of di- and trinucleotides and 7mers are approximately independent. From this calculation of the probability of matching any given 7mer, we calculated the probability of the 7-nt sequence occurring in the UTR, based on length.

P(UTR_contains_7mer)
= 1 - P(UTR_does_not_contain_7mer)
= 1 - (1-P(7nt_match))^(number_of_7mers)
= 1 - (1-P(7nt_match))^(UTR_length - 6)

To validate that this approach was accurately estimating the likelihood of a given miRNA seed to match to a given UTR, we calculated the number of targets (either conserved or nonconserved) for each miRNA in the annotated mouse and human UTR sets. The total expected number of UTRs targeted by all miRNA families was 95% of the actual observed number of targets in mouse and 93% of the actual observed number of targets in human, and the Pearson correlation between number of targets for each individual miRNA in the expected and observed sets was 0.93 for both mouse and human. When we repeated this analysis for our control set of shuffled miRNA seeds, the total expected number of UTRs targeted by all miRNA families was 99% and 100% of the actual observed number of targets in human and mouse, respectively, and the correlations were 0.92 for human and 0.93 for mouse. Across all human UTRs, 15% of miRNA target sites are conserved between human, mouse, rat, and dog, and of these conserved sites, about half that number are conserved above background expectation (Lewis et al., 2005). The discrepancy between total numbers of observed and expected targets for real miRNA families in one genome was primarily accounted for by the ~7.5% of targets with sites conserved above background.

The choice of a trinucleotide model over mononucleotide and dinucleotide models was based on the empirical performance of each model in estimating the observed number of targets matching each miRNA in one genome. To illustrate the higher-order effects captured by the trinucleotide model over lower-order models, scatterplots displaying the estimated and observed numbers of targets for each miRNA family in mouse are shown in fig. S5C.

To calculate the expected probability of a given miRNA matching a conserved site in a UTR, we calculated the probability of the miRNA matching a single 7mer in an analogous manner (the nucleotide frequencies were obtained from the human or mouse annotated UTR used as the basis for the alignment), but used the actual number of conserved 7mers in the aligned UTR, instead of the UTR length:

```
P(UTR_contains_conserved_7mer)
= 1 - P(UTR_does_not_contain_conserved_7mer)
= 1 - (1-P(7nt_match))^(number_of_aligned_conserved_7mers)
```

Dinucleotide and trinucleotide probabilities were estimated from the dinucleotide and trinucleotide frequencies for the entire UTR sequence, as opposed to using just conserved regions, because human UTRs are, on average, ~1000 nucleotides in length, while the number of conserved 7mers averages only ~70, too short for accurately estimating trinucleotide frequencies.

In contrast to the results in one genome, where the expected and observed numbers of targets correlated well, there was a marked enrichment for conserved targets of real miRNAs over expected. When considering all 73 vertebrate families, the ratio of the observed number of conserved targets to the expected number was 2.0 : 1, whereas the ratio for the controls was 0.93 : 1. The signal-to-noise values for each individual miRNA obtained via this approach approximated those obtained from TargetscanS (Lewis et al., 2005).

**Gene Density Maps (Figs. 1, 3, S2, S5, and S6)**

A stepwise construction of a gene density map from Fig. 1A is illustrated in fig. S2A.

For each miRNA, the Observed targets gene-density map reflected the actual distribution of genes with sites matching the miRNA, based on the target finding approaches discussed above, whereas the Expected targets (or Background) map reflected the expected distribution of genes with sites matching the miRNA, based on properties of their UTRs (length, conservation, and trinucleotide composition) that influence the likelihood of a match occurring by chance.

To calculate the Observed gene density map, the position of each of the genes targeted by the miRNA was assigned in accordance with its absolute expression ($x$-axis) and its relative expression ($y$-axis) as illustrated (fig. S2A). Maps were smoothed using a squared Euclidean kernel function, with each target gene contributing a density of $1/(r^2 + k)$ to each cell on the heatmap, where $r^2$ was the squared Euclidean distance between the coordinates of the cell and the coordinates of the target gene, and $k$

was a constant smoothing factor. The relatively large values for the smoothing constant *k* (61 X 0.4 for the mouse atlas, 24 X 0.2 for the C2C12 time course) were necessary for effective visualization, because a typical miRNA has more than an order of magnitude fewer conserved targets than the total number of cells on the density map. All density maps were normalized to a mean density of 1.0 (green), and the colors represent positive (red) and negative (blue) deviations from mean density.

The Expected gene density maps were calculated analogously, using all genes and scaling the contribution of each gene by its fractional expected probability of matching the miRNA by chance. These expected probabilities were calculated as described above, and take into account the influence of UTR length (for the analysis of nonconserved targets), number of conserved 7mers (for the analysis of conserved targets), and trinucleotide composition (for the analysis of both nonconserved and conserved targets.)

The differences between the density maps (displayed in Figs. 1 and 3 in the main text) were calculated by subtraction of the density of the Expected (or Background) map from the density of the Observed map at each of the 61 X 61 cells in the two density maps (fig. S2 and S5). Local differences in density (both in the original and subtracted density maps) indicate differences relative to the mean density. The subtracted density maps are not sensitive to the order in which the smoothing function and the density subtraction were applied, because the density in each cell is the result of summing the contributions of each gene, applied using the kernel function.

For the analysis of conserved targets (Fig. 1 and fig. S2), instead of normalizing both the Observed and Background gene density maps to have the same mean density, the Background maps were scaled to a reduced intensity, based on the signal-to-noise of the miRNA, with the intent of subtracting out the density contributed by spurious targets conserved due to chance. Signal-to-noise was calculated as the number of observed conserved targets divided by the number of expected conserved targets (described in the previous section on calculating background). The signal-to-noise values for each of the six miRNAs used in the analysis were as follows: miR-133 (3.0), miR-1 (2.8), miR-122 (0.8), miR-142 (2.1) miR-9 (4.4), miR-124 (4.2). Because the signal-to-noise was below 1 for miR-122, the Background map for miR-122 was normalized to have the same mean density as the Observed map, as if signal-to-noise was 1.0. The other five subtracted gene density maps shown in Figure 1A reflected the distribution of predicted targets above estimated noise.

For the analyses of nonconserved targets (Figs. 3, S5, S6), both the Observed and Expected gene density maps were normalized to the same mean density. Thus, the subtracted maps reflected the relative differences in distribution between observed and expected nonconserved targets.

We caution that the gene density maps are strictly a visualization tool, and we employ them for the purpose of displaying general trends. The construction of the gene density map in figure 2A shows, for instance, that although the conserved targets of miR-133 tend to cluster in the lower right corner, miR-133 targets can be present in any region of the gene density map. In particular, all quantitative results and tests of statistical significance are derived directly from the absolute and relative bin indices of genes, without using the gene density map.

**Modified Kolmogorov-Smirnoff Test**

To characterize a signal for a particular miRNA in a particular tissue, we first assigned each gene to one of 61 equal-sized bins based on its relative expression in that tissue. This was the same procedure used to construct the gene density maps, except that genes were only binned along the relative axis. Genes with low relative expression were placed in low-numbered bins, while genes with high relative expression were placed in high-numbered bins. Each bin contained the same number of genes from the entire gene set, while the actual number of targeted genes in each bin varied.

Because different tissues preferentially express genes with different UTR lengths and trinucleotide compositions, and both of these variables affect the likelihood of matching to a particular 7-nt sequence, it was necessary to correct for these effects by estimating the expected number of target genes in each bin. For each gene in the entire gene set, we estimated the probability of targeting by a miRNA as a function of the gene's UTR length and trinucleotide composition (see above). Summing the probabilities in each bin gave the expected number of targets in each bin. This expected distribution was then normalized, setting the total number of expected targets summed across all bins equal to the total number of observed targets. Because the numbers of observed and expected targets were usually approximately equal, normalization amounted to multiplying by a number near 1.0 in most cases.

We used a modified Kolmogorov-Smirnoff test to compare expected and observed distributions (Mootha et al., 2003). Figure 3A (top panels) shows the results of subtracting the number of expected targets in each bin from the number of observed targets in each bin. The one-sided discrete Kolmogorv-Smirnoff test statistic was calculated by taking the running tally of the difference in each bin, across the entire distribution (Fig. 3A, red line in bottom panel) and using the largest cumulative negative difference as the KS test statistic. The negative displacement from zero on the *y*-axis in Figure 3A indicated the number of genes that were overrepresented on the left side of the distribution. To move the KS statistic back to zero, a corresponding number of targeted genes would have to be shifted from the left side of the distribution to the right side of the distribution.

To assess the significance of KS test statistic values, we counted the number of genes targeted, and selected an equal number of genes from the entire gene set as a control cohort. Genes targeted by the miRNA were allowed to be selected for the control set. The likelihood of a gene being selected for the control set was proportional to its probability of matching the miRNA by chance, as a function of its UTR length and trinucleotide composition. For each miRNA, we generated 1,000 control cohorts and obtained discrete KS test statistics for these controls (for ease of visualization only 100 control cohorts are shown in Figure 4A.) We used the KS test statistic values of the 1,000 control cohorts to build an empirical background distribution, from which a p-value for the KS test statistic value of the actual miRNA was determined. For KS test statistics beyond the 98th percentile of the empirical distribution (i.e., those more significant than 980 of the 1000 controls), there were insufficient numbers of controls to accurately estimate *P* values. For each of these, we fit the asymptotic KS test statistic tail probability $Q = e^{-2nx\text{^}2}$ (van der Waerden), where *x* is the value of the KS test statistic, to the tail of our empirical distribution to derive an approximate *P* value for the miRNA-tissue pair. The parameter *n* was estimated by finding the value of *n* at which the 98th percentile of the theoretical tail probability matched that of the empirical distribution.

Using the set of 73 miRNA families conserved in mammals and zebrafish, and the 61 tissues of the mouse atlas, we performed 73 x 61 KS tests, i.e., each test involved comparing a particular miRNA-tissue pair to 1,000 control cohorts (fig. S7; a partial, clustered version is displayed in Fig. 3B). In general, we observed significant KS-test values for miRNA-tissue pairs when the miRNA is expressed specifically and strongly in that tissue. We caution, however, that a significant signal for a particular miRNA-tissue pair does not always indicate that the miRNA is expressed in that particular tissue. For instance, miR-10 may be expressed in spinal cord neurons, which express many genes in common with cerebral cortex neurons, perhaps explaining the signal seen for a wide range of brain regions. Selection against acquiring a target site in these genes would be apparent beyond that specific cell type to cell types sharing similar expression profiles. Conversely, lack of a signal in a particular tissue may be due to heterogeneity in the tissue, because the genes most highly expressed in the tissue may come from cells in which the miRNA is absent. Heterogenous tissues such as lung, prostate, etc., gave weak signals for all miRNAs.

### Estimation of Selective Avoidance

The 73 miRNA families in our analysis had an average of 1,087 human and 1,050 mouse targets among our set of 8,551 genes, with an average overlap of 367 genes that were shared among the two species. To calculate the number of genes avoiding target sites during evolution, we considered the genes that were above median in terms of both relative and absolute expression in the tissue of miRNA expression (genes falling in the upper right quadrant of the gene density map, Fig. 3B.) We calculated both the observed and expected number of targets for this set of genes, in each case considering only the subset of genes from the total set which were not targeted by that miRNA in mouse. Although the sites were in human and not mouse UTRs, it was the mouse (not human) expression data that was used to determine which genes fell into the upper right quadrant, effectively preventing direct mRNA-destabilizing effects from contributing to the signal. The difference between the expected and observed numbers of targets represented our estimate for the number of genes affected by selective avoidance. The probability for a gene to contain a site matching the miRNA was calculated as described in the previous section (Background Estimation), and was dependent on UTR length and trinucleotide composition. Because the expected and observed numbers of genes for the total gene set differed slightly, we scaled the expected numbers of genes so that the total number of expected targets would equal the total number of observed targets when considering all 8,551 genes in the gene set. The signal for selective avoidance would therefore be constituted by the relative depletion of genes with sites within the upper right quadrant of the gene density map in Figure 3B (i.e. those genes which were expressed both strongly and specifically in this tissue.)

For example, to estimate the number of genes avoiding miR-133 sites in skeletal muscle, we considered the 2,661 genes of the mouse expression atlas that were above median in both relative and absolute expression. Of these genes, 207 were targets in mouse and thus were excluded from the analysis; such genes could potentially show confounding effects due to direct miRNA-mediated mRNA degradation. Of the remaining 2,454 genes, we observed that 156 had miR-133 target sites in their orthologous human UTRs, whereas we would have expected 188.4 genes to be targeted, based on the expected probabilities calculated for their UTR length and trinucleotide composition. Dividing these two numbers, this estimates that in the upper right quadrant, ~17.2% of genes are avoiding miR-133 target sites. Given that there are 2,454 genes in the quadrant, we conclude that ~420 genes expressed in muscle are under selection to avoid 7-nt miR-133 target sites.

### Estimation of the Extent of Target Depletion

To estimate the maximal extent of target depletion, we started with the mouse-only nonconserved analysis (Fig. 3A), and focused on the subset of genes that were most highly and specifically expressed in tissues with our six miRNAs (defined as genes that were in the top ten percent of genes in both relative expression rank and absolute expression rank.) We calculated the expected number of nonconserved targets among this subset, and compared this figure to the actual number of nonconserved targets in the subset, normalizing the total number of expected targets to reflect the total number of observed targets when summed across all genes. For miR-133 there were 403 genes that were both highly and specifically expressed in muscle, and of these, we observed 10 targets while expecting 24.3, giving us a depletion of 59%. The other microRNAs had depletion percentages as follows: miR-1, 43%; miR-122, 57%; miR-142, 54%; miR-9, 42%, and miR-124, 31%.

We chose only a small subset of genes that were highly expressed in both absolute and relative terms in order to account for the possibility of mRNA degradation effects. For messages in which miRNA targeting might cause mRNA degradation, both the absolute and relative ranks of the gene would be reduced. The result would be an effect in which targets were shifted incrementally to the lower left corner of the gene density map, and therefore looking in the middle of the map would be misleading, because target depletion would be partially obscured by genes of formerly higher rank cascading down. Thus, to quantify the maximal extent of target depletion, we looked at the genes that were most highly and specifically expressed, because there was no possibility of other genes falling into that region during such a cascade. The numbers we derived represent lower limits for the percentage of 7-mer sites responsive to highly expressed miRNAs because some targets may be translationally repressed with little or no changes in mRNA levels and insufficient time or selective pressure for site loss.

### Combining Mouse and Human UTR Information

Because of the noise in performing sequence analysis in one genome, we incorporated both human and mouse sequence information in our analysis for Figure 4, counting a gene to be a miRNA target if it contained a 7-nt target site in its UTR in either species. The expected probability of a gene being targeted by a miRNA was calculated as above, but trinucleotide frequencies were tallied only after the

UTRs were filtered so that long runs of conserved sequence (stretches of 7 or more nucleotides conserved in four genomes) were counted only once.

The total expected number of UTRs targeted by all miRNA families was 98% of the actual observed number of targets in the mouse + human analysis. This compared to 93% in mouse and 95% in human. The Pearson correlation between the number of targets for each individual miRNA in the expected and observed sets was 0.93 for the human set, 0.93 for the mouse set, and 0.95 for the combined mouse and human set.

When performing the human-only and mouse-only nonconserved analyses in Figures 3, S5, and S6, we excluded genes from the analysis that were targeted in the other species. In cases where genes had highly conserved UTRs, this meant that if the gene lacked a site in one species, it had a substantially reduced likelihood of having that site in the other species. To account for this, we did not tally nucleotide counts from long runs of conserved sequence (again, defined as stretches of 7 or more conserved nucleotides) for purposes of determining UTR length and nucleotide frequencies in both the human-only and mouse-only nonconserved analyses.

**C2C12 Myotube Differentiation Timecourse**

The expression data used in the analysis of C2C12 murine myoblast cell line differentiation (Rao et al., data not shown) consisted of 24 individual microarray experiments hybridized to the Affymetrix U74Av2 chip, reflecting eight time points assayed in triplicate. The first three time points (days -2, -1, 0) reflect gene expression prior to the onset of differentiation, and the latter five time points (days 2, 4, 6, 8, 10) reflect gene expression during the course of differentiation. The U74Cv2 chip data was not incorporated into our analysis, because it was missing the three experiments at the day 8 time point. The data were normalized using Affymetrix Microarray Suite 5.0 (MAS5) software, and we selected 4,965 genes for our analysis, using the same criteria as we followed in choosing genes for the main analysis.

Because of the smaller number of samples, we treated each of the 24 individual microarray experiments as its own separate sample for the gene density map analysis shown in Figure 1B (i.e, we did not merge the triplicate experiments). The gene density maps for miR-1, miR-133, and let-7 conserved targets were otherwise constructed in the same manner as in the main analysis, producing smaller maps of size 24 X 24 instead of 61 X 61. For each time point, the gene density maps from the three triplicates were averaged to produce the composite maps shown in Figure 1B. miR-1 and miR-133 are two muscle-specific miRNAs that accumulate beginning at day 0 and increase over the course of C2C12 differentiation (Rao et al., data not shown). The non-muscle-specific microRNA let-7 is shown alongside for comparison.

The mean change in the expression levels of the miRNA targets over the course of differentiation was calculated (fig. S3) as the geometric mean of the targets before differentiation (days -2, -1, 0) divided by the geometric mean of the targets after differentiation (days 8, 10); only genes that were expressed above median both before and after differentiation were included in the calculation. For each miRNA family with at least 100 conserved targets among the 4,965 genes, we calculated the mean change in expression levels of their targets, and found that miR-1 and miR-133 targets decreased by the greatest magnitude, with miR-1 targets decreasing an average of 23%, and miR-133 targets decreasing an average of 16% (fig. S3D.) In comparison, the typical decrease in expression for conserved targets of other miRNAs was ~5%, which we attribute to a propensity for differentiated myotubes to express genes with shorter, less well-conserved UTRs compared to undifferentiated myoblasts.

To evaluate the significance of the decrease in mean expression observed for miR-1 and miR-133 conserved targets, we repeated our analysis with 10,000 control cohorts for each miRNA, in a manner analogous to the modified Kolmogorov-Smirnoff test. We counted the number of genes targeted and selected an equal number of genes for each control cohort. Genes were randomly selected to populate the cohort based on their probability of having a conserved site to the miRNA by chance (see Background Estimation.) Only genes that were expressed above median both before and after differentiation were included in the analysis. For each miRNA, we derived an empirical background distribution describing the mean expression change due to chance, and used it to estimate a P-value for the observed decrease in the mean expression of the miRNA's targets. The P-value for miR-1 was < 0.0001, indicating that the decrease in expression of miR-1 targets was more significant than all 10000 control cohorts, while the P-value for miR-133 was 0.0100. The conserved targets of the other miRNAs were not significantly downregulated; the next most significant were the targets of miR-24, with P-value of 0.2524.

We also extended our analysis to nonconserved targets, calculating the mean decrease in expression in the same manner as for the conserved targets, and the using control cohorts to evaluate significance. Nonconserved targets of miR-1 and miR-133 decreased an average of 7% and 8%, respectively, neither of which were significant, due to the general tendency of genes with longer

UTRs to be expressed at lower levels in myotubes compared to myoblasts. Nonconserved targets of the other miRNAs also were not significantly downregulated.

While the decrease in expression of the conserved targets of miR-1 and miR-133 is highly significant, the significance comes from the consistency with which each target gene is downregulated, as opposed to large changes of two-fold or more. Hence, lists of genes with the largest foldchanges in expression have little overlap with the conserved targets of miR-1 and miR-133.

### Secondary Structures Flanking Conserved and Nonconserved Sites

79 mammalian miRNA families (Lewis et al., 2003) were searched against a database of multiz alignments of 3'UTR sequences constructed by identifying the annotated 3'UTR regions for 22383 RefSeq mRNAs (Pruitt et al., 2005) mapped by the UCSC genome browser [(Karolchik et al., 2003); genome.cse.ucsc.edu]. 7- and 8-nt sites conserved in human/mouse/rat/dog/chicken containing Watson-Crick pairing to bases 2-7 of the miRNA supplemented by either or both a Watson-Crick match to base 8 or an adenosine across from position 1 of the miRNA were identified. Those 7- and 8-nt sites found in human 3'UTR sequence but not observed to be conserved in the 5-vertebrate alignments were collected and included as the nonconserved set. In addition, a control set consisting of 4 cohorts corresponding to each of the 79 miRNA sequences were searched against the alignments and conserved and nonconserved sets were collected. Sets of control sequences were constructed for the 79 miRNA families so as to preserve properties affecting the likelihood of finding a match and score in a single genome using the TargetScan algorithm (Lewis et al., 2003). Notably, these control sequences preserved the predicted free energies associated with pairing to the miRNA seed region.

Zhao et al. (Zhao et al., 2005) report that the predicted secondary structures of sequences immediately flanking authentic miRNA binding sites have significantly less predicted stability than do average 3'UTR fragments. To explore this claim that authentic targets might be associated with less stable predicted secondary structures in mRNAs, we evaluated the predicted folding energies of sequences surrounding sites found to be conserved in 3'UTR alignments of 5 vertebrates and the remaining sites found in human. To enable evaluation of predicted folding energies of sequences flanking the 7- or 8-nt matches, only those sites located >70 nt downstream of the 3'UTR start in human and those sequences located >70 nt upstream of the 3'UTR terminus in human were included in the analysis. These 70-nt fragments were folded using the RNAfold routine from the Vienna RNA package (Hofacker, 1994) and the average folding free energy of the upstream and downstream fragments were calculated for both the 5-vertebrate conserved set and the nonconserved set. Results obtained when examining regions flanking conserved sites corresponding to real miRNAs were indistinguishable from those for the conserved sites corresponding to control cohort sequences.

These sets of sites also were evaluated for accessibility using a method resembling one used to predict miRNA target sites in *Drosophila* 3'UTRs (Robins et al., 2005). A 100-nt region surrounding each site was folded using RNAfold from the Vienna RNA package (Hofacker, 1994) and the predicted local structure at the 7- or 8-nt sites was searched for sequences of three consecutive unpaired bases. The set of conserved sites in 5 vertebrates was enriched for open structure relative to the set of nonconserved sites. However, the results obtained when identifying sites for sets of control cohort sequences that preserve the binding free energy of the seed region (in addition to properties affecting the likelihood of finding a target site in human 3'UTRs) were highly similar to those found for the real miRNAs. In summary, our analyses indicate that non-occlusive secondary structure, as measured by previously reported algorithms (Robins et al., 2005; Zhao et al., 2005), does not influence miRNA-directed targeting in mammals.

### Selection of predicted targets and nonconserved cohorts for reporter assays

Nonconserved TargetScan-like targets were selected randomly from human 3'UTR sequences that had human TargetScan scores within the range of those of miR-1 predicted targets with experimental support (Lewis et al., 2003), but were not TargetScan predictions because they did not score above the cutoffs in mouse or rat. Four of the six were not scored in mouse or rat because they lacked seed matches in the orthologous mouse or rat UTR. The other two (N1 and N3) were scored in both mouse and rat but had scores below the cutoffs.

TargetScanS predictions were randomly selected from a list of predicted human targets (Lewis et al., 2005) that had exactly two sites (7- or 8-nt matches to the seed region) in the 3'UTR, which were within ~1 kb of each other. Nonconserved TargetScanS-like targets were selected to resemble the TargetScanS predictions in human, in that they had two sites in the human 3'UTR, which fell within orthologous aligned regions of the human, mouse, rat and dog genomes. However, the aligned segments corresponding to both human sites were diverged to include mismatched nucleotides in mouse, rat or dog sequences. In all cases, both sites were concurrently disrupted in at least one of the mouse, rat or dog orthologous sequences. In two of 17 cases (N18 and N19), there was an additional non-aligned site within the 3'UTR of the ortholog lacking the two aligned sites. However, these two cases do not bias our interpretation, because only one of the two mediated repression.

To simplify the experimental analysis, we choose to examine UTR fragments with two matches to the same miRNA family, even though most UTRs with a conserved match to a miRNA family do not have a second conserved match to the same miRNA family (Lewis et al., 2005). This simplification was justified based on the observation that UTRs with a conserved match to one miRNA usually have a second conserved match to a second miRNA (Lewis et al., 2005), and in cells in which both miRNAs are expressed the repression presumably would be equivalent to that observed with two matches to the same miRNA (Doench and Sharp, 2004). The same would be true for nonconserved matches, in that more than 90% of UTRs have nonconserved matches to multiple miRNAs. One concern was that UTRs with two nonconserved sites to the same miRNA might be more likely to be important species-specific targets. To address this possibility, we investigated whether such UTRs occur more or less frequently than would be expected by chance, comparing UTRs containing multiple matches to miRNAs with the number containing multiple matches to control sequences of similar overall abundance that do not match miRNAs. For both control sequences and miRNA matches, we plotted the number of UTRs with one match versus the number with more than one match, and found that the double-site UTRs occurred in the same relative proportion for the control sequences as for miR-1 and miR-124. Therefore, UTRs with multiple miRNA matches occur as frequently as expected by chance, indicating that there is no strong selection for or against multiple occurrences over selection acting on single occurrences.

*Renilla* reporter plasmids were constructed by insertion of PCR-amplified 3'UTR fragments into a p-RL-SV40-derived vector (Promega). Mutant derivatives were constructed by QuikChange site-directed mutagenesis (Stratagene). Insert sequences were confirmed by sequencing and are provided in Table S1.

**Transfection and luciferase assays**

HeLa cells were transfected using Lipofectamine 2000 (Invitrogen) in 24-well plates (~0.5 x $10^5$ cells / well) with 25 ng firefly luciferase control reporter (pIS0, (Lewis et al., 2003)), 10 ng *Renilla* luciferase reporter, 1.25 μg pUC19 and an appropriate amount of miRNA duplex. miR-1 duplex comprised oligonucleotides: 5'-UGGAAUGUAAAGAAGUAUGUA-3' and 5'-CAUACUUCUUUACAUUCAAUA-3'; miR-124 duplex comprised: 5'-UAAGGCACGCGGUGAAUGCCA-3' and 5'-GCAUUCACCGCGUGCCUUAAU-3'. Firefly and *Renilla* luciferase activities were measured 24 hours after transfection with the Dual-luciferase assay

(Promega). *Renilla* activity was normalized to firefly activity to control for transfection efficiency. Values plotted in Fig. 1 are geometric means of replicate values.

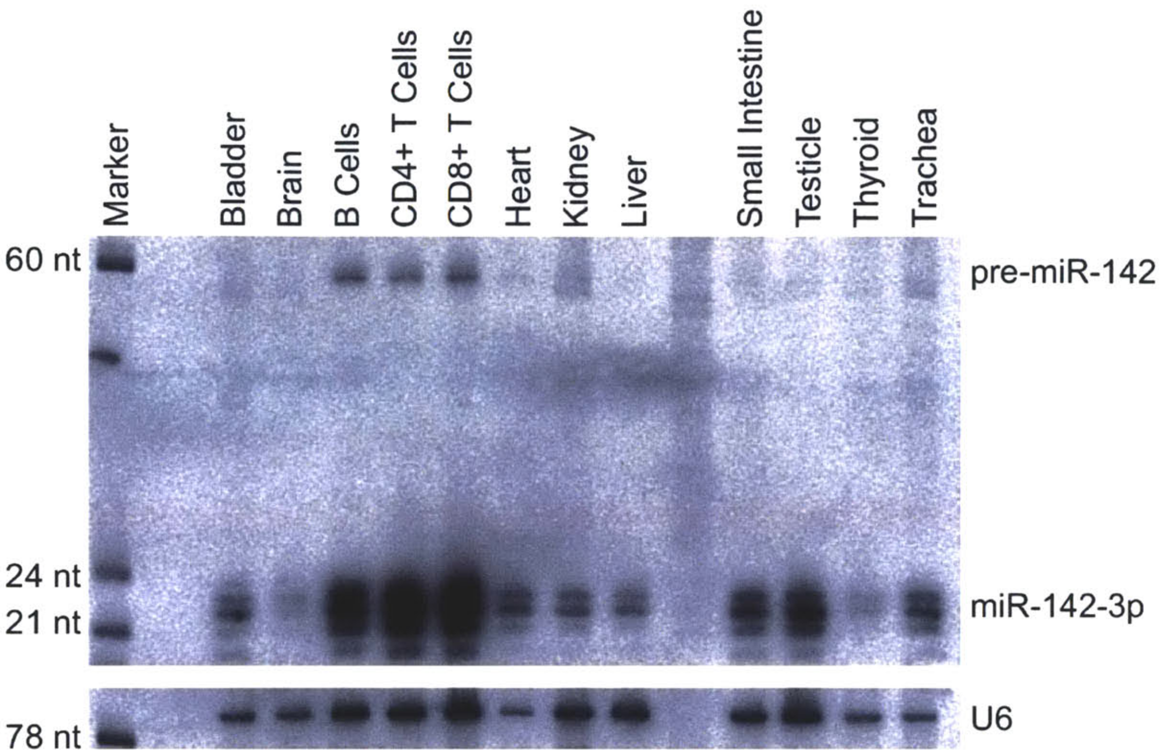
### Northern blotting

For Fig. 4D, 9 μg total RNA was loaded per lane. All tissues were from rat. RNA from cortex, cerebellum, spinal cord, hippocampus, hypothalamus, pituitary, olfactory epithelium and dorsal root ganglia was purchased from Analytical Biological Services (Wilmington, DE); RNA from heart, liver and spleen was purchased from Ambion (Austin, TX). Northern blotting was performed as described previously [(Lau et al., 2001); http://web.wi.mit.edu/bartel/pub/protocols/], with the following DNA oligo probes:

miR-7a: CAACAAAATCACTAGTCTTCCA
miR-7b: AACAAAATCACAAGTCTTCCA
miR-124: TGGCATTCACCGCGTGCCTTAA
U6 snRNA: TTGCGTGTCATCCTTGCGCAGG

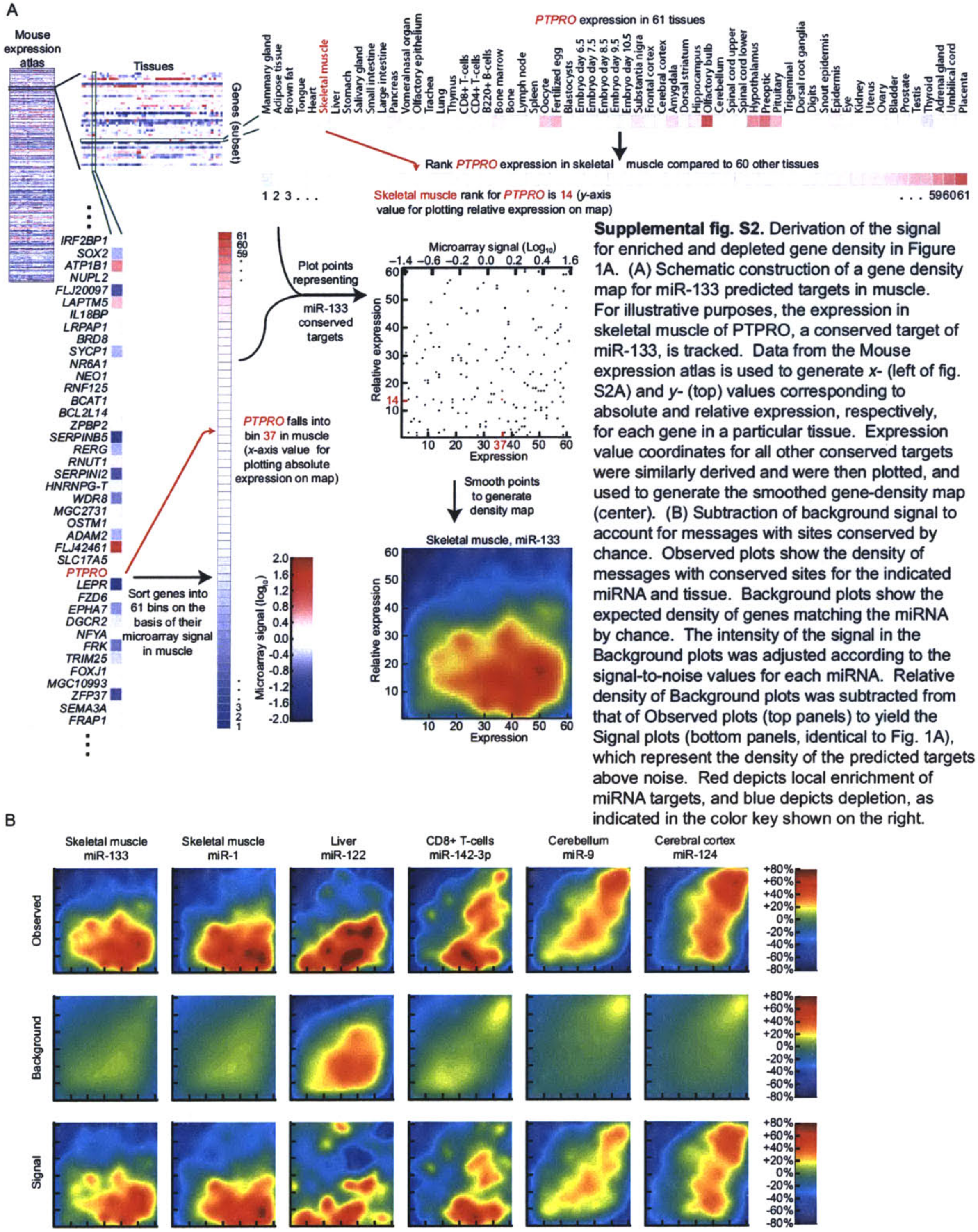
miR-7 was probed with a combination of miR-7a and miR-7b probes. For comparison, the blot was stripped and reprobed for miR-124, then U6 snRNA.

**Supplemental Figure S1**



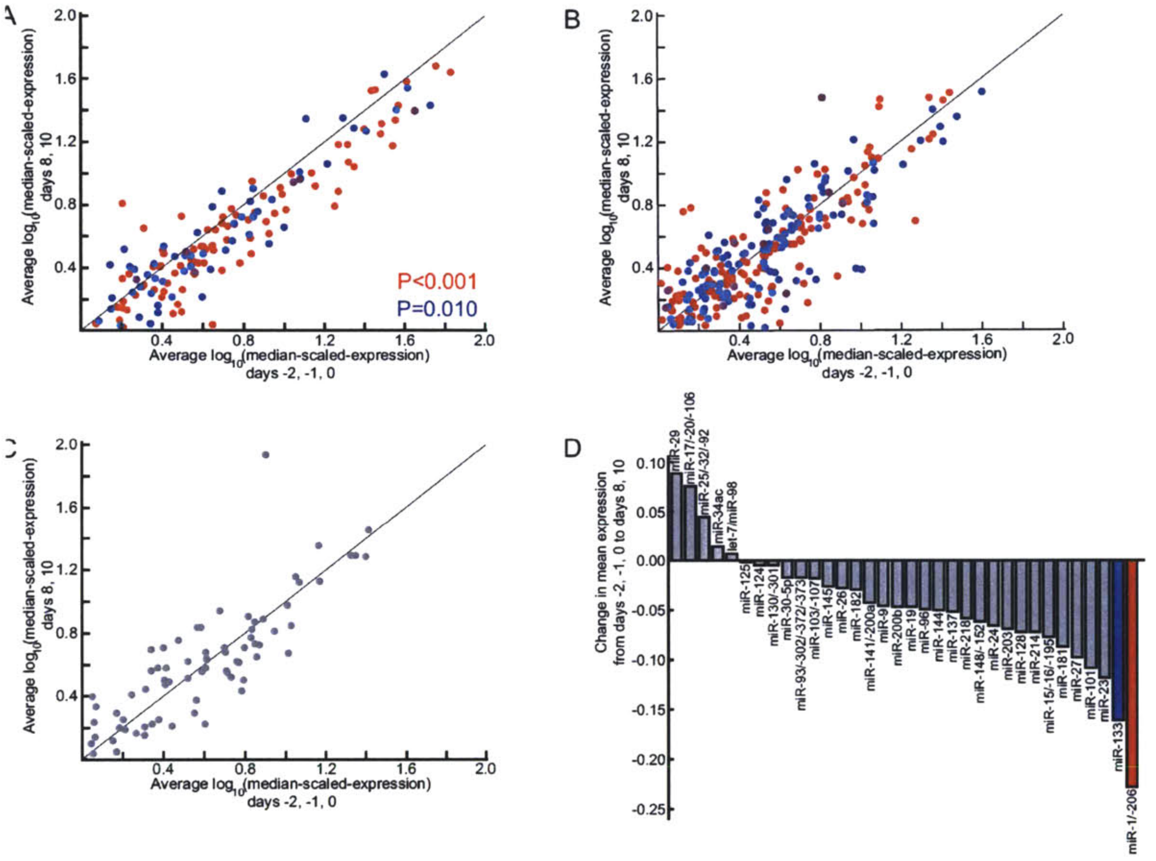
**Supplemental fig. S1.** miR-142-3p accumulates to high levels in human B-cells, CD4+ T cells and CD8+ T cells from peripheral blood. Detection of pre-miR-142 suggests active expression of miR-142-3p in purified, differentiated peripheral blood cells. Small RNA blotting was performed using 15 μg human total RNA per lane, the human miR-142-3p miRCURY LNA probe (Exiqon, Vedbaek, Denmark), and a U6 DNA oligo probe (TTGCGTGTCATCCTTGCGCAGG), as previously described [(S14); http://web.wi.mit.edu/bartel/pub/protocols/]. RNA from peripheral human B-cells, CD4+ T-cells, and CD8+ T-cells was purchased from AllCells (Berkeley, CA). All other samples were purchased from Ambion (Austin, TX).

## Supplemental Figure S2



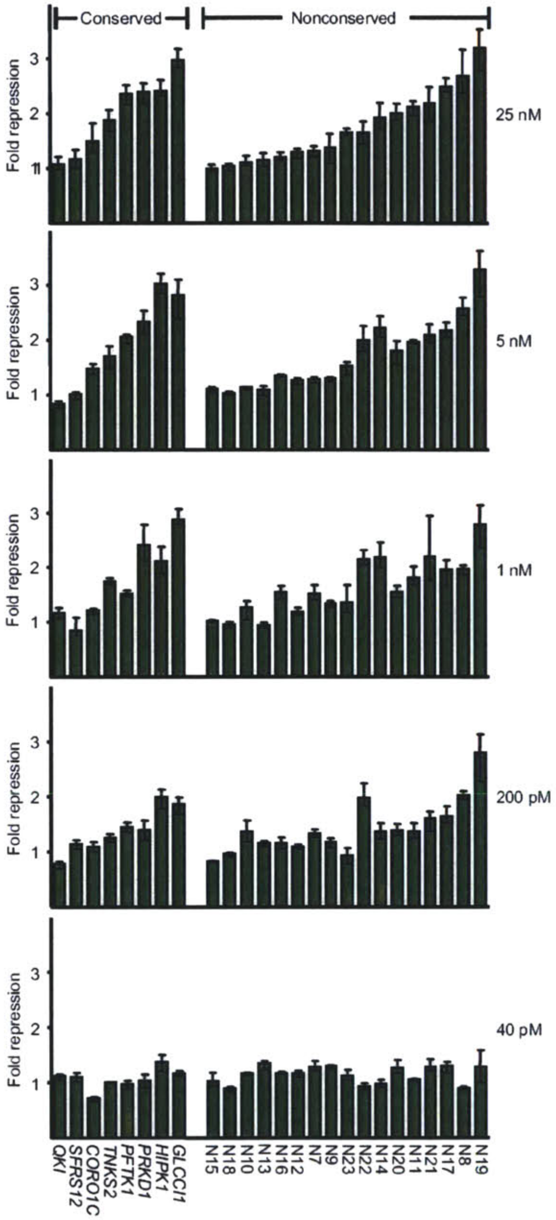
**Supplemental fig. S2.** Derivation of the signal for enriched and depleted gene density in Figure 1A. (A) Schematic construction of a gene density map for miR-133 predicted targets in muscle. For illustrative purposes, the expression in skeletal muscle of PTPRO, a conserved target of miR-133, is tracked. Data from the Mouse expression atlas is used to generate *x*- (left of fig. S2A) and *y*- (top) values corresponding to absolute and relative expression, respectively, for each gene in a particular tissue. Expression value coordinates for all other conserved targets were similarly derived and were then plotted, and used to generate the smoothed gene-density map (center). (B) Subtraction of background signal to account for messages with sites conserved by chance. Observed plots show the density of messages with conserved sites for the indicated miRNA and tissue. Background plots show the expected density of genes matching the miRNA by chance. The intensity of the signal in the Background plots was adjusted according to the signal-to-noise values for each miRNA. Relative density of Background plots was subtracted from that of Observed plots (top panels) to yield the Signal plots (bottom panels, identical to Fig. 1A), which represent the density of the predicted targets above noise. Red depicts local enrichment of miRNA targets, and blue depicts depletion, as indicated in the color key shown on the right.

## Supplemental Figure S3



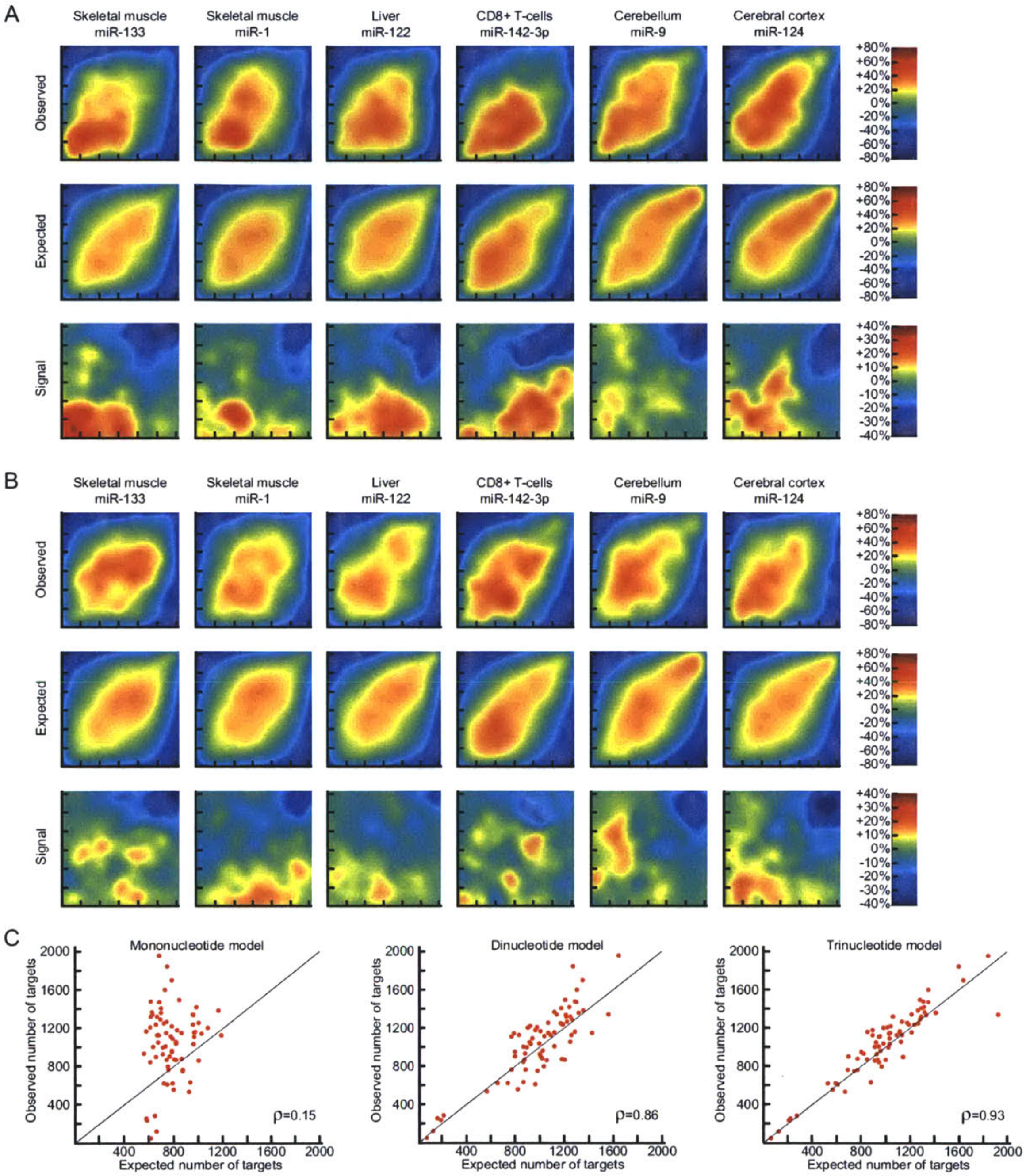
**Supplemental fig. S3.** Decreased expression of miR-1 and miR-133 predicted targets during myoblast differentiation. The scatter plots display genes that are expressed above median levels on the microarray both before and after differentiation. This set was enriched for conserved targets of miR-1 and miR-133 (60% were expressed both before and after) versus nonconserved targets of miR-1 and miR-133 (43% were expressed both before and after). (A) Conserved targets of miR-1 (red), miR-133 (blue) or both (purple), expressed above median both before and after differentiation. *x* axis: expression before differentiation. *y* axis: expression after differentiation. Conserved targets of miR-1 and miR-133 are consistently shifted from the midline, reflecting a decrease in their expression levels over the course of differentiation. The decrease in the expression of the 95 miR-1 targets averaged 23%, with a range of 7% to 36% at 25th and 75th percentiles. The decrease in the expression of the 62 miR-133 targets averaged 16%, with a range of -4% to 31% at 25th and 75th percentiles. Mean changes in the expression of miRNA targets were measured by the change in the geometric mean of their expression values. Because a substantial fraction of the myoblasts do not differentiate into myotubes, the decrease is expected to be greater in the subset of cells that differentiate. Bootstrap P values were derived using 10,000 control cohorts. (B) The 377 nonconserved targets of miR-1 and miR-133. Expression of these nonconserved targets decreased subtly (7% and 8%, respectively, *P* values not significant) (C) Conserved let-7 targets did not markedly change. (D) Changes in mean expression of the conserved targets of 34 miRNA families. Values indicate the average decrease over the course of differentiation. The 34 miRNA families shown are the subset of the 73 miRNA families in our analysis that had at least 100 conserved targets among the 4965 genes in the C2C12 analysis.

# Supplemental Figure S4



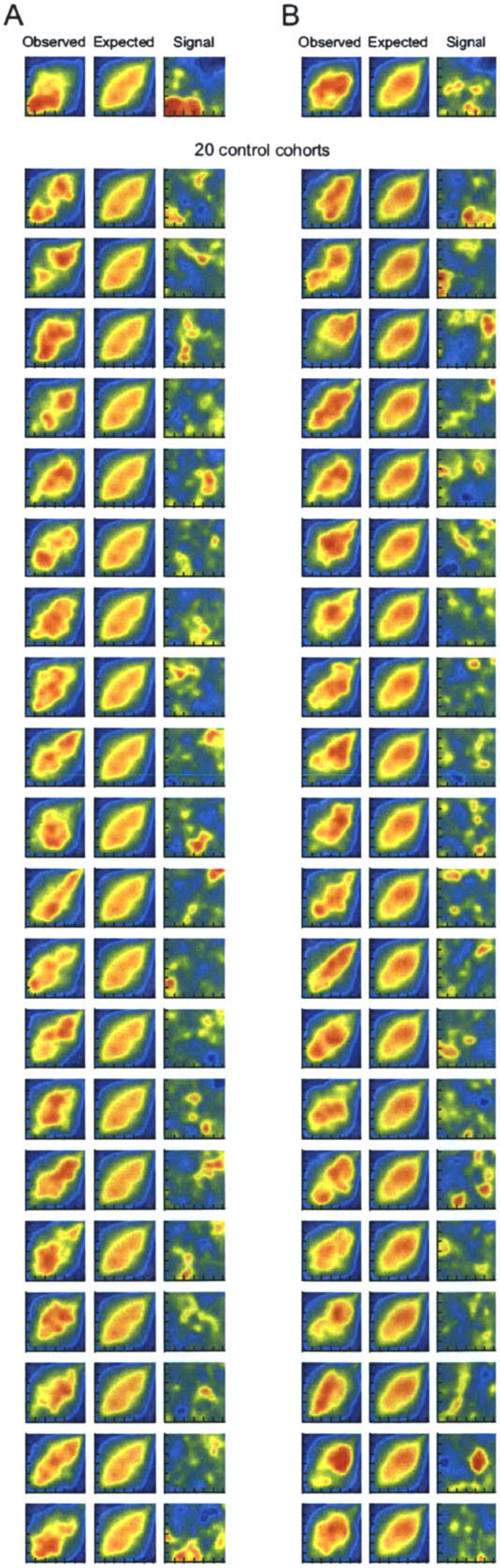
**Supplemental fig. S4.** MicroRNA titration of miRNA-mediated repression of reporter genes containing 3'UTR segments of target genes. All targets from Fig. 2B were re-assayed as in Fig. 2, except that miRNA concentrations were titrated to indicated levels. Luciferase values were processed as in Fig. 2, then values with cognate miRNA were normalized to those with non-cognate miRNA. Three replicates were performed (error bars represent largest and smallest values), except for the top panel, which was derived from 12 replicate values shown in Fig. 2B (error bars represent the third highest and lowest values).

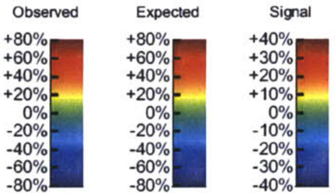
## Supplemental Figure S5



**Supplemental fig. S5.** Derivation of the signal for enriched and depleted gene density in Figure 3. (A) Derivation for Fig. 3A. (B) Derivation for Fig. 3B. Observed plots show the expression of messages with nonconserved sites for the indicated miRNA and tissue, generated as illustrated in figure S2A. Expected plots show the background distribution of genes likely to have a nonconserved match to the miRNA by chance. Relative density of each Expected plot was subtracted from that of the corresponding Observed plot (top panels) to yield the Signal plot (bottom panels, identical to Fig. 3), which represents the density of the nonconserved targets relative to expectation. Red depicts local enrichment of miRNA targets, and blue depicts depletion, as indicated in the color key shown on the right. (C) Comparison of the performance of mono-, di- and tri-nucleotide models for predicting the observed numbers of targets in mouse. Individual points correspond to the 73 miRNA families used for the analysis. ρ, Pearson correlation between observed and expected number of targets.

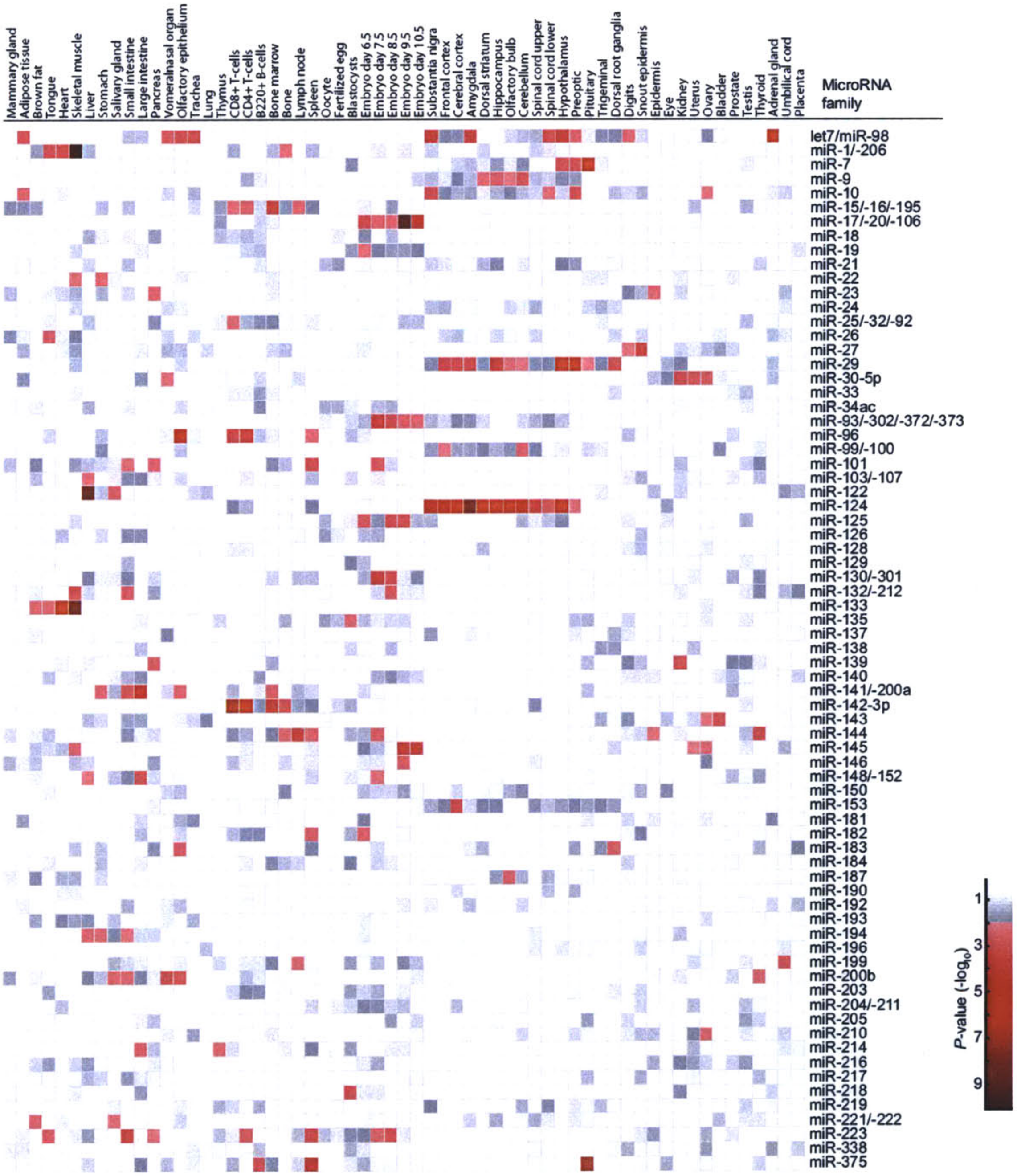
## Supplemental Figure S6





**Supplemental fig. S6.** Analysis of messages with nonconserved miR-133 sites in muscle (fig. S5), repeated with 20 control cohorts to illustrate the variability of the signal expected by chance. Each control cohort was populated with the same number of genes as for the Observed plot in figure S5. Genes for the control cohorts were selected according to their probability of containing the miRNA target site, on the basis of UTR length and nucleotide composition.

## Supplemental Figure S7



**Supplemental fig. S7.** Complete map showing KS-test *P* values for each tissue-miRNA pair (61 tissues, 73 miRNAs) like that shown in Figure 4B. Darker areas denote increasingly significant values (Table S2), as indicated in color key.

## Supporting Online Text

### Specificity determents beyond 7- or 8-nt matches to miRNA seed regions

We present experimental and computational evidence indicating that determinants outside the matches to the seed region play only small roles in specifying targeting. Additional specificity determinants proposed previously include pairing to the 3' portion of the miRNA, mRNA secondary structure that could occlude miRNA pairing, and sites for RNA-binding proteins that might occlude or recruit the miRNA-programmed silencing complex. Although early computational analyses, including ours, assumed that pairing to the 3' portion of mammalian miRNAs would usually provide added specificity (John et al., 2004; Lewis et al., 2003), more recent studies suggest that 3' pairing is primarily important for sites that do not have perfect seed matches, which appear to be much less prevalent than those that have perfect matches (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005). In contrast to the emerging consensus regarding the role of 3' pairing, two recent studies have suggested that non-occlusive mRNA secondary structure is a major determinant of targeting (Robins et al., 2005; Zhao et al., 2005). To investigate this possibility, we implemented these two algorithms with the idea that if they successfully identified non-occlusive secondary structure important for targeting that the conserved sites would preferentially fall into these open areas when compared to the nonconserved sites (Materials and Methods). A slight preference for conserved sights was detected, but with further investigation this preference did not appear to be associated with miRNA targeting, in that it persisted when we replaced the miRNAs with non-miRNA control sequences. We conclude that, mRNA secondary structure, as measured by previously reported algorithms (Robins et al., 2005; Zhao et al., 2005), does not perceptively influence miRNA-directed targeting in mammals, when considering results summed over all targets.

Although our data support the idea that overall specificity determinants outside the 7- or 8-nt match to the seed region overall play relatively minor roles in specifying targeting by highly expressed miRNAs, for any individual site under study they might still play important roles. Although messages preferentially co-expressed with miRNAs are generally ~50% depleted in miRNA matches, they are not totally depleted, and only 13/17 fragments with nonconserved TargetScanS-like sites mediated repression. We speculate that some of the non-responsive sites are in regions of the UTR made less accessible by secondary structure or protein binding. Furthermore, determinants outside the 7- or 8-nt match might explain some of the variability in repression observed in the reporter assay, ranging from <1.3-fold to 4-fold.

### Number of genes avoiding miRNA targeting

The finding that ~170 to ~440 genes of the mouse expression atlas appear to be selectively avoiding targeting to the six miRNA families of Figure 3 raised the question of how many genes of the atlas are avoiding targeting to all miRNAs. This question cannot be addressed by simply summing the antitargets of each miRNA family because the antitargets of one miRNA might be the same as those of another. For example, the antitargets of miR-133 would be expected to largely overlap with those of miR-1 because these two miRNAs are expressed in the same tissues. However, the antitargets for miRNAs expressed in different tissues cannot all overlap; when estimating the number of genes selectively avoiding targeting by a miRNA, we use relative rank, and a gene cannot be relatively high in all tissues. When considering that a single miRNA can have more than 400 antitargets and that there are numerous tissues and cell types that specifically express miRNAs, we estimate that the aggregate number of antitargets is on the order of thousands, not hundreds. Note that an antitarget of one miRNA family can be a conserved target of another family and a nonconserved target of a third family.

### Signatures for sites in coding sequence

To examine targeting within the coding sequence, we repeated the analysis of Figure 4B using coding sequences rather than 3'UTRs. Some miRNA families, including let-7, miR-9, miR-29, miR-122, miR-124, miR-142, miR-133, miR-125 gave robust, accurate signatures, but the P-values were generally less significant than for the 3'UTR analysis, and other miRNAs gave spurious signatures. We conclude that substantial targeting involving perfect seed matches occurs in coding sequence but that 3'UTRs are more hospitable for targeting. This result agrees with previous target-prediction analysis, which showed that 8-nt sites within coding sequences were under selective pressure to preserve miRNA pairing but that this signal for conserved targeting, although present in coding sequence, is highest in 3'UTRs (Lewis et al., 2005).

### Additional considerations and implications

Our analyses of the expression of messages with conserved and nonconserved sites reports propensities and trends. It is important to bear in mind that some messages do not follow the dominant trends. For example, the miR-133 analysis in skeletal muscle showed a propensity of messages with conserved sites to be expressed in muscle, but expressed at lower levels in muscle compared to other tissues of the atlas (Fig. 1A). As mentioned in the text, this trend, together with the trend during myoblast differentiation, is concordant with the ideas that miRNAs often 1) dampen the output of preexisting messages to facilitate a more rapid and robust transition to a new expression program, 2) optimize protein output without eliminating it entirely, and 3) destabilize many target messages to further define tissue-specific transcript profiles. However, messages with conserved sites populate all regions of the gene density map (fig. S2A), with some having no detectable expression in muscle or myoblasts (as modeled by C2C12 cells). Because it is doubtful that all, or even most, of these sites are conserved by chance, we conclude that a sizable minority of conserved targets represent exceptions to the principles enumerated above. These include messages that are destabilized to imperceptible levels with the help of miRNA-mediated repression—a class of targets that might be particularly abundant among targets of the miR-302 family (Fig. 1C). They also include failsafe targets, which are messages that are nearly completely repressed at the transcriptional level but require miRNAs to assure fidelity of this repression (Hornstein et al., 2005).

On the whole, analyzing the apparent imprint of miRNAs on mRNA expression and evolution revealed the spatial and temporal expression of miRNAs during mammalian development (Fig 4). This striking correspondence between the signatures and the expression patterns indicated that the signals we observed in Figures 1 and 3 have biological meaning and conversely that miRNA expression patterns have biological meaning, i.e., that miRNAs generally are active in the tissues where they are expressed and are not sequestered in an inactive form.

Our results have ramifications for interpreting results of reporter assays and TargetScanS—two of the main tools for studying and identifying mammalian miRNA targets. In our heterologous reporter assay, conserved sites mediate repression indistinguishable from that of nonconserved sites (Fig. 2 and fig. S4). Although not every message with conserved sites is an authentic target, and some that have nonconserved sites might be authentic species-specific targets, it is reasonable to propose that those with conserved sites are substantially enriched in biological targets compared to those with nonconserved sites. The observation that conserved and nonconserved sites mediate similar repression in the reporter system calls into question the utility of such a reporter system for distinguishing biological targets from messages with fortuitous pairing to the miRNA. Although nonconserved sites are less likely than conserved sites to be biological (in large part because they are less likely to be in mRNAs that are coexpressed with the miRNA), the abundance of nonconserved sites and our in vivo evidence for function of such sites (Fig. 3 and 4) suggest frequent nonconserved repression. Therefore, the TargetScanS predictions represent only a fraction (probably a minority) of targets repressed in the animal. A more complete list can be compiled by considering the messages coexpressed with the miRNA that have a conserved or nonconserved 7-nt TargetScanS-like site. Nonetheless, a focus on conserved predictions enriches for interactions that evolution has preserved and are therefore most likely to be consequential. Furthermore, the use of conservation as the criterion for distinguishing features of miRNA targeting from equally plausible fortuitous features has been informative for discovering general principles of miRNA targeting in the past [e.g., in defining Watson-Crick seed pairing (Lewis et al., 2003) and the A anchor (Lewis et al., 2005) as generally important for targeting] and will undoubtedly reveal additional insights in the future.

## Supporting References

1. A. I. Su *et al.*, *Proc. Natl. Acad. Sci. USA* **101**, 6062 (2004).
2. D. Karolchik *et al.*, *Nucleic Acids Res* **31**, 51 (2003).
3. M. Blanchette *et al.*, *Genome Res* **14**, 708 (2004).
4. S. Griffiths-Jones, *Nucleic Acids Res* **32**, D109 (2004).
5. B. P. Lewis, C. B. Burge, D. P. Bartel, *Cell* **120**, 15 (2005).
6. V. K. Mootha *et al.*, *Nat Genet* **34**, 267 (2003).
7. B. L. van der Waerden, *Mathmatical Statistics* (Springer-Verlag, Berlin, 1969), page 74.
8. P. K. Rao, M. Farkhondeh, S. Baskerville, H. F. Lodish, (data not shown).
9. B. P. Lewis, I. H. Shih, M. W. Jones-Rhoades, D. P. Bartel, C. B. Burge, *Cell* **115**, 787 (2003).
10. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res* **33**, D501 (2005).
11. Y. Zhao, E. Samal, D. Srivastava, *Nature* **436**, 214 (2005).
12. I. Hofacker, Fontana, W, Stadler, PF, Bonhoeffer, S, Tacker, M, Shuster P, *Monatshefte fur Chemie*, 167 (1994).
13. H. Robins, Y. Li, R. W. Padgett, *Proc. Natl. Acad. Sci. USA* **102**, 4006 (2005).
14. J. G. Doench, P. A. Sharp, *Genes Dev.* **18**, 504 (2004).
15. N. C. Lau, L. P. Lim, E. G. Weinstein, D. P. Bartel, *Science* **294**, 858 (2001).
16. B. John *et al.*, *PLoS Biol.* **2**, e363 (2004).
17. A. Krek *et al.*, *Nat. Genet.* **37**, 495 (2005).
18. J. Brennecke, A. Stark, R. B. Russell, S. M. Cohen, *PLoS Biol.* **3**, e85 (2005).
19. E. Hornstein *et al.*, *Nature*, in press (2005).

# Chapter II.

The following chapter includes the manuscript, figures, and supplementary data for the paper titled "MicroRNA targeting specificity in mammals: determinants beyond seed pairing", published in *Molecular Cell* in 2007. I performed all of the computational analyses in the paper. On the experimental side, co-author Andrew Grimson performed the luciferase assays with the assistance of Wendy Johnston. Our collaborators at Rosetta Inpharmatics, Philip Garrett-Engele and Lee Lim, performed microRNA transfection and array experiments. Andrew Grimson, David Bartel, and I each took turns writing and revising the text.

# MicroRNA Targeting Specificity in Mammals: Determinants Beyond Seed Pairing

**Andrew Grimson[1,2,5], Kyle Kai-How Farh[1,2,3,5], Wendy K. Johnston[1,2], Philip Garrett-Engele[4], Lee P. Lim[4]*, David P. Bartel[1,2]***

[1]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA

[3]Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[4]Rosetta Inpharmatics (wholly owned subsidiary of Merck and Co.), 401 Terry Avenue N, Seattle, WA 98109, USA

[5]These authors contributed equally to this work.

*Contact: lee_lim@merck.com (L.P.L.), dbartel@wi.mit.edu (D.P.B.)

## Summary

**Mammalian microRNAs (miRNAs) pair to 3'UTRs of mRNAs to direct their posttranscriptional repression. Important for target recognition are ~7-nt sites that match the seed region of the miRNA. However, these seed matches are not always sufficient for repression, indicating that other characteristics help specify targeting. By combining computational and experimental approaches, we uncovered five general features of site context that boost site efficacy: AU-rich nucleotide composition near the site, proximity to sites for co-expressed miRNAs (which leads to cooperative action), proximity to residues pairing to miRNA nucleotides 13-16, and positioning within the 3'UTR at least 15 nt from the stop codon and away from the center of long UTRs. A model combining these context determinants quantitatively predicts site performance both for exogenously added miRNAs and for endogenous miRNA-message interactions. Because it predicts site efficacy without recourse to evolutionary conservation, the model also identifies effective nonconserved sites and siRNA off-targets.**

# Introduction

MicroRNAs are ~22-nt endogenous RNAs deriving from transcripts that form characteristic hairpin precursor structures (Bartel, 2004). The first miRNAs were discovered in *C. elegans*, where they were found to suppress the protein output of mRNAs via incomplete base pairing with sites located in the 3'UTRs of the target messages (Lee et al., 1993; Reinhart et al., 2000; Wightman et al., 1993). These RNAs were later found to be part of a much larger family, with many members deeply conserved through the vertebrate and metazoan lineages (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001).

A central goal for understanding the functions of all these tiny non-coding RNAs has been to understand how they recognize their target messages. Conserved Watson-Crick pairing to the 5' region of the miRNA, which includes the miRNA seed, enables prediction of targets above the background of false-positive predictions, indicating the importance of this region for miRNA target recognition (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Lewis et al., 2003). More than a third of human genes appear to have been under selective pressure to maintain their pairing to miRNA seeds (Lewis et al., 2005), and many messages that either decrease upon miRNA ectopic expression or increase upon miRNA knock-down have matches to the miRNA seed (Giraldez et al., 2006; Krutzfeldt et al., 2005; Lim et al., 2005).

Messages downregulated after introducing a miRNA are most associated with four types of sites, which are in agreement with those anticipated from preferential conservation of sites in orthologous UTRs (Farh et al., in preparation). These include one 6mer, two 7mers, and one 8mer (Figure 1A). The 6mer is the perfect 6-nt match to the miRNA seed (miRNA nucleotides 2-7) (Lewis et al., 2005). The best 7mer site, referred to here as the 7mer-m8 site, contains the seed match augmented by a match to miRNA nucleotide 8 (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Lewis et al., 2003). Also effective is another 7mer, the 7mer-A1 site, which contains the seed match augmented by an A at target position 1 (Lewis et al., 2005). The 8mer site comprises the seed match flanked by both the match at position 8 and the A at position 1 (Lewis et al., 2005).

Reporter assays indicate that pairing to the seed is not only important for recognition but that in some cases 7-8mer sites appear sufficient (Brennecke et al., 2005; Doench and Sharp, 2004; Lai et al., 2005). However, seed sites do not always confer repression, and when repression occurs, the degree of repression is highly variable in different UTR contexts. In the animal, messages preferentially expressed in the same tissue as a highly expressed miRNA have 3'UTRs that are ~50% depleted in

7mer sites, presumably because these messages have important roles in that tissue, and during the course of evolution they have avoided acquiring sites to coexpressed miRNAs that would compromise their function (Farh et al., 2005). However, these UTRs are not completely devoid of 7mer sites, suggesting that even among miRNA sites in vivo, similar variability in site efficacy also exists, such that identical 7mer sites may or may not be consequential depending on the UTR context in which they arise. Thus, having a canonical 7-8mer site is clearly important but often not sufficient for detectable downregulation; unknown context determinants outside of seed sites must also be playing important roles for target recognition.

Without knowing these context determinants, experimentalists face a difficult challenge. Where should they begin when seeking to understand the molecular mechanisms of phenotypes arising from miRNA knockdown or ectopic expression studies? One approach is to focus only on the conserved sites, but so many messages are under selective pressure to maintain pairing to miRNAs—more than 200 messages, on average, for each highly conserved mammalian miRNA family—that this approach only serves to narrow the field, while potentially excluding some of the more rapidly evolving and species-specific interactions (Farh et al., 2005; Giraldez et al., 2006; Krek et al., 2005; Krutzfeldt et al., 2005; Lewis et al., 2005).

We set out to discover the context features that help specify miRNA targeting, with the idea that such insights into target recognition would increase the ability to predict functional sites, both those that are conserved and those that are not. We found five independent features that influenced targeting, each of which had both experimental and computational support. Combining these determinants, we constructed a model of miRNA regulation capable of quantitatively predicting the performance of miRNA sites based solely on sequence, and we confirmed experimentally the predictive power of this model for both exogenously added miRNAs and for endogenous miRNA-message interactions. Because our approach accurately distinguishes effective from non-effective sites without regard to site conservation, it also provides the basis for identifying siRNA off-targets.

## Results and Discussion

### Closely Spaced Sites Often Act Synergistically

We started with mRNA microarray data from 11 miRNA transfection experiments. A large majority (75%) of the downregulated messages detected on microarrays have canonical 7-8mer sites in their 3'UTRs (Farh et al., in preparation). However, only a minority of the messages possessing single sites in their 3'UTRs displayed detectable destabilization on the array (19%, 25%, and 43% for the 7mer-A1, 7mer-m8, and 8mer, respectively, Figure 1B), which suggested that analyzing the differences between those messages that were detectably downregulated and those that were not could help identify context features important for target recognition.

As anticipated from previous studies (Brennecke et al., 2005; Doench and Sharp, 2004; Lai et al., 2005), multiple sites were associated with greater mRNA destabilization in our transfection experiments (Figures 1B and 1C). To determine whether cooperative action of dual sites might enhance this greater destabilization, we systematically examined the repression observed for genes with two sites to the same miRNA, comparing it to that simulated based on genes with one site. When considering all genes, the repression observed for those with two-sites was almost exactly that expected if the two sites had contributed independently to repression; that is, the repression for a gene with two sites matched the result calculated by multiplying the repression from two single sites (Figure 1D). This multiplicative effect, a hallmark of independent and noncooperative action, was observed previously in a heterologous reporter assay designed to model miRNA repression (Doench and Sharp, 2004).

We observed one notable exception to the overall tendency of apparently independent action: when the two sites were close together, the repression tended to be greater than that expected from the independent contributions of two single sites (Figure 1E). Examining the conservation of sites in orthologous UTRs of human, mouse, rat and dog, we found that when plotting the number of co-conserved sites for authentic miRNAs, after subtracting the average number of co-conserved sites for control cohorts, the greatest enrichment was for closely spaced dual sites (Figure 1F). Although most of the co-conserved sites were at longer intervals (because longer intervals greatly outnumber shorter intervals), the observed enrichment compared to any other specific intervals indicated a detectable biological preference for short intervals. Such cases with short inter-site spacing appear to be the more effective ones and thus would be easiest to identify genetically. In agreement with this idea, the *C. elegans lin-4*:*lin-14* and *lsy-6*:*cog-1* interactions and the *Drosophila* miR-4:*Bearded* interactions all involve dual conserved 8mer sites with short inter-site spacing (Johnston and Hobert, 2003; Lai et al., 2005; Lee et al., 1993; Wightman et al., 1993); see targetscan.org for minor refinements to *C. elegans* site annotations.

We investigated this phenomenon further using reporter assays, examining UTR fragments containing two sites and asking whether the observed repression from these two sites was greater than that expected from the sites individually (calculating expected repression as the product of the repression values when measured for each of the two sites in isolation). The two UTR fragments for which observed repression deviated most from that expected corresponded to the two shortest inter-site distances examined (Figure S1). For example, two proximal miR-124 sites individually mediated only subtle down-regulation, whereas both sites together mediated more robust down-regulation, which was significantly cooperative (Figure 1G, left-most). To test whether the proximity of the two sites was required for the observed cooperativity, four different sets of constructs that increased inter-site spacing were generated. Increasing spacing from 19 to 56 intervening nt, using either of two different insertion sequences, fully abrogated cooperativity, whereas increasing it to only 34 nt did not (Figure 1G). Cooperativity was maintained when both sites were changed to miR-1 sites (Figure 1H). Importantly, repression from UTRs containing either an intact upstream or downstream site alone (purple or yellow bars, respectively) was not significantly altered by the different inter-site sequences employed. Together, results from conservation and reporters indicated that the cooperativity of close sites applied to both mRNA destabilization and repression at the protein level.

The opportunities for cooperative miRNA function would dramatically increase if sites to two different miRNAs could act cooperatively. To test this possibility, we tested reporters based on UTR fragments containing closely spaced sites for miR-1 and miR-133, two miRNAs co-expressed in muscle cells (Figure 1I). The one with 8-nt inter-site spacing exhibited cooperative repression in response to a mixture of miR-1 and miR-133. The one with 4-nt spacing did not, which we attribute to spacing that is too close, in that cooperativity was achieved by extending the spacing by another 6 nt. To test more generally whether sites to different miRNAs act cooperatively when closely spaced, we examined whether sites to our transfected miRNAs were more effective if they were near to sites for miRNAs endogenously expressed in the cells. Repression was strongest when spacing between the endogenous and the transfected miRNA sites was between 8 and 39 nt (Figure 1J).

We conclude that one of the important context determinants that influences efficacy of sites is their proximity to sites for coexpressed miRNAs. A recent report, published while our paper was in review, reached similar conclusions (Doench and Sharp, 2004). Cooperative miRNA function implies a mechanism whereby repression can become more sensitive to small changes in miRNA levels. Moreover, cooperativity of sites for co-expressed miRNAs greatly enhances the regulatory effect and utility of combinatorial miRNA expression.

### Additional Watson-Crick Pairing at Nucleotides 12-17 Enhances miRNA Targeting

We next searched for evidence that pairing to the 3' portion of the miRNA might provide another context determinant that could enhance repression of canonical 7- or 8mer sites. When starting with previously developed energy-based rubrics for predicting and scoring 3' supplementary pairing (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Lewis et al., 2003) those sites with better scores were no more effective; indeed, increased supplementary pairing, as measured by these rubrics, appeared to disfavor efficacy (Figure S2A). We also examined the results of a different approach that predicted sites with extensive Watson-Crick and G:U pairing along the length of the miRNA and tolerated imperfect seed matches (Miranda et al., 2006). However, after excluding those predictions with canonical seed sites, the remainder performed no better than messages without sites (Figure S2B). We therefore searched for a rubric that accomplished its intended purpose of identifying productive 3' pairing, evaluating site efficacy while systematically varying position, continuity, and identity (± G:U pairs) of pairing. Overall, consequential pairing preferentially involved Watson-Crick pairing to miRNA nucleotides 12-17, most especially nucleotides 13-16 (Figures 2A). Watson-Crick pairing to four contiguous nucleotides was most associated with down-regulation if it started at position 13 (Figure 2B). Regarding the position of the UTR segment that paired to the miRNA 3' region, the most optimal arrangement placed it directly opposite the miRNA strand, although several neighboring registers were also effective (Figure 2C).

Turning to evolutionary conservation, Watson-Crick pairing to four contiguous miRNA nucleotides was substantially more conserved when this pairing started at miRNA positions 12, 13, or 14 (Figure 2D), and those same nucleotides at the core of effective 3' pairing were the best conserved miRNA nucleotides outside of the seed region (Figure 2E). Thus, site conservation results were in accord with the experimental results from the array, as would be expected if this newly identified supplemental pairing influences protein output in the animal.

The similarities between 3' pairing and seed pairing were striking. Analogous to seed pairing, 3' pairing was relatively insensitive to predicted thermodynamic stability and instead quite sensitive to geometry, preferring contiguous Watson-Crick pairs uninterrupted by wobbles, bulges, or other mismatches. Also like seed pairing, 3' pairing was sensitive to position, with pairing at the 3' core (positions 13-16) being more important for efficacy than pairing to other positions. Nonetheless, the position requirement was less stringent than for seed pairing, with several nearby registers tolerated (Figure 2B-D).

Using the guidelines derived from analyses of site efficacy and conservation, we developed a simple scheme for scoring 3' pairing based on rewarding pairing throughout the 3' end of the miRNA, but with particular emphasis on the 13-16 nucleotide region. Comparing the efficacy of high-scoring sites with that of low-scoring sites revealed that 3' pairing was an effective determinant, most particularly for 7mer-m8 sites (Figure 2F), although the magnitude was less than the difference between a 7mer site and an 8mer site.

We suspect that very extensive 3' pairing might be more effective than that observed in Figure 2F, but that very extensive pairing is too rare to be reliably evaluated in our array analysis. Extensive 3' pairing also appears to be utilized relatively rarely during biological targeting in mammals. For example, the numbers of canonical sites (6-8mers) with extensive 3' pairing (comprising ≥ 5 contiguous, well-positioned pairs) that were conserved above background averaged two per miRNA family. Nonetheless, we anticipate our newly developed guidelines for detecting consequential 3' pairing will help identify unusual but important cases in which extensive 3' pairing is crucial for mammalian target repression. For example, 3' pairing can help compensate for imperfect seed pairing (Brennecke et al., 2005; Doench and Sharp, 2004), and a few sites with extensive 3' compensatory pairing, including the *let-7* sites in *C. elegans lin-41*, and the miR-196 site in mammalian *HoxB8* are known to function in mammalian cells (Lewis et al., 2003; Reinhart et al., 2000; Yekta et al., 2004). In all of these cases, the sites would have exceptionally high scores using our rubric (≥10 contiguous Watson-Crick base pairs).

## Effective Sites Preferentially Reside Within a Locally AU-rich Context

We next searched for nucleotide composition properties shared by 3'UTRs of downregulated messages. When sites were divided into functional and nonfunctional sites based on their performance on the microarray, we found that the nucleotides immediately flanking the functional sites were highly enriched for A and U content relative to the nonfunctional sites (Figure 3A, green plot). This phenomenon of high local AU content was important in the immediate vicinity of the site and then fell off quickly. A comparison of the local nucleotide composition adjacent to conserved and nonconserved miRNA sites agreed well with the array data; the nucleotides immediately flanking the conserved sites were highly enriched for A and U content relative to those flanking the nonconserved sites (Figure 3A, blue plot).

We developed a rubric that considered the composition of residues 30 nt upstream and 30 nt downstream of the seed site, with weighting tailing off with the inverse of the distance from the seed site (Figure 3B). The 7mers scoring in the top quartile by our rubric appeared at least as effective as 8mers scoring in the bottom two quartiles, illustrating the substantial influence of local AU composition for site efficacy (Figures 3C).

As an additional test of whether sites within high local AU density were more effective in the animal, we performed site-depletion analysis. This analysis was based on the finding that messages preferentially expressed in the same tissue as the miRNA are depleted for sites matching that miRNA, because these highly expressed messages have been under selective pressure to avoid deleterious repression (Farh et al., 2005). Sites within high local AU density, as measured by our rubric, were significantly more depleted in messages preferentially co-expressed with the miRNA compared to sites within low AU density (Figure 3D, $P < 10^{-6}$, 1-sided K-S test). Because site depletion, as with site conservation, is a function of protein downregulation in the animal, our results indicated that local AU content impacts not only mRNA destabilization but also protein expression. We conclude that AU composition in the immediate vicinity of the 7-8mer match is a major context feature influencing mammalian miRNA targeting specificity.

High global AU composition within 3'UTRs also correlates with a higher density of conserved miRNA complementary sites, as well as higher 3'UTR conservation more broadly (Robins and Press, 2005). When considering nucleotide composition of the entire 3'UTR, we observed a correlation between global AU content and efficacy on the array. However, this correlation was not as strong as for local composition, and after accounting for local AU context, no significant residual correlation remained for global AU composition, indicating that global AU composition does not directly influence site efficacy.

Because the preference for A's and U's in the vicinity of the site appeared to involve no more than fortuitous Watson-Crick pairing, this local AU determinant resembled the preference for A across from nucleotide 1, thereby extending the apparent non-Watson-Crick component of site recognition far beyond position 1. Indeed, the preference for A at this position might be considered a component of the local AU effect, skewed to favor A over U. Local AU density also explained why previous methods designed to identify and score 3' pairing were counterproductive for finding more effective sites (see previous section). Each of these methods uses predicted folding free energy to score 3' pairing (Brennecke et al., 2005; Krek et al., 2005; Lewis et al., 2005; Lewis et al., 2003). Seed sites present in regions of higher local GC content will tend to have predicted 3' pairing with better folding free energy due to fortuitous GC base-pairing, when in fact they are significantly less effective.

**Effective Sites Preferentially Reside in the 3'UTR but not too Close to the Stop Codon**

Although the majority of investigation into miRNA function has been for sites located in 3'UTRs, the 5'UTRs and open reading frames (ORFs) of mammalian genes contain, on average, twice the sequence length as 3'UTRs, and artificial siRNAs can direct cleavage at perfectly complementary sites throughout the message. For single 8mer sites, we observed no detectable efficacy in 5'UTRs, detectable but marginal efficacy in ORFs, and high efficacy in 3'UTRs (Figure 4A). These results mirrored those from previous site-conservation analysis, expression arrays, and site-depletion analysis (Farh et al., 2005; Lewis et al., 2005; Lim et al., 2005).

To explore the large difference between efficacy of sites in the ORF and near the beginning of the 3'UTR, we focused on the performance of sites located near the transition of these two regions. When plotting the number of sites conserved above chance, we found that this number, which was low in the ORF, remained low in the first ~15 nt following the stop codon, and thus the transition between ORF targeting and 3'UTR targeting did not occur precisely at the stop codon, but was offset ~15 nt downstream (Figure 4B). In agreement with this finding, 3'UTR sites within 15 nt of the stop codon were less effective on the array, compared to sites elsewhere in the 3'UTR ($P < 0.01$ for a 1-sided K-S test comparing expression values for messages with an 8mer, data not shown).

To confirm these findings on the protein level, activity was compared for reporters that were identical except at two nucleotides that disrupted two stop codons, thereby extending the ORF by 69 residues and bringing into service a stop codon falling within 8 nt of a miR-124 site (Figure 5C). These two point substitutions, both more than 70 nt upstream of the site, specifically abrogated function of the miR-124 site, while having no effect on the function of a downstream miR-124 site (Figure 4C). These results confirmed that the segment immediately following the stop codon was inhospitable for targeting.

### Effective Sites Preferentially Reside Near Both Ends of the 3'UTR

We next examined whether the location of the site within the remainder of the 3'UTR influenced performance. Sites residing near the two ends of long UTRs were substantially more effective than those near the center (Figures 5A). Moreover, site-conservation analysis revealed that more sites were selectively maintained near the ends than in the central region (Figure 5B). Site-depletion in messages preferentially co-expressed with miRNAs was also more severe near the ends of long UTRs than near the center, indicating that newly emergent sites are more likely to be functional in the animal if they fall near the ends (Figure 5C, $P = 0.032$, 1-sided K-S test). Thus, all three lines of evidence—the experimental approach, which monitored mRNA destabilization, plus the two computational approaches, which addressed protein output in the animal—indicated that the UTR quartiles near the ORF and near the poly(A) tail were more hospitable for effective targeting than were the two central quartiles. This effect was most pronounced for longer UTRs (>1300 nt).

### A Quantitative Model for Site Efficacy

To develop a quantitative tool for predicting the efficacy of single sites, irrespective of their evolutionary conservation, we used linear regression to model the relationship between downregulation on the array and the context of single 7mer-m8 sites. To build a model general to various cell types, proximity to sites for co-expressed miRNAs was not considered. Sites in 5'UTRs, in ORFs, or within 15 nt of the stop codon were excluded, in accord with our results showing that such sites were generally not effective (Figure 4). For additional pairing to the 3' region of the miRNA, scores were calculated as described in Figure 2F, with the score equal to the greatest number of contiguously paired bases, weighted towards pairing at nucleotides 13-16 (Figure 6A). For local AU composition, scores were calculated as described in Figure 3B, with a higher score indicating greater local AU composition in the region flanking the site (Figure 6B). For position effects, scores were based on the distance in nucleotides between the site and the closest end of the 3'UTR (Figure 6C).

The scores for each context feature were not significantly co-correlated (Figure S4), indicating that their effects could be considered independently and combined into a single model (Figure 6D). When tested on an independent series of siRNA transfections this model accurately anticipated the messages destabilized in response to the siRNAs (Figure 6E), and the individual features of the model were each effective on their own (Figure S5). This result showed that our model was predictive for data that had not been used in its derivation and demonstrated that it applied to siRNA off-targeting as well as miRNA targeting.

Using the same approach, we considered the 8mer, 7mer-A1, and 6mer sites and combined the models for these sites into a single unified model. To address the question of how evolutionary conservation impinged on the model, the effects were plotted separately for conserved and nonconserved sites (Figure 6F). For both conserved and nonconserved sites the unified model yielded a clear correspondence between the predicted and observed effects on message stability. Nonetheless, conserved sites with the same score were slightly more effective than were the nonconserved sites. Perhaps additional, uncharacterized context features remain to be discovered and have been under selective pressure for optimization, leading to the greater observed efficacy of the conserved sites. Alternatively, our model might not be parameterized to fully capture the effects of known features, or for conserved sites evolution might have combined known factors into more optimal, synergistic arrangements.

Despite the difference in magnitude of mRNA destabilization for conserved and nonconserved sites, correspondence between predicted and observed effects for these two classes of sites was indistinguishable in terms of its slope. The strong correspondence when considering only the nonconserved sites alleviated concern that our context determinants might have merely pointed to conserved sites and thereby artifactually predicted repression primarily through unknown determinants associated with conserved sites. Such a concern would have been particularly relevant for the AU-content and UTR-position features because these features correlated with all conserved sites, not just those matching miRNAs (Supplementary discussion). Further alleviating this concern was the performance with siRNA off-targets, which proved that the model as well its component features functioned well for sites that have not been under any selective pressure for targeting efficacy (Figures 6E and S5). Moreover, local AU content and UTR position were each supported by site-depletion analysis, which is also independent of conservation and has the added benefit of speaking to targeting efficacy in the animal (Figures 3D and 5C).

To illustrate the utility of our model for predicting which sites are more likely to mediate repression, we examined the destabilization of messages with single sites predicted to be in favorable or unfavorable contexts (Figure 6G). For messages with 7mer-m8 sites, the minimal fraction of messages downregulated on the array was 0.49 for the sites in favorable contexts, a substantial discrimination when compared to the 0.09 value observed for sites in unfavorable contexts. Analogous discrimination was observed for the other types of canonical sites.

To evaluate the model further, we constructed 25 reporters and 25 mutant-control reporters designed to test the efficacy of single 7mer-m8 miR-25 sites in either a favorable or an unfavorable context, as predicted by the model. miR-25, which was not among the miRNAs used in the array experiments, was chosen to confirm that the model extended to other miRNAs, and assays were performed in HEK293 cells to confirm that the model extended to cell types other than HeLa cells. UTRs were selected without regard to site conservation. In retrospect, more sites in favorable contexts were conserved, as was expected (6 of 13, compared to 1 of 12).

Nearly all of the sites in favorable contexts yielded significant repression (Figure 6H), whereas none of the sites in unfavorable contexts yielded detectable repression (Figure 6I). Importantly, of those 7mer-m8 sites predicted to be in favorable contexts, the fraction yielding repression in the reporter assay substantially exceeded the 0.49 value derived from the expression array data alone (Figure 6G,6H). Two factors explained this observation. The first was that the arrays underestimated the number of downregulated messages because of their large measurement noise. For example, in a scenario in which every message with a 7mer site was downregulated by 15%, the 7mer expression distribution would shift to the left, but because of the noise of the array experiment (as indicated by the spread of the no-site distribution), only 53% of the sites would be scored as effective. The second factor was that the array monitored only mRNA destabilization, whereas the reporter experiments monitored protein output. Therefore, messages that were repressed translationally with little or no change in mRNA level would yield detectable repression with the reporters but not the arrays. Regardless of the relative contributions of these two factors, the larger extent of repression observed at the protein level with the reporters compared to that observed at the RNA level with the arrays indicated that the context determinants uncovered in our study were general—they successfully predicted sites that mediated repression at the mRNA level and those that mediated it at the protein level. Their general relevance was expected because the context determinants were supported by the site-conservation and site-depletion analyses.

Because our approaches searched for general context determinants, we could not exclude the existence of additional, unrelated context determinants specific to targets repressed only at the protein level. However, we have no reason to suspect that such specific determinants exist. Indeed, the absence of detectable targeting for all the assayed sites predicted to fall in non-favorable contexts

(Figure 6I) indicated that if such unknown determinants specific to translational repression exist, they either are rarely present or have only negligible effects.

### Specificity Determinants apply to endogenous miRNA-message interactions

From the results of our conservation and depletion analyses, which examined the evolutionary effects of in vivo miRNA targeting, we anticipated that our model would apply not only to exogenously supplied miRNAs and siRNAs but also to endogenous miRNA-message interactions (Figures 2-6). To confirm this relevance to targeting in vivo, we used the model to predict the effects on endogenous messages after perturbing three miRNAs: inhibition of miR-122 in the adult mouse liver, knockout and rescue of miR-430 in the zebrafish embryo, and ablation of miR-155 in murine T cells (Krutzfeldt et al., 2005; Giraldez et al., 2006; Rodriguez et al., 2007). In each case, context scores of messages with single sites to the relevant miRNA corresponded significantly to the in vivo response, thereby confirming the predictive power of our model for miRNA targeting in vivo (Figure 7A-D; $P < 10^{-5}$ for each panel, $\rho$ = -0.24, 0.31, -0.14, -.26, respectively, Spearman rank correlation tests). We next evaluated individual context features of the combined model. For each feature, the predicted repression for messages with single miR-155 sites corresponded significantly to the in vivo response, thereby confirming the relevance of 3' pairing, local AU content and site position for endogenous targeting in the mouse (Figure 7E-G).

### Mechanistic Implications

The mechanism of specificity can be explained at least partly in terms of site accessibility and site affinity, which influence the association and dissociation of the silencing complex. For example, the differential efficacy of 8mer, 7mer, and 6mer sites presumably reflects differences in binding affinity. Supplemental pairing outside of the seed region, particularly to nucleotides 13-16 of the miRNA, could further decrease the dissociation rate of the bound silencing complex. In contrast, the local AU context determinant might be associated with weaker mRNA secondary structure in the vicinity of the site and thus increased accessibility to the seed site. Our observation that the local AU nucleotide composition bias centered on the seed site, and not on a larger region, supports the proposal that the seed site nucleates the initial interaction with the silencing complex, with additional pairing playing a role later in keeping the miRNA and mRNA complex more stably associated (Bartel, 2004).

Increased repression with multiple sites could be explained by an increased likelihood of any one site being bound or by a beneficial effect of having more than one silencing complex bound simultaneously. Our observation that overlapping or near-overlapping sites for two different miRNAs yielded less downregulation than did more distantly spaced sites favored the latter possibility. A beneficial effect of having multiple silencing complexes bound simultaneously implied that interactions with the downstream repressive machinery are limiting, either in quantity or duration, for repression. The finding that cooperativity extended to different sites for co-expressed miRNAs ruled out as the mechanism of cooperative action a local-concentration effect in which the second, nearby site merely re-captures a dissociating complex. Instead, the cooperative function of optimally spaced sites might be explained by cooperative contacts with the repressive machinery. Alternatively, it might be explained by binding at one site increasing binding at the other site, either through favorable contacts between silencing complexes or by displacing occlusive mRNA structure.

Additional evidence that miRNA sites must remain bound in order to confer repression came from the reduced site effectiveness in the 5'UTR and ORF, which presumably results from displacement of silencing complexes as the ribosome translocates from the cap-binding complex to the stop codon. We found that this effect extended beyond the stop codon and into the first ~15 nt of the 3'UTR. The length of ~15 nt was in agreement with crystallographic and functional studies demonstrating that the mRNA enters the ribosome ~15 nt downstream of the decoding site (Takyar et al., 2005; Yusupova et al., 2001), which would presumably strip off any silencing complex and block rebinding until the ribosome dissociated from the message. The apparent interference by the ribosome along its entire path of translation strongly implied that messages under miRNA control experience at least one complete round of translation prior to or concurrently with their repression, thereby disfavoring models in which an appreciable fraction of messages are sequestered prior to translation of a full-length protein. In contrast to natural miRNA sites, sites perfectly complementary to artificial siRNAs function well in ORFs and thus do not appear subject to ribosome interference. Perhaps the ribosome has more difficulty disrupting the more extensive pairing. Moreover, extensively paired sites might not need to remain associated because they result in catalytic cleavage, whereas 7-8mer sites might require longer periods of association to confer appreciable repression.

The increased efficacy of sites falling near the ends of long UTRs might be attributed either to proximity with the translation machinery or to increased site accessibility. Within the circularized structures of mRNAs, with the poly(A) tail interacting with the 5' cap, sites located in the middle of long 3'UTRs would be furthest from the translational machinery, whereas sites closer to the ORF and 5'UTR might be better situated to interact with the translation machinery and hence induce repression. Proximity would be greater within the 5'UTR or ORF, but such sites would face ribosome interference, leaving sites near the ends of the 3'UTR as the most optimal. Alternatively, if a long

UTR is visualized as a cloud of interconverting structures, bound by the ribosome on one end and the polyA-binding proteins on the other, the regions in the middle of the UTR would be expected to be less accessible because they would have opportunities to form occlusive interactions with segments from either side, whereas residues near the ends would not.

The strong influence of local AU content was another context determinant suggesting a role for occlusive UTR structure. We followed up on this possibility using predicted local secondary structure with published rubrics to score site accessibility ((Robins and Press, 2005). The method of Long et al. (2007) is not yet available for large-scale analyses, but we were able to evaluate the sites of Figure 6H using STarMir (sfold.wadsworth.org), which implements this algorithm. Of the thirteen sites we predicted to be in optimal context, STarMir predicted only one to be functional. When implementing the method of Zhao et al. in a genome-wide analysis, a correlation was observed between downregulation and weaker secondary structure in the vicinity of the site, but secondary structure prediction was less informative than was local AU content and had no utility after accounting for local AU content (Figure S3). Perhaps direct recognition of A's and U's flanking the seed sites is a component of target recognition, in which case scoring local AU content would provide a more reliable measure of this recognition feature than would secondary-structure predictions. Or perhaps because of RNA-binding proteins, RNA tertiary structure, and the compact but multiple competing conformations of arbitrary RNA sequence (Schultes et al., 2005), the details of intracellular UTR structures differ substantially from those of the predicted structures, such that scoring local AU content is significantly more reliable for predicting site accessibility.

To the extent that the context features revealed in our analyses reflected the negative effects of occlusive secondary structure and ribosome interference, we anticipate that they will apply not only to miRNA regulatory sites but also to a range of other elements, including binding sites for regulatory proteins. For instance, messages with many conserved regulatory elements would be associated with higher overall AU content, whereas messages selectively avoiding any regulatory targeting their 3'UTR would benefit from increased GC content that makes any newly emergent fortuitous sites less likely to function, thereby helping to explain the strong correlation between high nucleotide conservation and high global AU composition in 3'UTRs. Likewise, poorer accessibility of elements within 15 nt of the stop codon and near the center of long UTRs would explain why other 3'UTR regions have higher nucleotide conservation.

### A Resource for Prioritizing Conserved Sites and Predicting Functional Nonconserved Sites

In addition to providing insights and constraints for mechanistic models, our newly identified and parameterized context determinants provide valuable information for selecting which of the many mammalian miRNA:target relationships are most promising for experimental follow-up. Accordingly, for all conserved and nonconserved 7-8mer sites matching known miRNAs, each of these determinants is evaluated and reported (Figure S6; targetscan.org), with the goal of providing a resource for enabling even more rapid progress in understanding this recently appreciated mode of gene regulation.

## Experimental Procedures

### MicroRNA and mRNA Sequence Data

MicroRNAs conserved in human, mouse, rat, dog, and zebrafish were clustered into 73 families based on miRNA nucleotides 2-8 (Table S1). Human annotated 5'UTR, ORF, and 3'UTR sequences were obtained from RefSeq, and orthologous sequences in mouse, rat, and dog were derived from the UCSC genome browser multiZ multiple genome alignments (Blanchette et al., 2004). When multiple RefSeq identifiers mapped to a single Entrez Gene entry, the RefSeq annotation with the longest UTR was used.

### Conservation Analysis

Site-conservation analysis was as in Lewis et al. (2005), but with an improved method for choosing the control sequences used to estimate the background conservation (Farh et al., in preparation). As before, control motifs were chosen to have an abundance in human 3'UTRs as close as possible to that of the authentic miRNA sites. In addition, to account for the observation that the four nucleotides are conserved at different rates in mammalian 3'UTRs, control sequences were also limited to those that had identical or nearly identical (within one nucleotide) composition as the authentic sites.

### Array Experiments

HeLa cells were transfected with synthetic miRNA duplexes (Table S2), and after 12 and 24 hours mRNA was extracted and analyzed on Agilent arrays in duplicate, using Agilent array software to obtain values for intensity and change relative to mock transfection with associated *P* values, as described (Lim et al., 2005). Probes were mapped to their representative mRNA sequence through Entrez Gene. Only those probes that were above median intensity in at least half of all experiments were considered for the expression analysis, which limited the analysis to genes expressed at a sufficiently high level that down-regulation would be readily detectable. When multiple probes matched a single gene, their geometric mean and most significant *P* value were used. Array data from siRNA transfections was processed analogously.

For displaying the changes in mRNA abundance (Figure 1B), the down-regulation values for messages with a cognate site were binned such that each bin corresponded to a range of values of 0.1 on a $\log_2$ scale. Unless indicated otherwise, only messages with a single site to the cognate miRNA were considered. For instance, when evaluating single 8mer sites, only genes with a single 8mer in the UTR and no additional 7mers or 6mers were considered. Likewise, when a single 6mer was being considered, genes were excluded in which that 6mer was part of a canonical 7-8mer. Analyses for multiple sites (Figures 1C-1E, 1J) and additional specificity determinants (Figures 2-7) were performed in the same way, requiring the exact numbers of the motif being examined and no other canonical sites, except that additional 6mers were allowed in order to increase the sample sizes, based on the observation that the presence of additional 6mers had marginal effect in the context of stronger sites. To evaluate the significance of differences between fold-change distributions, K-S tests were performed as 1-sided tests when testing hypotheses and 2-sided tests when testing for differences.

### Calculating the Fraction of Genes Downregulated

To determine for each type of site the minimal fraction of genes down-regulated on the array, we compared the cumulative distribution of expression changes for messages with the site versus those with no canonical site, calculating the maximum positive cumulative difference between the two distributions. To correct for bumpiness in the cumulative distributions, we performed 100 permutations in which the down-regulation values for all genes were randomly shuffled, while maintaining the size of the original distributions, and the median maximum positive deviation for these 100 control permutations was subtracted from the value obtained from the real distributions.

### Depletion Analysis

Site-depletion analyses were as in Farh et al. (2005) but focused on a subset of tissues and tissue-specific miRNAs with very robust depletion signatures and employed only mouse UTR data. To test the hypothesis that sites falling in one particular context were more depleted than in the other context, direct comparisons were performed between sites in the two contexts by aggregating the relative expression values for targeted genes in each tissue expressing the cognate miRNA (boxed in the figures). A 1-sided K-S test evaluated the difference between the aggregate distributions.

### Simulating Repression from Dual Sites

The simulated dual-site distributions (green lines, Figure 1D-E) were derived by selecting two genes randomly from the single 7mer site distribution and summing their $\log_2$ expression changes. This procedure was repeated 100 times for each gene in the dual-site distribution. For comparison to the simulated distributions, the observed dual-site distributions were modified by selecting one gene randomly from the dual site distribution and one gene randomly from the no-site distribution and summing their $\log_2$ expression changes (blue lines, Figure 1A-B). This modification was necessary for comparison because by summing two distributions the variance of the resulting distribution greatly increased because microarray measurement noise was compounded. The distributions for single site plus no site and for no site plus no site are also shown for comparison. Because genes with dual sites typically had longer 3'UTRs than those with one or no sites, the sampling was not performed entirely at random; to control for possible length effects, pairs of randomly sampled genes were required to have UTR lengths within 30% of each other.

### Reporter assays

HeLa cells were transfected using Lipofectamine 2000 (Invitrogen) in 24-well plates (~0.3 x $10^5$ cells per well) with 25 ng firefly-luciferase control reporter (pIS0, (Lewis et al., 2003)), 10 ng SV40-*Renilla*-luciferase reporter, 1.25 µg pUC19 and miRNA duplex. Experiments transfecting a single miRNA used 25 nM miRNA duplex. Because co-transfection of one miRNA titrated repression by the other, relative miRNA concentrations were adjusted when co-transfecting miRNAs so that both were nearly equivalently active when combined. Experiments with two miRNAs were performed with 25 nM miR-1 and 1 nM miR-133 (Figure 1I). Luciferase activities were measured 24 hours after transfection with the Dual-luciferase assay (Promega). To control for transfection efficiency, *Renilla* activity was divided by firefly activity. For each construct assayed, multiple independent experiments, each comprising three replicate values, were combined. To combine replicate values from independent experiments, *Renilla* values were normalized to the geometric mean of *Renilla* values from otherwise identical constructs in which all sites were disrupted. Values plotted for each construct are the geometric mean of normalized *Renilla* values from transfection with the cognate miRNA divided by that value from transfection with the noncognate miRNA. Construct and miRNA duplex sequences are provided (Tables S3-S5). Reporter assays utilizing HEK293 cells were performed identically, except that transfection mixes (per well) contained 100 ng firefly-luciferase control reporter (pIS0), 100 ng TK-*Renilla*-luciferase reporter, 1 µg pUC19 and 25 nM miRNA duplex.

## Acknowledgments

We are grateful to J. Schelter and P. Linsley for providing the microarray data for the miR-133 transfection and to the Rosetta Gene Expression Laboratory for microarray work. Supported by an NIH grant (D.B.) and NIH postdoctoral fellowship (A.G.). D.B. is an HHMI investigator.

## References

Ambros, V. (1989). A hierarchy of regulatory genes controls a larva-to-adult developmental switch in C. elegans. Cell *57*, 49-57.

Ambros, V. (2004). The functions of animal microRNAs. Nature *431*, 350-355.

Ambros, V., Bartel, B., Bartel, D. P., Burge, C. B., Carrington, J. C., Chen, X., Dreyfuss, G., Eddy, S. R., Griffiths-Jones, S., Marshall, M., *et al.* (2003). A uniform system for microRNA annotation. Rna *9*, 277-279.

Ambros, V., and Horvitz, H. R. (1984). Heterochronic mutants of the nematode Caenorhabditis elegans. Science *226*, 409-416.

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (submitted). The impact of microRNAs on protein output.

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.

Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. Cell *116*, 281-297.

Bartel, D. P., and Chen, C. Z. (2004). Micromanagers of gene expression: the potentially widespread influence of metazoan microRNAs. Nat Rev Genet *5*, 396-400.

Baskerville, S., and Bartel, D. P. (2005). Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes. RNA *11*, 241-247.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D., *et al.* (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res *14*, 708-715.

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

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

Chalfie, M., Horvitz, H. R., and Sulston, J. E. (1981). Mutations that lead to reiterations in the cell lineages of *C. elegans*. Cell *24*, 59-69.

Chen, K., and Rajewsky, N. (2006). Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes. Cold Spring Harb Symp Quant Biol *71*, 149-156.

Didiano, D., and Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol *13*, 849-851.

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

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

Eisenberg, E., and Levanon, E. Y. (2003). Human housekeeping genes are compact. Trends Genet *19*, 362-365.

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

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.

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (submitted).

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.

Griffiths-Jones, S. (2004). The microRNA Registry. Nucleic Acids Res *32*, D109-111.

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

He, L., Thomson, J. M., Hemann, M. T., Hernando-Monge, E., Mu, D., Goodson, S., Powers, S., Cordon-Cardo, C., Lowe, S. W., Hannon, G. J., and Hammond, S. M. (2005). A microRNA polycistron as a potential human oncogene. Nature *435*, 828-833.

Hofacker, I., Fontana, W, Stadler, PF, Bonhoeffer, S, Tacker, M, Shuster P (1994). Fast folding and comparison of RNA secondary structures. Monatshefte fur Chemie, 167-168.

Hornstein, E., Mansfield, J. H., Yekta, S., Hu, J. K.-H., Harfe, B. D., McManus, M. T., Baskerville, S., Bartel, D. P., and Tabin, C. J. (2005). miR-196 acts upstream of Hoxb-8 and Shh in limb development. Nature, in press.

Houbaviy, H. B., Murray, M. F., and Sharp, P. A. (2003). Embryonic stem cell-specific MicroRNAs. Dev Cell *5*, 351-358.

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

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.

Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., Di Padova, F., Lin, S. C., Gram, H., and Han, J. (2005). Involvement of microRNA in AU-rich element-mediated mRNA instability. Cell *120*, 623-634.

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

Johnnidis, J. B., Harris, M. H., Wheeler, R. T., Stehling-Sun, S., Lam, M. H., Kirak, O., Brummelkamp, T. R., Fleming, M. D., and Camargo, F. D. (2008). Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. Nature *451*, 1125-1129.

Johnston, R. J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. Nature *426*, 845-849.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T. S., Hinrichs, A., Lu, Y. T., Roskin, K. M., Schwartz, M., Sugnet, C. W., Thomas, D. J., *et al.* (2003). The UCSC Genome Browser Database. Nucleic Acids Res *31*, 51-54.

Kiriakidou, M., Nelson, P. T., Kouranov, A., Fitziev, P., Bouyioukos, C., Mourelatos, Z., and Hatzigeorgiou, A. (2004). A combined computational-experimental approach predicts human microRNA targets. Genes Dev *18*, 1165-1178.

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

Krutzfeldt, J., Rajewsky, N., Braich, R., Rajeev, K. G., Tuschl, T., Manoharan, M., and Stoffel, M. (2005). Silencing of microRNAs in vivo with 'antagomirs'. Nature *438*, 685-689.

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

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

Lai, E. C. (2002). Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. Nat Genet *30*, 363-364.

Lai, E. C., Burks, C., and Posakony, J. W. (1998). The K box, a conserved 3' UTR sequence motif, negatively regulates accumulation of *Enhancer of split* Complex transcripts. Development *125*, 4077-4088.

Lai, E. C., and Posakony, J. W. (1997). The Bearded box, a novel 3' UTR sequence motif, mediates negative post-transcriptional regulation of *Bearded* and *Enhancer of split* Complex gene expression. Development *124*, 4847-4856.

Lai, E. C., Tam, B., and Rubin, G. M. (2005). Pervasive regulation of *Drosophila* Notch target genes by GY-box-, Brd-box-, and K-box-class microRNAs. Genes Dev *19*, 1067-1080.

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science *294*, 858-862.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science *294*, 862-864.

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., 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.

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.

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.

Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. Nat Struct Mol Biol *14*, 287-294.

MacIsaac, K. D., and Fraenkel, E. (2006). Practical strategies for discovering regulatory DNA sequence motifs. PLoS Comput Biol *2*, e36.

Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J., *et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

Martick, M., Horan, L. H., Noller, H. F., and Scott, W. G. (2008). A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. Nature *454*, 899-902.

Miranda, K. C., Huynh, T., Tay, Y., Ang, Y. S., Tam, W. L., Thomson, A. M., Lim, B., and Rigoutsos, I. (2006). A pattern-based method for the identification of microRNA binding sites and their corresponding heteroduplexes. Cell *126*, 1203-1217.

Mootha, V. K., Lindgren, C. M., Eriksson, K. F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstrale, M., Laurila, E., *et al.* (2003). PGC-1alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. Nat Genet *34*, 267-273.

Ota, A., Tagawa, H., Karnan, S., Tsuzuki, S., Karpas, A., Kira, S., Yoshida, Y., and Seto, M. (2004). Identification and characterization of a novel gene, C13orf25, as a target for 13q31-q32 amplification in malignant lymphoma. Cancer Res *64*, 3087-3095.

Pruitt, K. D., Tatusova, T., and Maglott, D. R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501-504.

Rajewsky, N., and Socci, N. D. (2004). Computational identification of microRNA targets. Dev Biol *267*, 529-535.

Rao, P. K., Farkhondeh, M., Baskerville, S., and Lodish, H. F. (data not shown).

Rao, P. K., Farkhondeh, M., Baskerville, S., and Lodish, H. F. (unpublished data).

Reinhart, B. J., Slack, F. J., Basson, M., Pasquinelli, A. E., Bettinger, J. C., Rougvie, A. E., Horvitz, H. R., and Ruvkun, G. (2000). The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature *403*, 901-906.

Reinhart, B. J., Weinstein, E. G., Rhoades, M. W., Bartel, B., and Bartel, D. P. (2002). MicroRNAs in plants. Genes Dev *16*, 1616-1626.

Rhoades, M. W., Reinhart, B. J., Lim, L. P., Burge, C. B., Bartel, B., and Bartel, D. P. (2002). Prediction of plant microRNA targets. Cell *110*, 513-520.

Robins, H., Li, Y., and Padgett, R. W. (2005). Incorporating structure to predict microRNA targets. Proc Natl Acad Sci U S A *102*, 4006-4009.

Robins, H., and Press, W. H. (2005). Human microRNAs target a functionally distinct population of genes with AT-rich 3' UTRs. Proc Natl Acad Sci U S A *102*, 15557-15562.

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, O., Jr., 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.

Schuler, A., Schwieger, M., Engelmann, A., Weber, K., Horn, S., Muller, U., Arnold, M. A., Olson, E. N., and Stocking, C. (2008). The MADS transcription factor Mef2c is a pivotal modulator of myeloid cell fate. Blood *111*, 4532-4541.

Schultes, E. A., Spasic, A., Mohanty, U., and Bartel, D. P. (2005). Compact and ordered collapse of randomly generated RNA sequences. Nat Struct Mol Biol *12*, 1130-1136.

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

Sempere, L. F., Freemantle, S., Pitha-Rowe, I., Moss, E., Dmitrovsky, E., and Ambros, V. (2004). Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. Genome Biol *5*, R13.

Shkumatava, A., Stark, A., and Bartel, D. P. (submitted).

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.

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

Su, A. I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K. A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G., *et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067.

Suh, M. R., Lee, Y., Kim, J. Y., Kim, S. K., Moon, S. H., Lee, J. Y., Cha, K. Y., Chung, H. M., Yoon, H. S., Moon, S. Y., *et al.* (2004). Human embryonic stem cells express a unique set of microRNAs. Dev Biol *270*, 488-498.

Takyar, S., Hickerson, R. P., and Noller, H. F. (2005). mRNA helicase activity of the ribosome. Cell *120*, 49-58.

Tomczak, K. K., Marinescu, V. D., Ramoni, M. F., Sanoudou, D., Montanaro, F., Han, M., Kunkel, L. M., Kohane, I. S., and Beggs, A. H. (2004). Expression profiling and identification of novel genes involved in myogenic differentiation. Faseb J *18*, 403-405.

Tsang, J., Zhu, J., and van Oudenaarden, A. (2007). MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol Cell *26*, 753-767.

van der Waerden, B. L. (1969). Mathmatical Statistics (Berlin, Springer-Verlag).

Wei, X., Sun, W., Fan, R., Hahn, J., Joetham, A., Li, G., Webb, S., Garrington, T., Dakhama, A., Lucas, J., *et al.* (2003). MEF2C regulates c-Jun but not TNF-alpha gene expression in stimulated mast cells. Eur J Immunol *33*, 2903-2909.

Wienholds, E., Kloosterman, W. P., Miska, E., Alvarez-Saavedra, E., Berezikov, E., de Bruijn, E., Horvitz, H. R., Kauppinen, S., and Plasterk, R. H. (2005). MicroRNA expression in zebrafish embryonic development. Science *309*, 310-311.

Wightman, B., Burglin, T. R., Gatto, J., Arasu, P., and Ruvkun, G. (1991). Negative regulatory sequences in the *lin-14* 3'-untranslated region are necessary to generate a temporal switch during *Caenorhabditis elegans* development. Genes Dev *5*, 1813-1824.

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

Xie, X., Lu, J., Kulbokas, E. J., Golub, T. R., Mootha, V., Lindblad-Toh, K., Lander, E. S., and Kellis, M. (2005). Systematic discovery of regulatory motifs in human promoters and 3' UTRs by comparison of several mammals. Nature *434*, 338-345.

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

Yusupova, G. Z., Yusupov, M. M., Cate, J. H., and Noller, H. F. (2001). The path of messenger RNA through the ribosome. Cell *106*, 233-241.

Zhao, Y., Ransom, J. F., Li, A., Vedantham, V., von Drehle, M., Muth, A. N., Tsuchihashi, T., McManus, M. T., Schwartz, R. J., and Srivastava, D. (2007). Dysregulation of cardiogenesis, cardiac conduction, and cell cycle in mice lacking miRNA-1-2. Cell *129*, 303-317.

Zhao, Y., Samal, E., and Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets *Hand2* during cardiogenesis. Nature *436*, 214-220.

## Figure Legends

**Figure 1.** Downregulation of messages with 6-8mer sites.

(A) Canonical miRNA complementary sites.

(B) Effectiveness of single canonical sites. Changes in abundance of mRNAs following miRNA transfection were monitored with microarrays. Distributions of changes (0.1 unit bins) for messages containing the indicated single sites in their UTRs are shown (left), together with the cumulative distributions (right). The dashed line in the cumulative distributions indicates that 27% of mRNAs with UTRs containing a single 8mer were down-regulated at least 29% ($2^{-0.5} = 0.71$). Results of eleven experiments, each performed in duplicate and each transfecting a duplex for a different miRNA (Table S2), were consolidated. Although the results shown were an amalgam of the data from all 11 miRNAs, the relative strengths of the different sites were consistent when examining each transfection individually. For the cumulative plots, the minimal fraction of downregulated genes in that distribution is reported (parentheses), based on comparison with the no site distribution. Repression from UTRs containing an 8mer site was significantly more than that from UTRs with a 7mer-m1 site ($P < 10^{-20}$, 1-sided K-S test); similar comparisons between UTRs containing a 7mer-m8 site versus a 7mer-A1 site, a 7mer-A1 versus a 6mer, and, a 6mer versus no site were also significant ($P < 10^{-6}$, $P < 10^{-20}$ and $P < 10^{-31}$, respectively).

(C) Increased effectiveness of dual sites. Changes in mRNA abundance following miRNA transfection, represented as in B, except mRNAs with 3'UTRs containing the indicated pairs of sites were monitored. Repression from UTRs containing both an 8mer and either a 7mer or 8mer site was significantly more than that from UTRs with two 7mer-m8 sites ($P < 10^{-3}$, 1-sided K-S test); similar comparisons between UTRs containing two 7mer-m8 sites versus two 7mer-A1 sites, two 7mer-A1 versus two 6mer, and, two 6mer versus no site were also significant ($P = 0.034$, $P < 10^{-11}$ and $P < 10^{-6}$ respectively).

(D) Independence of most dual sites. Cumulative distributions of changes in mRNA levels following miRNA transfection for messages containing the indicated combinations of miRNA binding sites. Simulated values for 3'UTRs containing two 7mer sites (green) were calculated by combining the effect of two single 7mers; actual values for 3'UTRs containing two 7mers are in blue and those with length-matched UTRs containing single 7mers are in purple; otherwise as shown for Figure 1B. Because UTRs with single sites were chosen to have the same lengths as those with dual sites, the repression by single sites differed slightly from that observed for UTRs more generally (Figure 1B).

(E) Synergism between closely spaced sites. Cumulative distributions of changes in mRNA levels as for (D), except the plot for two observed sites (blue) only considered 3'UTRs containing two closely spaced 7mers (within 100 nt of each other). Repression from UTRs containing two adjacent sites was significantly increased compared to simulated UTRs containing one plus one site ($P = 0.040$, 1000 resampling iterations), whereas repression from UTRs containing two sites irrespective of distance did not significantly differ from simulated UTRs containing one plus one site ($P = 0.81$).

(F) Selective maintenance of dual sites spaced at different intervals. Human 3'UTRs with exactly two 7mer sites were binned based on inter-site distance (counting the number of nucleotides between the 3' nt of the first site and the 5' nt of the second site). The number of conserved dual sites exceeding the background (as estimated from the average of control cohorts) was plotted after performing for each bin site-conservation analysis analogous to that in Lewis et al (2005), using the miRNA families conserved broadly among vertebrates (Table S1).

(G) MicroRNA-mediated repression of luciferase reporter genes fused to 3'UTR fragments containing two miR-124 target sites with different spacing intervals. After normalizing to the transfection control, luciferase activity from HeLa cells cotransfected with each reporter construct and its cognate miRNA (miR-124) was normalized to that from cotransfection of each reporter with its non-cognate miRNA (miR-1). Plotted are the normalized values, with error bars representing the third largest and third smallest values among 12 replicates. *P* values (Wilcoxon rank-sum test) indicate whether repression from a reporter containing both sites (blue) was significantly greater than expected from multiplicative effects (green). For modeling independent activity of sites, repression expected from a reporter with two sites was the product of repression observed from otherwise identical reporters containing single intact sites (purple and yellow; pairing off the repression values in the order that they were generated). Reporters of the rightmost quintet were identical to those of the leftmost quintet, except the point substitutions disrupting target sites differed.

(H) Repression observed for the reporter constructs as in the leftmost quintet of (G) but modified such that both miR-124 sites were substituted with miR-1 sites. miR-196 served as the noncognate miRNA.

(I) Repression of reporter constructs containing 3'UTR fragments with naturally closely spaced miR-1 and miR-133 sites, compared to that of mutant derivatives of these fragments. A mixture of miR-1 and miR-133 was co-transfected as the cognate miRNA, and miR-196 served as the noncognate control.

(J) Cooperativity between sites to transfected and endogenous miRNAs in HeLa. Endogenous sites considered were those for *let-7* RNA, miR-16, miR-21, miR-23, miR-24, miR-27, and miR-30 (Tom Tuschl, personal communication). 7mer-m8 sites at a cooperative distance (>7 nt and <40 nt) from an endogenous miRNA 7-8mer site were significantly more downregulated than sites that were either too close to an endogenous miRNA (≤7 nt, including overlapping sites; $P = 0.0054$, one-sided K-S test), or not close to an endogenous site (≥40 nt, or no endogenous site, $P = 0.036$, one-sided K-S test).

**Figure 2.** Characteristics of beneficial 3' pairing.

(A) Pairing scheme, highlighting the core region of productive 3' pairing between the miRNA (orange) and 3'UTR (green). In grey are residues whose pairing was less correlated with increased efficacy or conservation.

(B) Preferential position of supplementary pairing within the miRNA 3' region. Starting with 7mer-m8 sites for representative vertebrate miRNAs of Table S1, a 4mer window was slid across the 3' end of the miRNA, searching for its Watson-Crick match in the opposing region of the message. A 3' pairing score was assigned to each miRNA 4mer, crediting one point for each contiguous pair within the 4mer, a half point for extending the contiguous pairing 1 nt upstream and a half point for extending it 1 nt downstream. The position of the miRNA 4mer and its complement in the message were allowed to be offset, but a ½ point penalty was assessed for each nucleotide of offset beyond ± 2 nt, and pairing to message positions already paired to the miRNA seed region (1-8) was disallowed. When the 4mer had alternative regions of contiguous pairing, it was assigned the highest of the alternative scores. For each position, the Spearman correlation between the score and down-regulation on the array was determined and its *P* value plotted.

(C) Preferred offsets between paired regions of miRNA and mRNA. Analysis of (B) was repeated using a 4mer fixed at miRNA nucleotides 13-16, and the correlation between score and down-regulation was evaluated for alternative pairing offsets. A positive offset shifts the miRNA 3' pairing segment to the right, relative to the message.

(D) Preferential positions of conserved pairing. Human 3'UTRs were scanned for sites with at least a 6mer seed match to miRNAs of Table S1. For each contiguous 4mer beginning at nucleotide 9 of the miRNA, we searched for the complementary Watson-Crick 4mer directly opposite in the message, allowing for a ± 2-nt offset and excluding overlap into nucleotides 1-8. Those cases in which the seed and its supplemental 4mer were co-conserved were considered conserved instances of 3' pairing. This was compared to control chimeric miRNA sequences with the same 5' seed sequence but with the 3' end of a different miRNA, and signal/background was calculated. The analysis was repeated for 3mers and 5mers.

(E) 4mers conserved among paralogous human miRNAs. For each position the number of human miRNA families that have a perfectly conserved 4mer is indicated; families with only a single human paralog were excluded.

(F) Effectiveness of 7mer-m8 sites compensated with either good or poor 3' pairing. Distributions (left) and cumulative distributions (right) of changes in abundance of mRNAs following miRNA transfection were monitored with microarrays, and are displayed as in Figure 1B, including for reference the results of canonical 8mer sites. Sites with good pairing were those with scores ≥ 4 and sites with poor pairing were those with scores ≤ 2. Pairing score was determined as in (B) but using a fixed miRNA 4mer corresponding to nucleotides 13-16 and crediting contiguous pairing elsewhere with ½ point per pair. Sites with good pairing were significantly more effective than sites with poor pairing ($P = 0.007$, one-sided K-S test).

**Figure 3.** Substantial influence of local AU content.

(A) Preferred nucleotide composition in the vicinity of effective and conserved 8mer sites. For the site-efficacy analysis (green), sites associated with the greatest down-regulation in the expression arrays (top third of sites when ranking for down-regulation) were compared with those associated with least down-regulation (bottom third of sites). At each position, counting from the site, the fractional difference in AU composition within a ±5-nt window is plotted. The analysis was repeated for conservation (blue), comparing nucleotide composition flanking conserved sites versus nonconserved sites for the 11 miRNA families (Table S1).

(B) Weighting of the AU composition for the different types of sites. For each position within 30 nt upstream and downstream of the site, the presence of an A or a U increased the score for the site by an amount proportional to the height of the bar for that nucleotide. When moving away from the site, the weight (bar height) decreased with the inverse of the distance from the site. For example, the weight of the nucleotides downstream of the 8mer were 1/2, 1/3, 1/4, 1/5 ... that of the nucleotide upstream of the 8mer.

(C) Effectiveness of 8mer (left) and 7mer-m8 (right) sites with varying local AU content. Sites were separated into four quartiles according to rubrics of (B). For reference, the results of another canonical site is included (grey), otherwise as in Figure 1B. For both site types, the quartile with the highest local AU content was significantly more downregulated the quartile with the lowest local AU content ($P < 10^{-7}$ for 8mer sites, $P < 10^{-29}$ for 7mer-m1 sites).

(D) Site-depletion analysis implying greater efficacy of sites emerging with high local AU composition. To evaluate the efficacy of sites in different local AU contexts, sites were partitioned at the median into two equal-sized groups based on their local AU content, and the depletion analysis was performed for each tissue-miRNA pair. $P$ values indicate the extent of depletion for individual pairs. The boxes on the grid indicate those pairs in which the miRNA is expressed, which was where highly expressed messages were expected to be most depleted in sites to that miRNA.

**Figure 4.** Poor efficacy of sites within 15 nt of the stop codon.

(A) Distributions (left) and cumulative distributions (right) of changes to the levels of messages with one 8mer site following miRNA transfection, analyzed and displayed as in Figure 1B. Messages with a site in their 3'UTR or ORF were significantly more repressed than those with no site ($P < 10^{-126}$ and $< 10^{-16}$, respectively), whereas those in 5'UTRs were not ($P = 0.181$).

(B) Conservation of 7-8mer sites in the region near the stop codon. Plotted are the number of sites conserved above background per nucleotide in the ±5-nt window centered on the indicated mRNA position, with position 0 being the first nucleotide of the 3'UTR. A site was scored as within the window if the 5'-most nucleotide of the 7-8mer was within the window.

(C) MicroRNA-mediated repression of luciferase reporter genes fused to 3'UTR fragments containing miR-124 sites or mutant derivatives. After normalizing for transfection, luciferase activity from HeLa cells cotransfected with each reporter construct and its cognate miRNA (green and blue bars) was normalized to that from co-transfection of each reporter with its non-cognate miRNA (purple and orange bars). Error bars represent the fourth largest and smallest values among 18 replicates. *P* values (Wilcoxon rank-sum test) indicate whether repression from reporters containing the original ORF (blue) was significantly greater than that with a reporter with an extended ORF that stops 9 nt from the upstream site (green).

**Figure 5**. Poorer performance of sites near the middle of long 3'UTRs.

(A) Fraction of sites associated with repression for 8mers residing at different positions in 3'UTRs. UTRs of at least 1300 nt with single 8mer sites were split into 20 equally spaced bins based on the relative position of the site (distance from stop codon divided by the UTR length). For each bin, the point is plotted that corresponded to the mean site position and the fraction of messages downregulated at a threshold $P < 0.05$ on the microarray.

(B) Number of sites conserved above background for 8mers residing at different positions in 3'UTRs. UTRs of at least 1300 nt were divided into 20 equally spaced bins, and each bin, the point is plotted that corresponded to the mean site position and the number of sites conserved above the background.

(C) Site-depletion analysis implying greater efficacy of sites near the ends of UTRs. To evaluate the efficacy of sites in different UTR regions, total 3'UTR sequence for mouse UTRs of at least 1300 nt was divided into the middle half (right panel) and the two remaining terminal quarters (left), and the depletion analysis was performed for each group analogously to Figure 3D.

**Figure 6.** A quantitative model that considers site context to predict site efficacy.

(A) Increased efficacy of 7mer-m8 sites with more 3' pairing. Messages with a single site were scored as described in Figure 2F, and for each score, the mean downregulation on the microarray is plotted.

The regression line is the best least-squares fit to the full data and represents the relationship between score and downregulation on the array. The average $\log_2$ fold change (–0.161) is indicated (dashed line). The few sites with pairing score <1 or >5 were folded into the first and last bins, respectively.

(B) Increased efficacy of 7mer-m8 sites within higher local AU content. Messages with a single site were split into 14 equally sized bins based on scoring described in Figure 3B, where 1.0 equaled the maximum possible score. For each bin, the point is plotted that corresponded to the mean score and the mean downregulation on the microarray, and the line was fit as in panel (A).

(C) Decreased efficacy of 7mer-m8 sites further from the 3'UTR ends. Messages with a single site were split into 14 equally sized bins based on their distance from the closest UTR ends, and the points and regression line were plotted as in panel (B).

(D) Correspondence between the predicted and observed efficacy of single 7mer-m8 sites. Messages with a single site were split into 14 equally sized bins based on their the context score, calculated for each message by predicting the offsets from the mean for the three context determinants (panels A-C) and then adding these three contributions to the average $\log_2$ fold change for 7mer-m8 sites (Figure S6). Points correspond to the mean score and mean downregulation for each bin. The regression line is the best least-squares fit to the full data.

(E) Performance of the combined model when applied to a dataset that was not used to derive the model. Shown are the cumulative distributions of changes in mRNA levels following siRNA transfection (Jackson et al. 2003) after first predicting the top and bottom quartiles using context scores calculated as in (D) for messages with single 7mer-m8 sites to the siRNAs (Figure S5); otherwise as shown for Figure 1B.

(F) A unified model that considers the differential efficacy of the 7-8mer canonical sites and the influence of context determinants. Human messages with a single 7mer-A1, 7mer-m8, or 8mer site were split into two sets based on whether the site was conserved in orthologous UTRs of mouse, rat and dog. For each site, a context score was calculated using the regressions of panels A-C and analogous ones for the other two types of sites (Table S6). For conserved (orange) and nonconserved (blue) sets, sites were divided into 14 bins based on context score, and the mean score and mean repression was plotted for each bin. The regression lines were fit to the full data.

(G) Performance of different types of sites predicted to be in favorable or unfavorable contexts, as ranked using the context scores of (F). Shown are the minimum number of downregulated messages, as calculated using cumulative distributions like that of (E).

(H) MicroRNA-mediated repression of luciferase reporter genes fused to 3'UTR fragments containing a single 7mer-m8 miR-25 site located in favorable context. After normalizing to the transfection control, luciferase activity from HEK293 cells cotransfected with each reporter construct and its cognate miRNA (miR-25) was normalized to that from cotransfection of a reporter with a mutated miR-25 site and cognate miRNA. Similarly, luciferase acitivity from cotransfected wild-type reporter constructs and non-cognate miRNA (miR-196) was normalized to that from cotransfection of mutant reporter constructs and non-cognate miRNA. Plotted are the normalized values, with error bars representing the third largest and third smallest values among 12 replicates, with significant repression when comparing results for the cognate and noncognate miRNA indicated (*, $P < 0.01$; **, $P < 0.001$; Wilcoxon rank-sum test). Below each gene name is the context score and the three context contributions used to calculate, as in (D), the context score for the miR-25 site.

(I) MicroRNA-mediated repression of luciferase reporter genes fused to 3'UTR fragments containing a single 7mer-m8 miR-25 target site located in unfavorable context, otherwise as in (H).

**Figure 7.** Relevance of the model and its component features for endogenous miRNA targeting.

(A) Correspondence between context score and in vivo efficacy for endogenous miR-430-target interactions in zebrafish. Messages containing 6mer, 7mer-A1, 7mer-m8, or 8mer sites were analyzed as in Figure 6F, using published array data (Giraldez et al. 2006) monitoring the stabilization of endogenous messages after removing miR-430 and other miRNAs.

(B) Correspondence between context score and in vivo efficacy for miR-430-target interactions after restoring miR-430 to embryos missing all miRNAs (Giraldez et al. 2006), analyzed as in (A).

(C) Correspondence between context score and in vivo efficacy for endogenous miR-122-target interactions in mouse liver. Analyzed as in (A), using published array data (Krutzfeldt et al. 2005) monitoring the stabilization of endogenous messages after inhibiting endogenous miR-122.

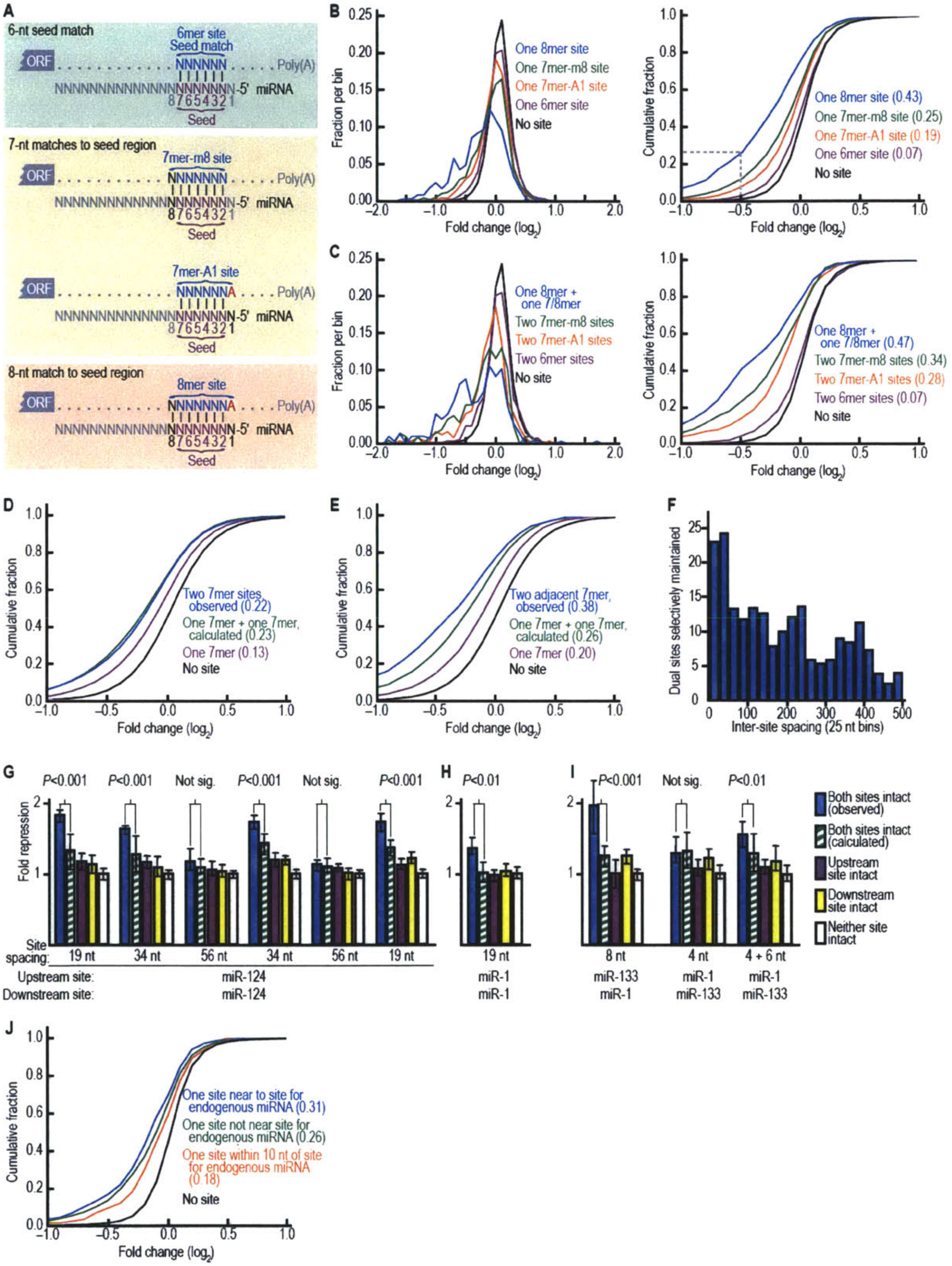
(D) Correspondence between context score and in vivo efficacy for endogenous miR-155-target interactions in mouse T cells. Analyzed as in (A), using published array data (Rodriguez et al. 2007) monitoring the stabilization of endogenous messages after deleting miR-155.

(E) Confirmation that the 3' pairing score correlates with endogenous targeting efficacy. Messages containing 6mer, 7mer-A1, 7mer-m8, or 8mer sites were analyzed as in Figure 6A, using published array data (Rodriguez et al. 2007) monitoring the stabilization of endogenous T-cell messages after deleting miR-155. To enable data for all four site types to be considered in aggregate, values were normalized by the average $\log_2$ fold change for each site type. Linear regression on all the data (red line), as in Figure 6A, yielded a significant correlation ($P < 10^{-3}$, Pearson correlation). As discussed for the siRNA transfection dataset (Figure S5A), this linear model appeared too simple because messages with sites scoring < 3 displayed no discernable trend (dashed blue line), whereas the handful of messages with high-scoring sites (<3% of sites have scores $\geq 4$) were associated with strong de-repression. Linear regression on only the data with scores $\geq 3$ (blue line) yielded a significant correlation ($P = 0.004$, Pearson correlation).

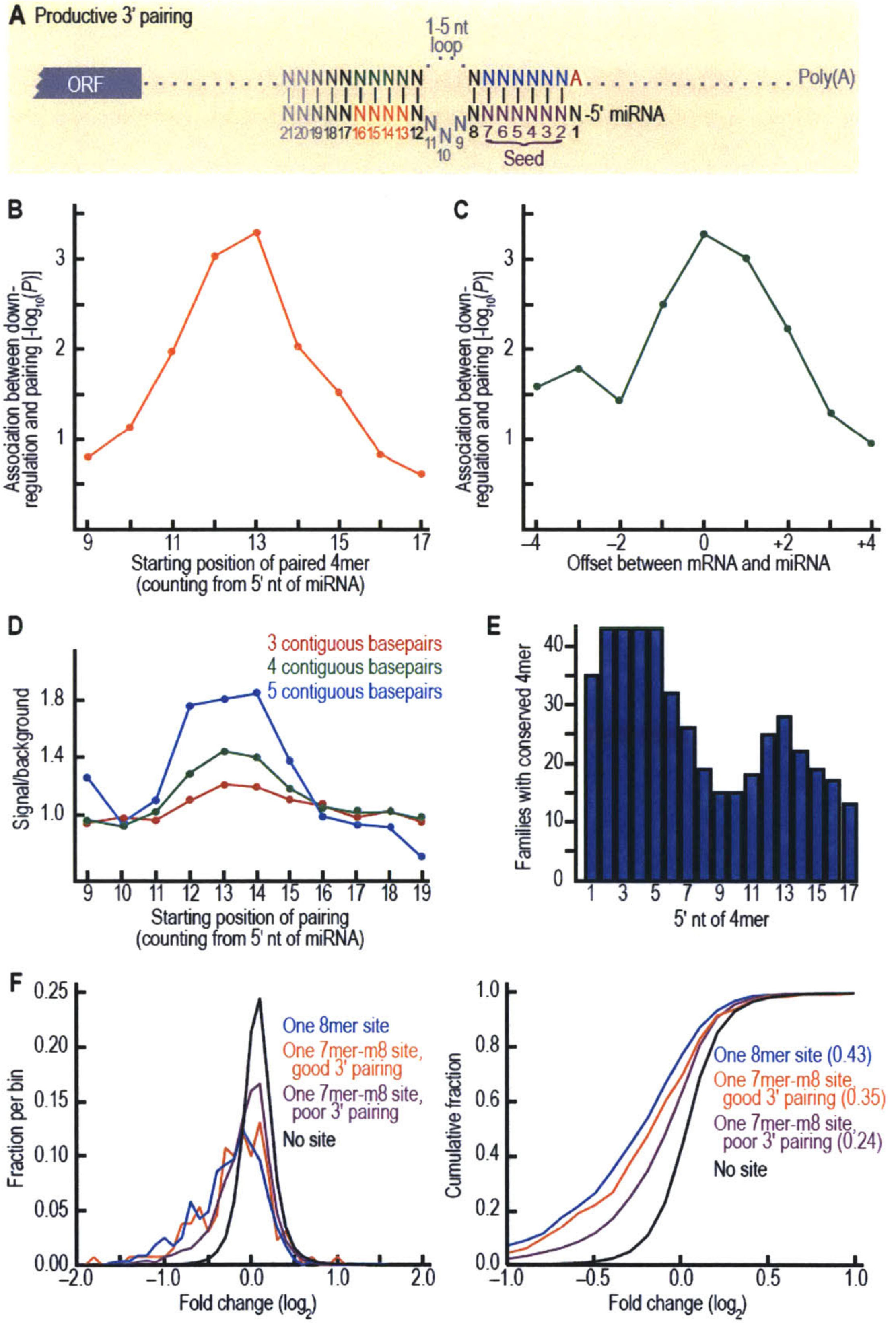
(F) Confirmation that local AU content correlates with endogenous targeting efficacy. Analyzed as in (A), except sites were scored for local AU content, as in Figure 6B ($P < 10^{-3}$, $\rho = 0.14$, Spearman correlation).

(G) Confirmation that site position correlates with endogenous targeting efficacy. Analyzed as in (A), except sites were scored for position, as in Figure 6C ($P < 10^{-3}$, $\rho = -0.14$, Spearman correlation).

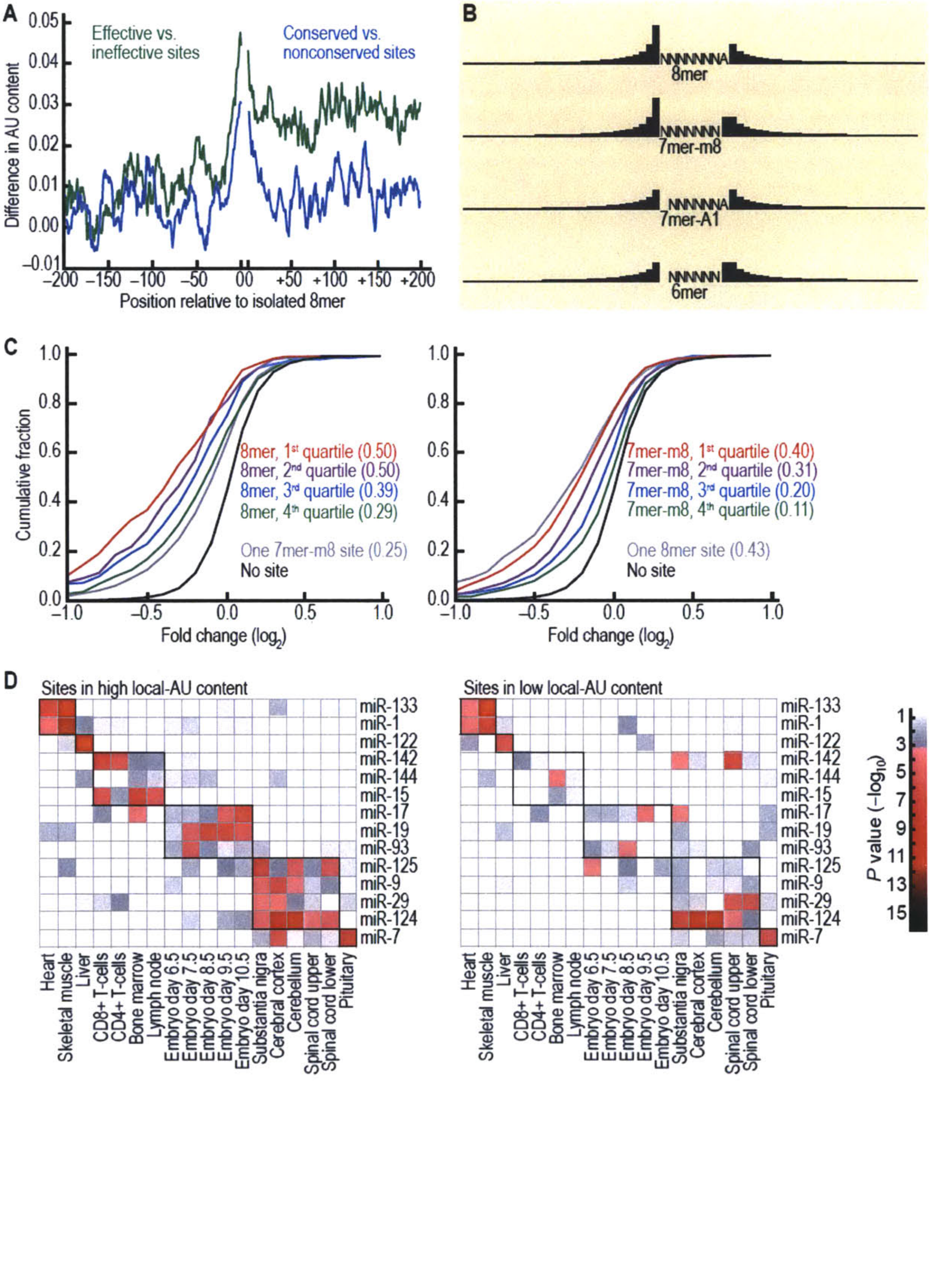
# Figure 1


A
6-nt seed match
6mer site
Seed match
ORF
NNNNNN
Poly(A)
NNNNNNNNNNNNNNNNNNNNN-5' miRNA
87654321
Seed
7-nt matches to seed region
7mer-m8 site
ORF
NNNNNNN
Poly(A)
NNNNNNNNNNNNNNNNNNNNN-5' miRNA
87654321
Seed
7mer-A1 site
ORF
NNNNNNA
Poly(A)
NNNNNNNNNNNNNNNNNNNNN-5' miRNA
87654321
Seed
8-nt match to seed region
8mer site
ORF
NNNNNNNA
Poly(A)
NNNNNNNNNNNNNNNNNNNNN-5' miRNA
87654321
Seed
B
Fraction per bin
0.25
0.20
0.15
0.10
0.05
0.00
−2.0
−1.0
0.0
1.0
2.0
Fold change (log2)
One 8mer site
One 7mer-m8 site
One 7mer-A1 site
One 6mer site
No site
Cumulative fraction
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Fold change (log2)
One 8mer site (0.43)
One 7mer-m8 site (0.25)
One 7mer-A1 site (0.19)
One 6mer site (0.07)
No site
C
Fraction per bin
0.25
0.20
0.15
0.10
0.05
0.00
−2.0
−1.0
0.0
1.0
2.0
Fold change (log2)
One 8mer + one 7/8mer
Two 7mer-m8 sites
Two 7mer-A1 sites
Two 6mer sites
No site
Cumulative fraction
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Fold change (log2)
One 8mer + one 7/8mer (0.47)
Two 7mer-m8 sites (0.34)
Two 7mer-A1 sites (0.28)
Two 6mer sites (0.07)
No site
D
Cumulative fraction
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Fold change (log2)
Two 7mer sites, observed (0.22)
One 7mer + one 7mer, calculated (0.23)
One 7mer (0.13)
No site
E
Cumulative fraction
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Fold change (log2)
Two adjacent 7mer, observed (0.38)
One 7mer + one 7mer, calculated (0.26)
One 7mer (0.20)
No site
F
Dual sites selectively maintained
25
20
15
10
5
0
0
100
200
300
400
500
Inter-site spacing (25 nt bins)
G
P<0.001
P<0.001
Not sig.
P<0.001
Not sig.
P<0.001
Fold repression
2
1
Site spacing:
19 nt
34 nt
56 nt
34 nt
56 nt
19 nt
Upstream site:
miR-124
Downstream site:
miR-124
H
P<0.01
2
1
19 nt
miR-1
miR-1
I
P<0.001
Not sig.
P<0.01
2
1
8 nt
4 nt
4 + 6 nt
miR-133
miR-1
miR-1
miR-1
miR-133
miR-133
Both sites intact (observed)
Both sites intact (calculated)
Upstream site intact
Downstream site intact
Neither site intact
J
Cumulative fraction
1.0
0.8
0.6
0.4
0.2
0.0
−1.0
−0.5
0.0
0.5
1.0
Fold change (log2)
One site near to site for endogenous miRNA (0.31)
One site not near site for endogenous miRNA (0.26)
One site within 10 nt of site for endogenous miRNA (0.18)
No site

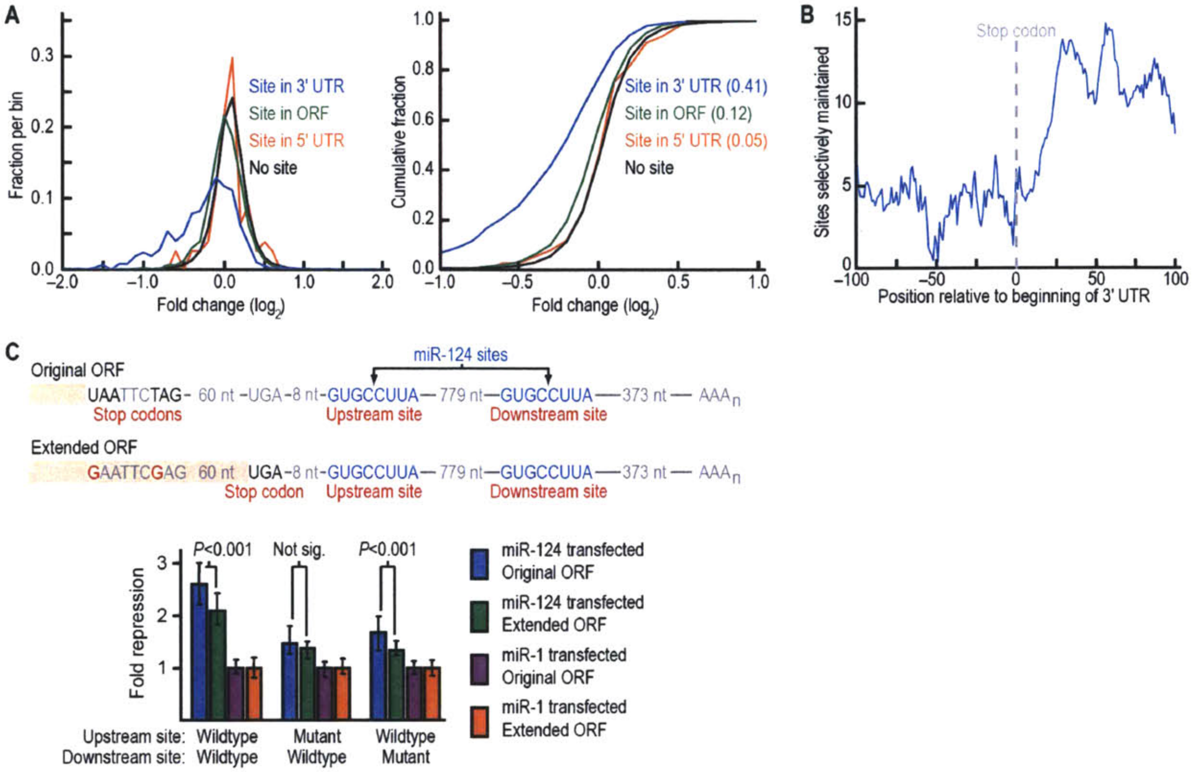
## Figure 2


A Productive 3' pairing
1-5 nt loop
ORF
NNNNNNNNNNN
NNNNNNNA
Poly(A)
NNNNNNNNNN
NNNNNNNNN -5' miRNA
21 20 19 18 17 16 15 14 13 12
N 11 N 10 N 9
8 7 6 5 4 3 2 1
Seed
B
Association between down-regulation and pairing [-log10(P)]
1
2
3
9
11
13
15
17
Starting position of paired 4mer (counting from 5' nt of miRNA)
C
Association between down-regulation and pairing [-log10(P)]
1
2
3
–4
–2
0
+2
+4
Offset between mRNA and miRNA
D
3 contiguous basepairs
4 contiguous basepairs
5 contiguous basepairs
Signal/background
1.0
1.4
1.8
9
10
11
12
13
14
15
16
17
18
19
Starting position of pairing (counting from 5' nt of miRNA)
E
Families with conserved 4mer
0
10
20
30
40
1
3
5
7
9
11
13
15
17
5' nt of 4mer
F
Fraction per bin
0.00
0.05
0.10
0.15
0.20
0.25
–2.0
–1.0
0.0
1.0
2.0
Fold change (log2)
One 8mer site
One 7mer-m8 site, good 3' pairing
One 7mer-m8 site, poor 3' pairing
No site
Cumulative fraction
0.0
0.2
0.4
0.6
0.8
1.0
–1.0
–0.5
0.0
0.5
1.0
Fold change (log2)
One 8mer site (0.43)
One 7mer-m8 site, good 3' pairing (0.35)
One 7mer-m8 site, poor 3' pairing (0.24)
No site

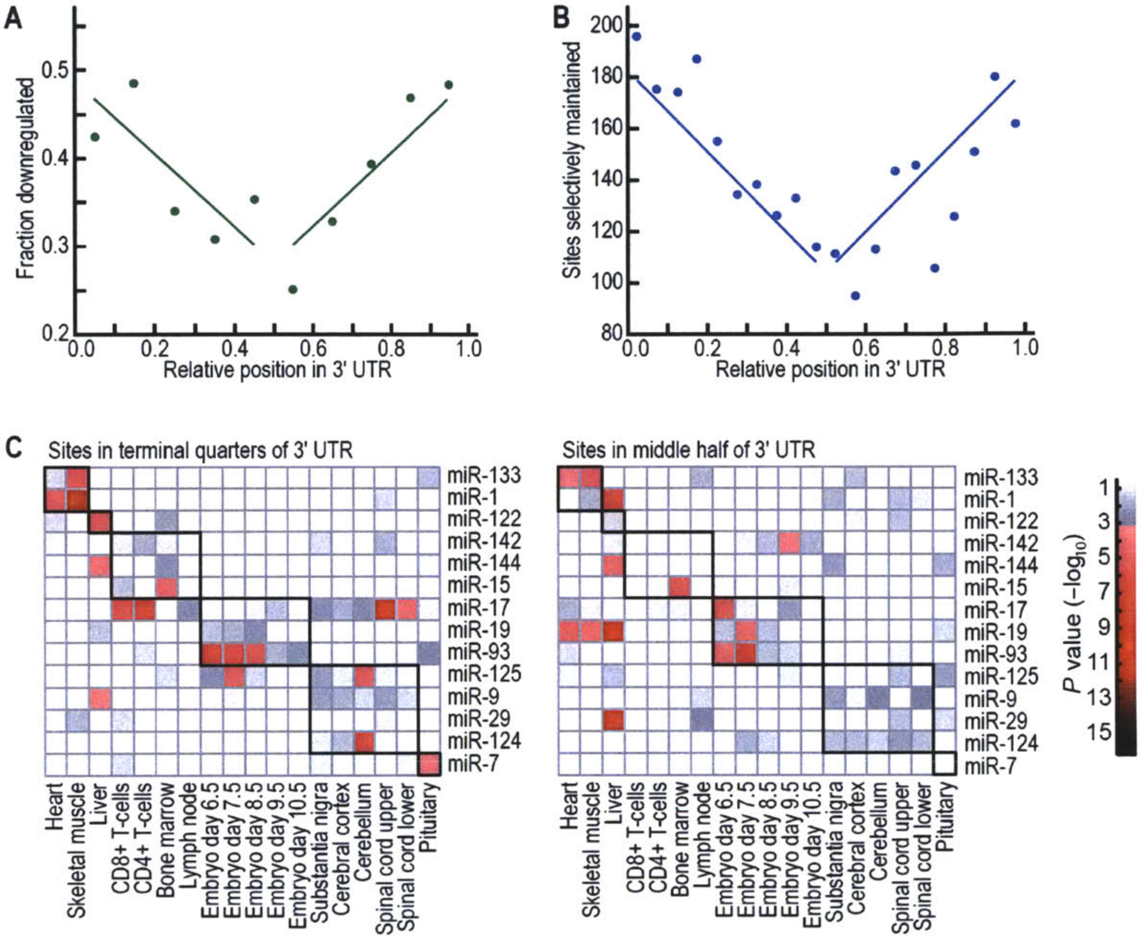
## Figure 3


A
0.05
0.04
0.03
0.02
0.01
0.00
−0.01
Difference in AU content
Effective vs. ineffective sites
Conserved vs. nonconserved sites
−200 −150 −100 −50 00 00 +50 +100 +150 +200
Position relative to isolated 8mer
B
NNNNNNA
8mer
NNNNNNN
7mer-m8
NNNNNNA
7mer-A1
NNNNNN
6mer
C
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative fraction
8mer, 1st quartile (0.50)
8mer, 2nd quartile (0.50)
8mer, 3rd quartile (0.39)
8mer, 4th quartile (0.29)
One 7mer-m8 site (0.25)
No site
−1.0 −0.5 0.0 0.5 1.0
Fold change (log2)
7mer-m8, 1st quartile (0.40)
7mer-m8, 2nd quartile (0.31)
7mer-m8, 3rd quartile (0.20)
7mer-m8, 4th quartile (0.11)
One 8mer site (0.43)
No site
Fold change (log2)
D
Sites in high local-AU content
Sites in low local-AU content
miR-133
miR-1
miR-122
miR-142
miR-144
miR-15
miR-17
miR-19
miR-93
miR-125
miR-9
miR-29
miR-124
miR-7
Heart
Skeletal muscle
Liver
CD8+ T-cells
CD4+ T-cells
Bone marrow
Lymph node
Embryo day 6.5
Embryo day 7.5
Embryo day 8.5
Embryo day 9.5
Embryo day 10.5
Substantia nigra
Cerebral cortex
Cerebellum
Spinal cord upper
Spinal cord lower
Pituitary
P value (−log10)
1
3
5
7
9
11
13
15

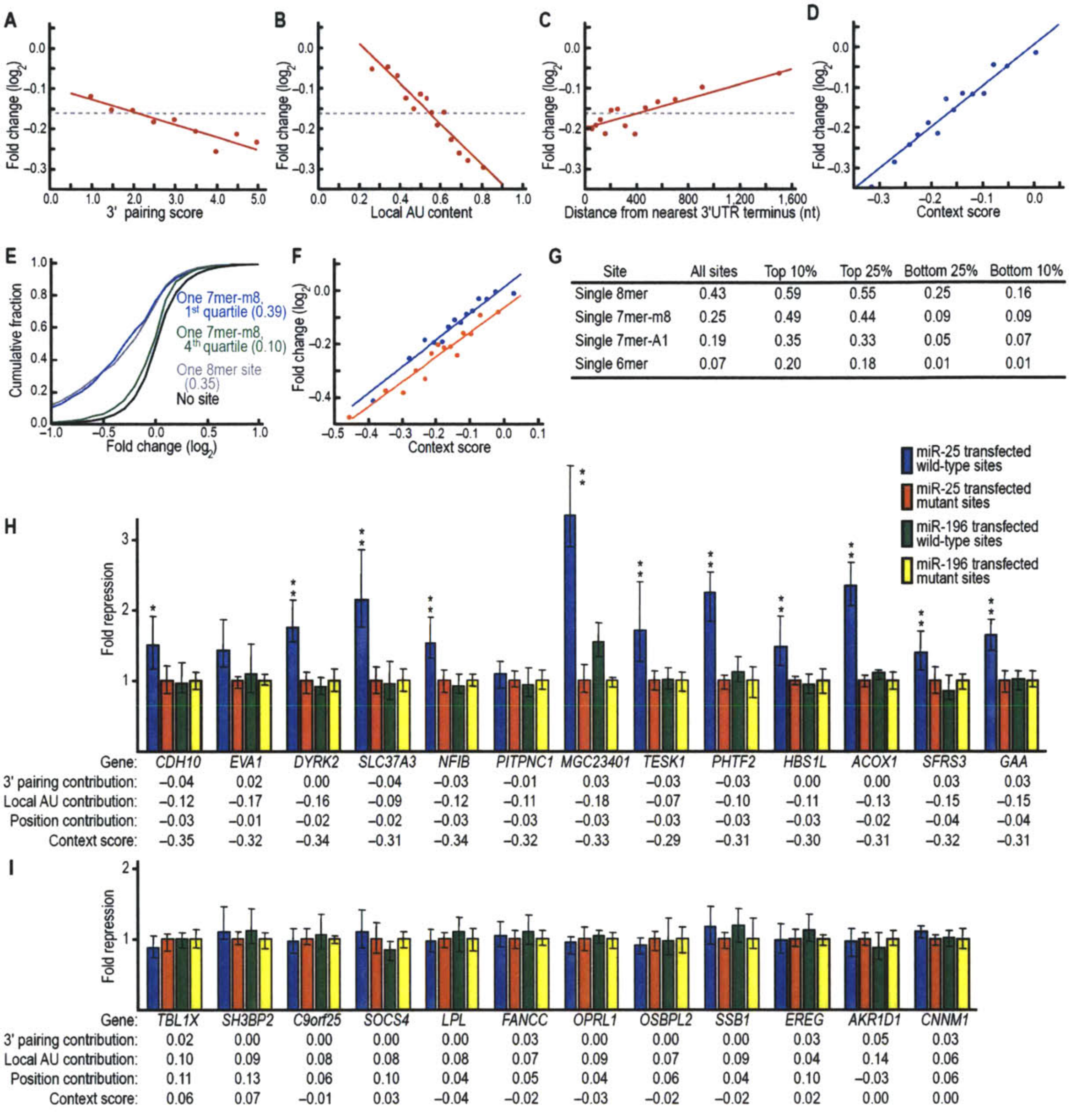
# Figure 4


A
Fraction per bin
0.0
0.1
0.2
0.3
−2.0
−1.0
0.0
1.0
2.0
Fold change (log$_2$)
Site in 3' UTR
Site in ORF
Site in 5' UTR
No site
Cumulative fraction
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5
1.0
Fold change (log$_2$)
Site in 3' UTR (0.41)
Site in ORF (0.12)
Site in 5' UTR (0.05)
No site
B
Sites selectively maintained
0
5
10
15
Stop codon
−100
−50
0
50
100
Position relative to beginning of 3' UTR
C
Original ORF
miR-124 sites
UAATTCTAG- 60 nt -UGA-8 nt-GUGCCUUA— 779 nt—GUGCCUUA—373 nt— AAA$_n$
Stop codons
Upstream site
Downstream site
Extended ORF
GAATTCGAG 60 nt UGA-8 nt-GUGCCUUA— 779 nt—GUGCCUUA—373 nt— AAA$_n$
Stop codon
Upstream site
Downstream site
Fold repression
1
2
3
$P$<0.001
Not sig.
$P$<0.001
Upstream site: Wildtype Mutant Wildtype
Downstream site: Wildtype Wildtype Mutant
miR-124 transfected Original ORF
miR-124 transfected Extended ORF
miR-1 transfected Original ORF
miR-1 transfected Extended ORF

## Figure 5


A
Fraction downregulated
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
Relative position in 3' UTR
B
Sites selectively maintained
80
100
120
140
160
180
200
0.0
0.2
0.4
0.6
0.8
1.0
Relative position in 3' UTR
C
Sites in terminal quarters of 3' UTR
Sites in middle half of 3' UTR
miR-133
miR-1
miR-122
miR-142
miR-144
miR-15
miR-17
miR-19
miR-93
miR-125
miR-9
miR-29
miR-124
miR-7
Heart
Skeletal muscle
Liver
CD8+ T-cells
CD4+ T-cells
Bone marrow
Lymph node
Embryo day 6.5
Embryo day 7.5
Embryo day 8.5
Embryo day 9.5
Embryo day 10.5
Substantia nigra
Cerebral cortex
Cerebellum
Spinal cord upper
Spinal cord lower
Pituitary
$P$ value ($-\log_{10}$)
1
3
5
7
9
11
13
15

# Figure 6



G

| Site | All sites | Top 10% | Top 25% | Bottom 25% | Bottom 10% |
|---|---|---|---|---|---|
| Single 8mer | 0.43 | 0.59 | 0.55 | 0.25 | 0.16 |
| Single 7mer-m8 | 0.25 | 0.49 | 0.44 | 0.09 | 0.09 |
| Single 7mer-A1 | 0.19 | 0.35 | 0.33 | 0.05 | 0.07 |
| Single 6mer | 0.07 | 0.20 | 0.18 | 0.01 | 0.01 |

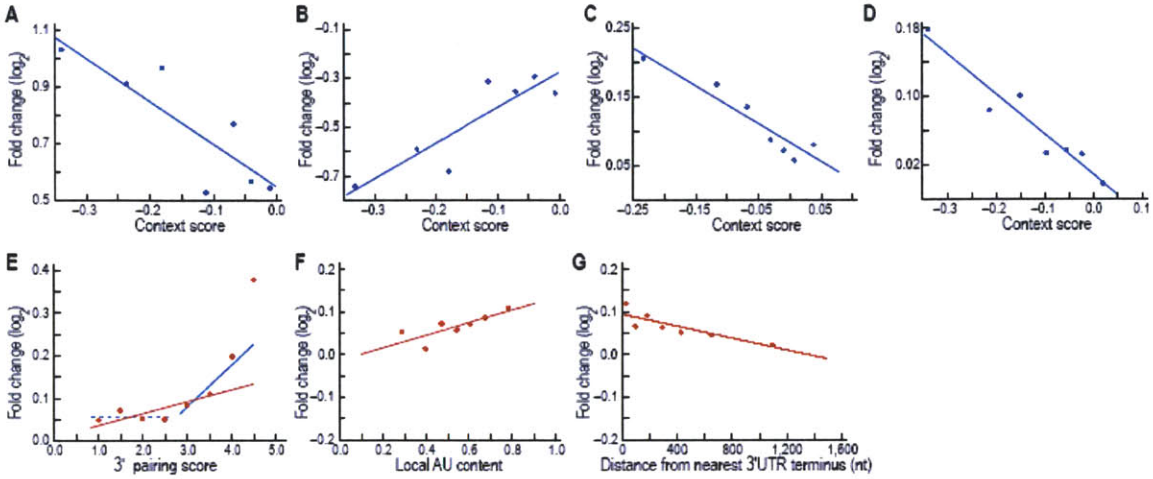
H

| Gene: | *CDH10* | *EVA1* | *DYRK2* | *SLC37A3* | *NFIB* | *PITPNC1* | *MGC23401* | *TESK1* | *PHTF2* | *HBS1L* | *ACOX1* | *SFRS3* | *GAA* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3' pairing contribution: | −0.04 | 0.02 | 0.00 | −0.04 | −0.03 | −0.01 | 0.03 | −0.03 | −0.03 | 0.00 | 0.00 | 0.03 | 0.03 |
| Local AU contribution: | −0.12 | −0.17 | −0.16 | −0.09 | −0.12 | −0.11 | −0.18 | −0.07 | −0.10 | −0.11 | −0.13 | −0.15 | −0.15 |
| Position contribution: | −0.03 | −0.01 | −0.02 | −0.02 | −0.03 | −0.03 | −0.03 | −0.03 | −0.03 | −0.03 | −0.02 | −0.04 | −0.04 |
| Context score: | −0.35 | −0.32 | −0.34 | −0.31 | −0.34 | −0.32 | −0.33 | −0.29 | −0.31 | −0.30 | −0.31 | −0.32 | −0.31 |

I

| Gene: | *TBL1X* | *SH3BP2* | *C9orf25* | *SOCS4* | *LPL* | *FANCC* | *OPRL1* | *OSBPL2* | *SSB1* | *EREG* | *AKR1D1* | *CNNM1* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 3' pairing contribution: | 0.02 | 0.00 | 0.00 | 0.00 | 0.00 | 0.03 | 0.00 | 0.00 | 0.00 | 0.03 | 0.05 | 0.03 |
| Local AU contribution: | 0.10 | 0.09 | 0.08 | 0.08 | 0.08 | 0.07 | 0.09 | 0.07 | 0.09 | 0.04 | 0.14 | 0.06 |
| Position contribution: | 0.11 | 0.13 | 0.06 | 0.10 | 0.04 | 0.05 | 0.04 | 0.06 | 0.04 | 0.10 | −0.03 | 0.06 |
| Context score: | 0.06 | 0.07 | −0.01 | 0.03 | −0.04 | −0.02 | −0.03 | −0.02 | −0.02 | 0.02 | 0.00 | 0.00 |

## Figure 7


A
B
C
D
E
F
G
Fold change (log2)
Context score
1.1
0.9
0.7
0.5
−0.3
−0.2
−0.1
0.0
−0.1
−0.3
−0.5
−0.7
0.25
0.15
0.05
−0.25
−0.15
−0.05
0.05
0.18
0.10
0.02
0.1
3' pairing score
0.4
0.3
0.2
0.1
0.0
1.0
2.0
3.0
4.0
5.0
Local AU content
0.2
0.1
0.0
−0.1
−0.2
0.4
0.6
0.8
1.0
Distance from nearest 3'UTR terminus (nt)
400
800
1,200
1,600

## Supplementary Data

### Signal for Site-Conservation Analyses

When examining conserved miRNA sites for favorable UTR contexts, we used the signal above background (sites selectively maintained) instead of the signal:background ratio. Signal above background indicates the selection of miRNA sites per UTR or segment of UTR, and thus it is the relevant measure of whether sites in a given context are more frequently subject to selection than sites in other contexts. In contrast, signal:background ratio (frequently also called the signal:noise ratio) is a measure of how well sites under selection can be distinguished from those conserved by chance. In UTR contexts that are enriched for conserved miRNA sites, there is also typically an increase in the conservation of other sequences that do not correspond to miRNA sites. In these circumstances, the signal above background improves because the increase in conserved miRNA sites outpaces the increase in conserved sequences that do not correspond to miRNA sites, but the signal:background ratio can remain constant or drop, despite the higher density of evolutionary selection for miRNA targeting. As a consequence, miRNA sites under selection can be paradoxically more difficult to predict with confidence when in favorable contexts because they tend to be associated with more background conservation than miRNA sites in poor contexts. For instance, Lewis et al. (2005) showed that within more highly conserved UTRs, the number of conserved miRNA sites above background increases, but the signal:background ratio drops because of the increase in background conservation.

Two phenomena can explain the association of greater background conservation with favorable context determinants. First, the selection acting on a miRNA site also acts to preserve the favorable context of the site, causing greater conservation in the vicinity of the site, although more limited than that for the site itself. This effect is compounded when a UTR has multiple conserved miRNA sites, and most UTRs with a conserved site to one miRNA family do have conserved sites to one or more additional miRNA families. Second, some UTR context determinants that encourage miRNA effectiveness likely generalize to RNA-protein interactions (e.g., to improve site accessibility), and a UTR regulated by miRNAs might also preferentially be regulated by proteins. As a result, these context determinants are associated with the conserved sequences that do not match miRNAs but can match the control sequences used to estimate the background conservation.

### Potential Correlations with Microarray Signal

We addressed the question of whether the level of mRNA expression, as measured by the intensity of the microarray signal, might correlate with and thereby confound interpretation of the specificity determinants. We examined the Spearman correlations for intensity vs. fold-change, considering each of the different canonical sites. The results were as follows:

8mer: rho = –0.093, $P = 0.0020$
7mer-m8: rho = –0.060, $P = 0.00098$
7mer-A1: rho = –0.037, $P = 0.042$
6mer: rho = –0.049, $P = 0.00013$

Thus downregulation was weakly correlated with higher intensity on the chip, presumably because the higher the intensity of the gene on the chip, the more likely that the gene is actually expressed in the cell, a prerequisite for downregulation. We minimized this effect by selecting for analysis only those genes that were expressed above median on the array, because genes expressed at levels lower than that would have a much smaller likelihood of going down. As a result, context determinants of our analysis were not correlated in a manor that would be a concern. For example, intensity and conservation are not significantly correlated. The other determinant that could have potentially been a concern was the local AU-effect. However, intensity and overall AU-richness were correlated in the wrong direction to explain the AU-effects (rho = –0.0454, indicating that AU-rich genes are expressed at slightly lower levels), and thus this was also not a concern.

**Table S1.** Vertebrate miRNA families. Listed are 73 human miRNAs, each representing a different family conserved in human, mouse, rat, dog and zebrafish.

| Family | 7mer-m8 site | Representative sequence | miRBase name | Accession |
|---|---|---|---|---|
| let-7 | CUACCUC | UGAGGUAGUAGGUUGUAUAGUU | hsa-let-7a | MIMAT0000062 |
| miR-1 | ACAUUCC | UGGAAUGUAAAGAAGUAUGUA | hsa-miR-1 | MIMAT0000416 |
| miR-7 | GUCUUCC | UGGAAGACUAGUGAUUUUGUUG | hsa-miR-7 | MIMAT0000252 |
| miR-9 | ACCAAAG | UCUUUGGUUAUCUAGCUGUAUGA | hsa-miR-9 | MIMAT0000441 |
| miR-10 | ACAGGGU | UACCCUGUAGAUCCGAAUUUGUG | hsa-miR-10a | MIMAT0000253 |
| miR-15 | UGCUGCU | UAGCAGCACAUAAUGGUUUGUG | hsa-miR-15a | MIMAT0000068 |
| miR-17 | GCACUUU | CAAAGUGCUUACAGUGCAGGUAGU | hsa-miR-17-5p | MIMAT0000070 |
| miR-18 | GCACCUU | UAAGGUGCAUCUAGUGCAGAUA | hsa-miR-18a | MIMAT0000072 |
| miR-19 | UUUGCAC | UGUGCAAAUCUAUGCAAAACUGA | hsa-miR-19a | MIMAT0000073 |
| miR-21 | AUAAGCU | UAGCUUAUCAGACUGAUGUUGA | hsa-miR-21 | MIMAT0000076 |
| miR-22 | GGCAGCU | AAGCUGCCAGUUGAAGAACUGU | hsa-miR-22 | MIMAT0000077 |
| miR-23 | AAUGUGA | AUCACAUUGCCAGGGAUUUCC | hsa-miR-23a | MIMAT0000078 |
| miR-24 | CUGAGCC | UGGCUCAGUUCAGCAGGAACAG | hsa-miR-24 | MIMAT0000080 |
| miR-26 | UACUUGA | UUCAAGUAAUUCAGGAUAGGUU | hsa-miR-26b | MIMAT0000083 |
| miR-27 | ACUGUGA | UUCACAGUGGCUAAGUUCUGC | hsa-miR-27b | MIMAT0000419 |
| miR-29 | UGGUGCU | UAGCACCAUCUGAAAUCGGUU | hsa-miR-29a | MIMAT0000086 |
| miR-30 | UGUUUAC | UGUAAACAUCCUCGACUGGAAG | hsa-miR-30a-5p | MIMAT0000087 |
| miR-31 | AUCUUGC | GGCAAGAUGCUGGCAUAGCUG | has-mir-31 | MIMAT0000089 |
| miR-33 | CAAUGCA | GUGCAUUGUAGUUGCAUUG | hsa-miR-33 | MIMAT0000091 |
| miR-34a | CACUGCC | UGGCAGUGUCUUAGCUGGUUGUU | hsa-miR-34a | MIMAT0000255 |
| miR-92 | GUGCAAU | UAUUGCACUUGUCCCGGCCUG | hsa-miR-92 | MIMAT0000092 |
| miR-93 | AGCACUU | AAAGUGCUGUUCGUGCAGGUAG | hsa-miR-93 | MIMAT0000093 |
| miR-96 | GUGCCAA | UUUGGCACUAGCACAUUUUUGC | hsa-miR-96 | MIMAT0000095 |
| miR-99 | UACGGGU | AACCCGUAGAUCCGAUCUUGUG | hsa-miR-99a | MIMAT0000097 |
| miR-101 | GUACUGU | UACAGUACUGUGAUAACUGAAG | hsa-miR-101 | MIMAT0000099 |
| miR-103 | AUGCUGC | AGCAGCAUUGUACAGGGCUAUGA | hsa-miR-103 | MIMAT0000101 |
| miR-122 | ACACUCC | UGGAGUGUGACAAUGGUGUUUGU | hsa-miR-122a | MIMAT0000421 |
| miR-124 | GUGCCUU | UAAGGCACGCGGUGAAUGCCA | hsa-miR-124a | MIMAT0000422 |
| miR-125 | CUCAGGG | UCCCUGAGACCCUUUAACCUGUG | hsa-miR-125a | MIMAT0000443 |
| miR-126 | CGGUACG | UCGUACCGUGAGUAAUAAUGC | hsa-miR-126 | MIMAT0000445 |
| miR-128 | CACUGUG | UCACAGUGAACCGGUCUCUUUU | hsa-miR-128a | MIMAT0000424 |
| miR-129 | GCAAAAA | CUUUUUGCGGUCUGGGCUUGC | hsa-miR-129 | MIMAT0000242 |
| miR-130 | UUGCACU | CAGUGCAAUGUUAAAAGGGCAU | hsa-miR-130a | MIMAT0000425 |
| miR-132 | GACUGUU | UAACAGUCUACAGCCAUGGUCG | hsa-miR-132 | MIMAT0000426 |
| miR-133 | GGGACCA | UUGGUCCCCUUCAACCAGCUGU | hsa-miR-133a | MIMAT0000427 |
| miR-135 | AAGCCAU | UAUGGCUUUUUAUUCCUAUGUGA | hsa-miR-135a | MIMAT0000428 |
| miR-137 | AAGCAAU | UAUUGCUUAAGAAUACGCGUAG | hsa-miR-137 | MIMAT0000429 |
| miR-138 | CACCAGC | AGCUGGUGUUGUGAAUC | hsa-miR-138 | MIMAT0000430 |
| miR-139 | ACUGUAG | UCUACAGUGCACGUGUCU | hsa-miR-139 | MIMAT0000250 |
| miR-140 | AAACCAC | AGUGGUUUUACCCUAUGGUAG | hsa-miR-140 | MIMAT0000431 |
| miR-141 | CAGUGUU | UAACACUGUCUGGUAAAGAUGG | hsa-miR-141 | MIMAT0000432 |
| miR-142 | ACACUAC | UGUAGUGUUUCCUACUUUAUGGA | hsa-miR-142-3p | MIMAT0000434 |
| miR-143 | UCAUCUC | UGAGAUGAAGCACUGUAGCUCA | hsa-miR-143 | MIMAT0000435 |
| miR-144 | AUACUGU | UACAGUAUAGAUGAUGUACUAG | hsa-miR-144 | MIMAT0000436 |
| miR-145 | AACUGGA | GUCCAGUUUUCCCAGGAAUCCCUU | hsa-miR-145 | MIMAT0000437 |
| miR-146 | AGUUCUC | UGAGAACUGAAUUCCAUGGGUU | hsa-miR-146a | MIMAT0000449 |
| miR-148 | UGCACUG | UCAGUGCACUACAGAACUUUGU | hsa-miR-148a | MIMAT0000243 |
| miR-150 | UUGGGAG | UCUCCCAACCCUUGUACCAGUG | hsa-miR-150 | MIMAT0000451 |
| miR-153 | CUAUGCA | UUGCAUAGUCACAAAAGUGA | hsa-miR-153 | MIMAT0000439 |
| miR-181 | UGAAUGU | AACAUUCAACGCUGUCGGUGAGU | hsa-miR-181a | MIMAT0000256 |

| | | | | |
|---|---|---|---|---|
| miR-182 | UUGCCAA | UUUGGCAAUGGUAGAACUCACA | hsa-miR-182 | MIMAT0000259 |
| miR-183 | GUGCCAU | UAUGGCACUGGUAGAAUUCACUG | hsa-miR-183 | MIMAT0000261 |
| miR-184 | UCCGUCC | UGGACGGAGAACUGAUAAGGGU | hsa-miR-184 | MIMAT0000454 |
| miR-187 | AGACACG | UCGUGUCUUGUGUUGCAGCCG | hsa-miR-187 | MIMAT0000262 |
| miR-192 | UAGGUCA | CUGACCUAUGAAUUGACAGCC | hsa-miR-192 | MIMAT0000222 |
| miR-193 | GGCCAGU | AACUGGCCUACAAAGUCCCAG | hsa-miR-193a | MIMAT0000459 |
| miR-194 | CUGUUAC | UGUAACAGCAACUCCAUGUGGA | hsa-miR-194 | MIMAT0000460 |
| miR-196 | ACUACCU | UAGGUAGUUUCAUGUUGUUGG | hsa-miR-196a | MIMAT0000226 |
| miR-199 | ACACUGG | CCCAGUGUUCAGACUACCUGUUC | hsa-miR-199a | MIMAT0000231 |
| miR-200b | CAGUAUU | UAAUACUGCCUGGUAAUGAUGAC | hsa-miR-200b | MIMAT0000318 |
| miR-203 | CAUUUCA | GUGAAAUGUUUAGGACCACUAG | hsa-miR-203 | MIMAT0000264 |
| miR-204 | AAAGGGA | UUCCCUUUGUCAUCCUAUGCCU | hsa-miR-204 | MIMAT0000265 |
| miR-205 | AUGAAGG | UCCUUCAUUCCACCGGAGUCUG | hsa-miR-205 | MIMAT0000266 |
| miR-210 | ACGCACA | CUGUGCGUGUGACAGCGGCUGA | hsa-miR-210 | MIMAT0000267 |
| miR-214 | CCUGCUG | ACAGCAGGCACAGACAGGCAG | hsa-miR-214 | MIMAT0000271 |
| miR-216 | UGAGAUU | UAAUCUCAGCUGGCAACUGUG | hsa-miR-216 | MIMAT0000273 |
| miR-217 | AUGCAGU | UACUGCAUCAGGAACUGAUUGGAU | hsa-miR-217 | MIMAT0000274 |
| miR-218 | AAGCACA | UUGUGCUUGAUCUAACCAUGU | hsa-miR-218 | MIMAT0000275 |
| miR-219 | GACAAUC | UGAUUGUCCAAACGCAAUUCU | hsa-miR-219 | MIMAT0000276 |
| miR-221 | AUGUAGC | AGCUACAUUGUCUGCUGGGUUUC | hsa-miR-221 | MIMAT0000278 |
| miR-223 | AACUGAC | UGUCAGUUUGUCAAAUACCCC | hsa-miR-223 | MIMAT0000280 |
| miR-338 | AUGCUGG | UCCAGCAUCAGUGAUUUUGUUGA | hsa-miR-338 | MIMAT0000763 |
| miR-375 | CGAACAA | UUUGUUCGUUCGGCUCGCGUGA | hsa-miR-375 | MIMAT0000728 |

**Table S2.** Synthetic miRNA duplexes used in microarray transfection experiments.

miR-1
```
  5'-UGGAAUGUAAAGAAGUAUGUA-3'
     | |||||||||||||||||
3'-AUAACUUACAUUUCUUCAUAC-5'
```

miR-7
```
  5'-UGGAAGACUAGUGAUUUUGUU-3'
     |••||||||||||||||||
3'-UAAUUUUCUGAUCACUAAAAC-5'
```

miR-9
```
  5'-UCUUUGGUUAUCUAGCUGUAUGA-3'
     | |||||||||||||||||||
3'-UAACAAACCAAUAGAUCGACAUA-5'
```

miR-122
```
  5'-UGGAGUGUGACAAUGGUGUUUGU-3'
     |••||||||||||||||||||
3'-UAAUUUCACACUGUUACCACAAA-5'
```

miR-124
```
  5'-UAAGGCACGCGGUGAAUGCCA-3'
     |||||||||||||||||||
3'-UAAUUCCGUGCGCCACUUACG-5'
```

miR-128
```
  5'-UCACAGUGAACCGGUCUCUUUU-3'
     | ||||||||||||||||||
3'-UAACUGUCACUUGGCCAGAGAA -5'
```

miR-132
```
  5'-UAACAGUCUACAGCCAUGGUCG-3'
     ||||||||||||||||||||
3'-UAAUUGUCAGAUGUCGGUACCA-5'
```

miR-133
```
  5'-UUGGUCCCCUUCAACCAGCUGU-3'
     ||••||||||||||||||||
3'-UAAAUUAGGGGAAGUUGGUCGA-5'
```

miR-142
```
  5'-UGUAGUGUUUCCUACUUUAUGGA-3'
     |•|||||||||||||||||||
3'-UAAUAUCACAAAGGAUGAAAUAC-5'
```

miR-148
```
  5'-UCAGUGCAUCACAGAACUUUGU-3'
     | ||||||||||||||||||
3'-UAACUCACGUAGUGUCUUGAAA-5'
```

miR-181
```
  5'-AACAUUCAACGCUGUCGGUGAGU-3'
     | |||||||||||||||||||
3'-UAUAGUAAGUUGCGACAGCCACU-5'
```

**Table S3.** Synthetic miRNA duplexes used in reporter experiments. miR-1, miR-124 and miR-133 duplexes are identical to those used in microarray transfection experiments (Table S2).

miR-25

```
  5'-CAUUGCACUUGUCUCGGUCUGA-3'
     | ||||||||||||||||||
3'-UAGAAACGUGAACAGAGCCAGA-5'
```

miR-196

```
  5'-UAGGUAGUUUCAUGUUGUUGGG-3'
     ||•|||||||||||||||||
3'-GAAUUCAUCAAAGUACAACAAC-
5'
```

**Table S4.** Sequences of UTR fragments assayed in Figures 1G-I and 4C.
Reporter plasmids encoded Renilla luciferase and were constructed in pIS2. pIS2 was derived from pRL-SV40 (Promega) by insertion of a multiple cloning site (shown below) within the region corresponding to the 3' UTR of the luciferase mRNA. Listed are sequences of UTR fragments, annotating their GenBank accession number, the restriction sites used in cloning (5' site – 3' site), the reporter plasmid name (in brackets), and the miRNA target sites (underlined). To disrupt miR-124 sites, the seed match TGCCTT was changed to TcggTT, except for the sites disrupted in the rightmost set of Figure 1G, which were instead changed to TGCCaa. To disrupt miR-1 sites, the seed match CATTCC was changed to CtgaCC. To disrupt miR-133 site, the seed match GGACCA was changed to GctgCA.

Please refer to the publication for the complete plasmid sequences.

**Table S5.** Sequences of UTR fragments assayed in Figure 6.
Reporter plasmids encoded Renilla luciferase and were constructed in pIS1. pIS1 was derived from pRL-TK (Promega) by insertion of a multiple cloning site (shown below) within the region corresponding to the 3' UTR of the luciferase mRNA. Listed are sequences of UTR fragments, annotating their name, the restriction sites used in cloning (5' site – 3' site), the reporter plasmid name (in brackets), and the miRNA target sites (underlined). To disrupt miR-25 sites, the seed match TGCAAT was changed to TcgtAT.

Please refer to the publication for the complete plasmid sequences.

**Table S6.** Context score parameters for different miRNA target sites, with Pearson correlation coefficient (ρ) and corresponding *P* values indicating the confidence in a non-zero slope.

**8mer**, mean value -0.31

| *Determinant* | *Slope* | *y-intercept* | ρ | P *value* |
|---|---|---|---|---|
| 3' pairing | –0.0041 | –0.299 | –0.01 | 0.80 |
| Local AU | –0.64 | 0.055 | –0.23 | $<10^{-13}$ |
| Position | 0.000172 | –0.38 | 0.18 | $<10^{-8}$ |

**7mer-m8**, mean value -0.161

| *Determinant* | *Slope* | *y-intercept* | ρ | P *value* |
|---|---|---|---|---|
| 3' pairing | –0.031 | –0.094 | –0.07 | $<10^{-3}$ |
| Local AU | –0.50 | 0.108 | –0.21 | $<10^{-15}$ |
| Position | 0.000091 | –0.198 | 0.11 | $<10^{-8}$ |

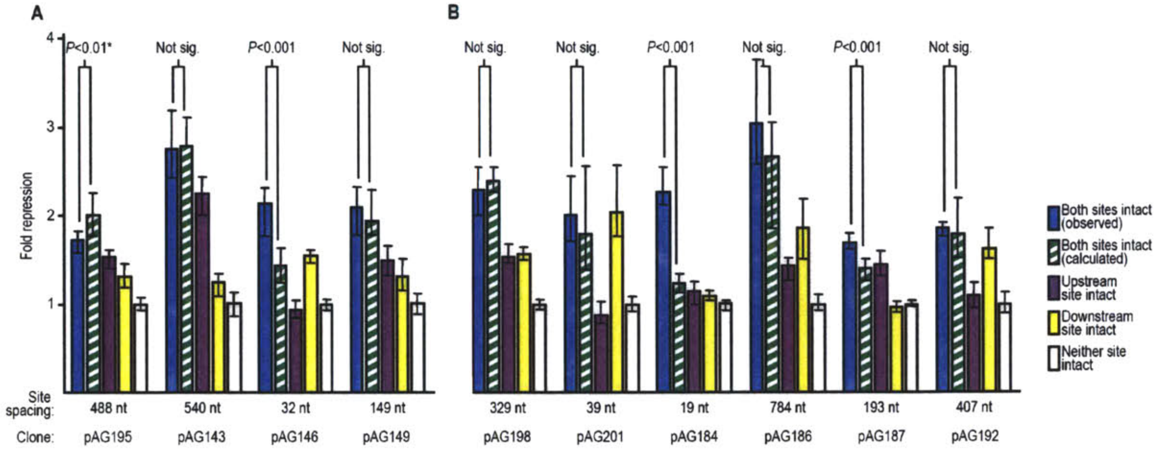
**7mer-A1**, mean value -0.099

| *Determinant* | *Slope* | *y-intercept* | ρ | P *value* |
|---|---|---|---|---|
| 3' pairing | –0.0211 | –0.0211 | –0.06 | $<10^{-2}$ |
| Local AU | –0.42 | 0.137 | –0.20 | $<10^{-15}$ |
| Position | 0.000072 | –0.131 | 0.10 | $<10^{-8}$ |

**6mer**, mean value -0.015

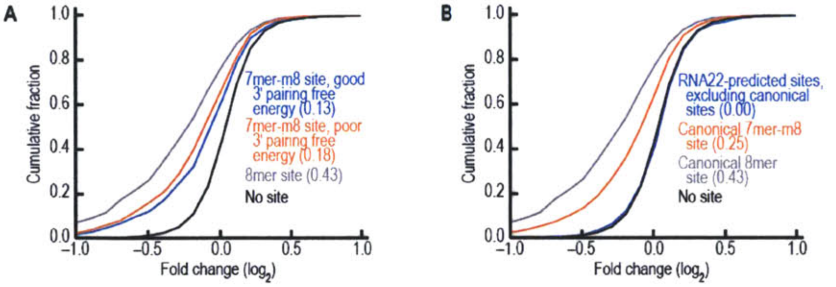
| *Determinant* | *Slope* | *y-intercept* | ρ | P *value* |
|---|---|---|---|---|
| 3' pairing | –0.00278 | –0.0091 | –0.01 | 0.52 |
| Local AU | –0.241 | 0.115 | –0.14 | $<10^{-15}$ |
| Position | 0.000049 | –0.033 | 0.07 | $<10^{-7}$ |

## Supplemental Figure S1



**Supplemental Figure 1.** MicroRNA-mediated repression of *Renilla* luciferase reporter genes fused to 3' UTR fragments containing two miR-1 (A) or miR-124 (B) sites, or mutant derivatives. After normalizing to the firefly transfection control, luciferase activity from HeLa cells cotransfected with each reporter construct and its cognate miRNA (A: miR-1; B: miR-124) was normalized to that from cotransfection of each reporter with a non-cognate miRNA (A: miR-124; B: miR-1). Plotted are the normalized values, with error bars representing the third largest and third smallest values among 12 replicates. *P* values (Wilcoxon rank-sum test) indicate whether repression from a reporter containing both sites (blue) was significantly greater than that expected from additive effects (green). For this additive model, repression expected from a reporter with two sites was the product of repression observed from otherwise identical reporters containing single intact sites (purple and yellow). The two UTR fragments containing the most closely spaced sites (pAG146 and pAG184) manifested significantly cooperative repression. pAG184 was selected for further study (Figure 1G). For six of the eight fragments containing less closely spaced sites, repression from constructs containing both sites (blue) did not statistically differ from that expected for additive repression (green). One exception was pAG187, which manifested significant cooperativity (pAG187), although the magnitude of the affect was relatively low. The other exception was pAG195, which manifested apparently negative cooperativity, although this result was of more borderline statistical significance. Parental clones are as described (Farh et al., 2005).

## Supplemental Figure S2

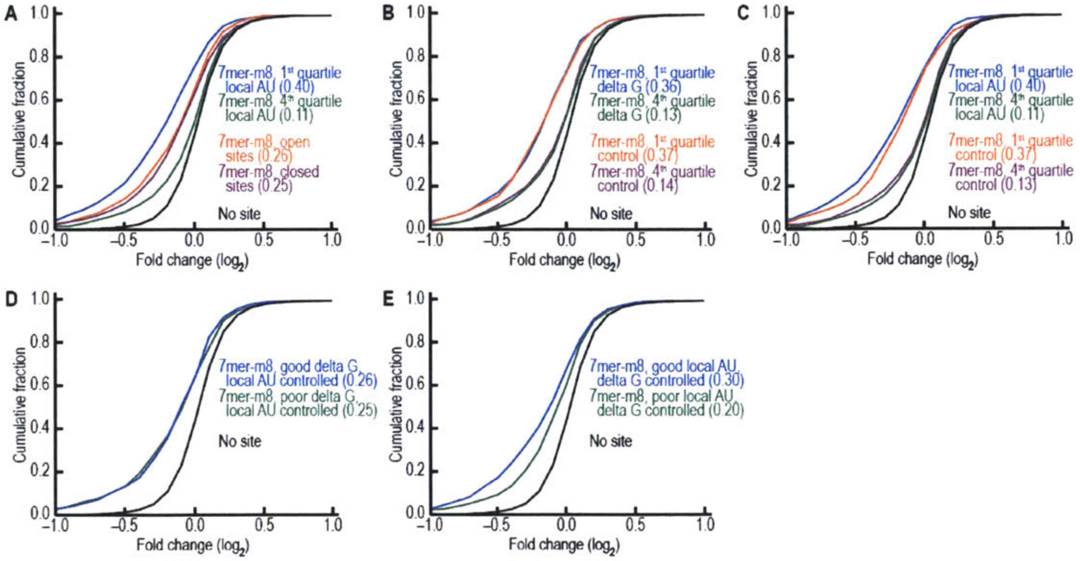


**Supplemental Figure 2.** Evaluation of previous algorithms designed to score productive pairing involving the 3' region of the miRNA.

(A) Performance of sites ranked using an energy-based rubric for predicting and scoring 3' supplementary pairing resembling those rubrics developed previously (Lewis et al., 2003; John et al., 2004; Krek et al., 2005). For each transfected miRNA, messages with single 7mer-m8 sites were identified together with the 20 UTR nucleotides upstream of each site, and folded with the miRNA using RNAhybrid (Rehmsmeier et al., 2004), without permitting the pairing of UTR nucleotides with each other or the pairing of miRNA nucleotides with each other. For each of the 11 miRNAs, messages were partitioned into four quartiles based on pairing free energy of their sites. Shown are the aggregate results for all 11 miRNAs for the quartile with the lowest deltaG values (most stable predicted pairing) and highest deltaG values. The quartile with the lowest deltaG values performed significantly worse than that with the highest deltaG values ($P = 0.0056$, two-sided K-S test). (Some previous methods that consider pairing to the 3' region of the miRNA normalize the predicted pairing free energy for each site to that of the fully paired miRNA. The results of this figure would have been the same if we had done similar normalization because for each of the 11 miRNAs considered, messages with sites were split evenly into the four equal quartiles before the sites for different miRNAs were combined.)

(B) Performance of sites with extensive Watson-Crick and G:U pairing along the length of the miRNA, but without canonical seed pairing, as proposed by Miranda et al. (2006). MicroRNA target predictions for the transfected miRNAs were obtained from the RNA22 web site. These predictions had been assigned to miRNAs using pairing parameters (G = 0, M = 14 and E = –25 Kcal/mol), which were more stringent than those used in the published work (Miranda et al., 2006), but we used them because the larger set described in the published work was not disclosed in November 2006. Predictions for miR-124 were not considered because these were for a miR-124 variant that was offset by one nucleotide from the miR-124 that we transfected. The predictions for the remaining 10 transfected miRNAs were filtered to remove those with canonical seed sites (1161 of the 1720 predictions for these 10 miRNAs had at least a 6mer site in their 3' UTR). The remainder consisted of sites with imperfect seed pairing and extensive 3' pairing. When considering their downregulation on the microarray, such sites performed no better than all the other genes with no seed site ($P = 0.096$, one-sided K-S test). The performance of canoncial 7mer-m8 and 8mer sites (Figure 1B) is shown for comparison. This result suggests that the many thousands of noncanonical predictions proposed by Miranda et al. (2006) did not include any more functional predictions than expected by chance.

## Supplemental Figure S3



**Supplemental Figure 3.** Evaluation of existing algorithms designed to score site accessibility by predicting UTR secondary structure.

(A) Performance of sites with three or more nucleotides in open structure, as determined by the algorithm of Robins et al. (2005). Messages with a single 7mer-m8 site were split with respect to open structure (defined as at least 3 nt continuous unpaired within site). Shown are cumulative distributions of mRNA changes on the array for messages in the highest and lowest quartiles. A slight difference was observed between open and closed sites, but this difference was not significant ($P = 0.066$, 1-sided K-S test). Performance of sites in the highest and lowest quartiles with respect to local AU content (Figure 3) is shown for comparison.

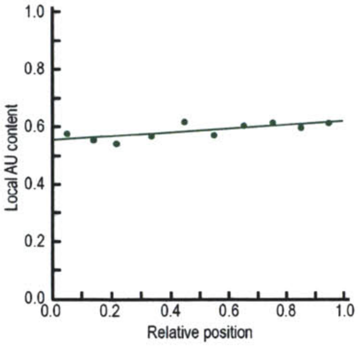
(B) Performance of sites flanked by low or high predicted stability of UTR regions flanking the seed site, as determined by the algorithm of Zhao et al. (2005, 2007). Messages with a single 7mer-m8 site were split into quartiles with respect to average predicted stability of 70-nt segments immediately flanking both sides of the site. Shown are cumulative distributions of mRNA changes on the array for messages in the highest and lowest quartiles. To control for the effects of global nucleotide composition, 10 random 70-nt regions in the same UTR were folded and messages were again split into quartiles with respect to average predicted stability. Among sites that were significantly downregulated on the array ($P < 0.01$), there was no significant difference in stability when comparing the 70-nt UTR regions flanking the site and randomly selected 70-nt regions from the same UTR ($P = 0.120$, Wilcoxon rank-sum test). Thus, the success of this algorithm in predicting down regulated messages could be attributed to a correlation between predicted stability near the site and more global properties of the UTR.

(C) Performance of sites in the highest and lowest quartiles with respect to local AU content, as scored in Figure 3B. To control for the effects of global nucleotide composition, 10 random positions in the same UTR were evaluated for local AU content. For sites that were significantly downregulated on the array ($P < 0.01$), the difference in AU content was significant when comparing the region immediately adjacent to the authentic site and randomly selected regions from the same UTR ($P = 0.0069$, Wilcoxon rank-sum test.) Although local AU content correlated with AU content throughout the remainder of the UTR, not all of the success of the algorithm could be attributed to a correlation with the more global property.

(D) Performance of sites flanked by low or high predicted stability of UTR regions flanking the seed site, after controlling for local and global AU content. To measure the residual effect of predicted secondary structure after controlling for local and global AU content, 1000 pairs of matched sites were picked randomly such that 1) their local AU content was within 5 percentiles of each other; 2) their global AU content was within 5 percentiles of each other; and 3) their average predicted folding stability in the regions flanking the sites, as determined by the algorithm of Zhao et al. (2005), was at least 30 percentiles apart. Repression was not significantly greater among sites with weaker predicted folding energy when AU content was held constant ($P = 0.088$, 1-sided K-S test). Thus, the specific predicted secondary structures scored by the algorithm of Zhao et al. were not informative after controlling for local AU content.

(E) Performance of sites with different local AU contents, after controlling for global AU content and predicted folding stability in the regions flanking the sites. 1000 pairs of matched sites were picked randomly such that 1) their global AU content was within 5 percentiles of each other; 2) their average predicted folding stability in the regions flanking the sites, as determined by the algorithm of Zhao et al. (2005), was within 5 percentiles of each other, and 3) their local AU content was at least 30 percentile apart. Repression was significantly greater among sites with high local AU content ($P < 10^{-7}$, 1-sided K-S test.). The Spearman correlation between predicted stability and global AU content was 0.6665, while the Spearman correlation between local AU content and global AU content was 0.6253. Because of the high co-correlation of these variables with global nucleotide content, random sampling of sequences in the same UTR was necessary to control for global AU and reveal underlying context determinants in the local neighborhood of effective sites.

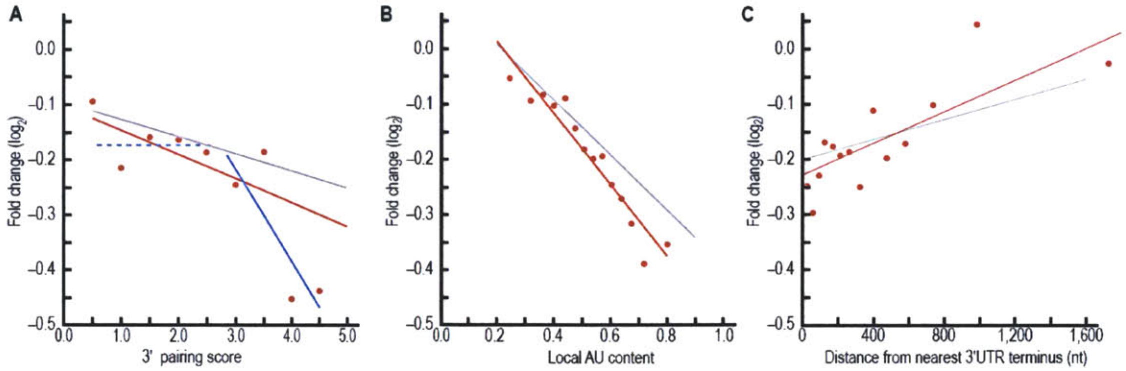
## Supplemental Figure S4



**Supplemental Figure 4.** Evaluation of correlations between determinants used to generate target site context scores. We observed a modest but significant (*P* = 0.0015) Spearman correlation between local AU content and position within 3' UTR (evaluated as in Figure 3B and 5, respectively), indicating a slight increase in local AU content as the local window traversed from the 5' to the 3' termini of the 3' UTR. This correlation could not account for the contribution of position to final context score, because sites located close to either terminus of the 3' UTR were most effective, rather than sites near the 3' terminus of the 3' UTR.

We evaluated whether scores for the three determinants used to generate the final context score were significantly correlated by Spearman correlation, and found that they were not: position score and 3' pairing score, *P* = 0.092; position score and local AU score, *P* = 0.51; local AU score and 3' pairing score, *P* = 0.40.

## Supplemental Figure S5



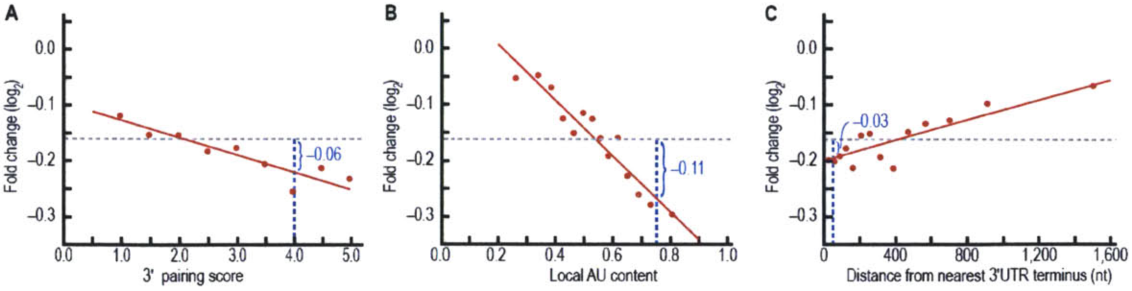
D

| siRNA | 7mer-m8 site | Sequence (sense strand) | Effective Strand |
|---|---|---|---|
| MAPK14-1 | CUGCGGU | CCUACAGAGAACUGCGGUU-dTdT | - |
| MAPK14-2 | AAUCACA | AUGUGAUUGGUCUGUUGGA-dTdT | + |
| MAPK14-3 | CUAAAGU | UUCUCCGAGGUCUAAAGUA-dTdT | - |
| MAPK14-4 | GACCUAA | UAAUUCACAGGGACCUAAA-dTdT | - |
| MAPK14-5 | UCCUUAU | CCAGUGGCCGAUCCUUAUG-dTdT | - |
| MAPK14-6 | AGUAGGC | UGCCUACUUUGCUCAGUAC-dTdT | + |
| MAPK14-7 | CUGAUGA | GUCAUCAG CUUUGUGCCAC-dTdT | + |
| MAPK14-8 | GGGAACU | GGCCUUUUCACGGGAACUC-dTdT | - |
| IGF1R-1 | UUACCGA | GCUCACGGUCAUUACCGAG-dTdT | - |
| IGF1R-2 | UCCUCAG | CCUGAGGAACAUUACUCGG-dTdT | + |
| IGF1R-3 | GGUCAGC | UGCUGACCUCUGUUACCUC-dTdT | + |
| IGF1R-4 | CCGUGUC | CGACACGGCCUGUGUAGCU-dTdT | + |
| IGF1R-5 | UGGCCGG | GAUGAUUCAGAUGGCCGGA-dTdT | - |
| IGF1R-6 | GCUGCAA | CUUGCAGCAACUGUGGGAC-dTdT | + |
| IGF1R-7 | CCGUGAG | CCUCACGGUCAUCCGCGGC-dTdT | + |
| IGF1R-12 | UCAGCAU | AAUGCUGACCUCUGUUACC-dTdT | + |
| IGF1R-13 | CGGUAAU | CAUUACCGAGUACUUGCUG-CU | + |
| IGF1R-16 | CCUCGGA | GGCCUCGAGAGCCUCGGAG-AC | - |

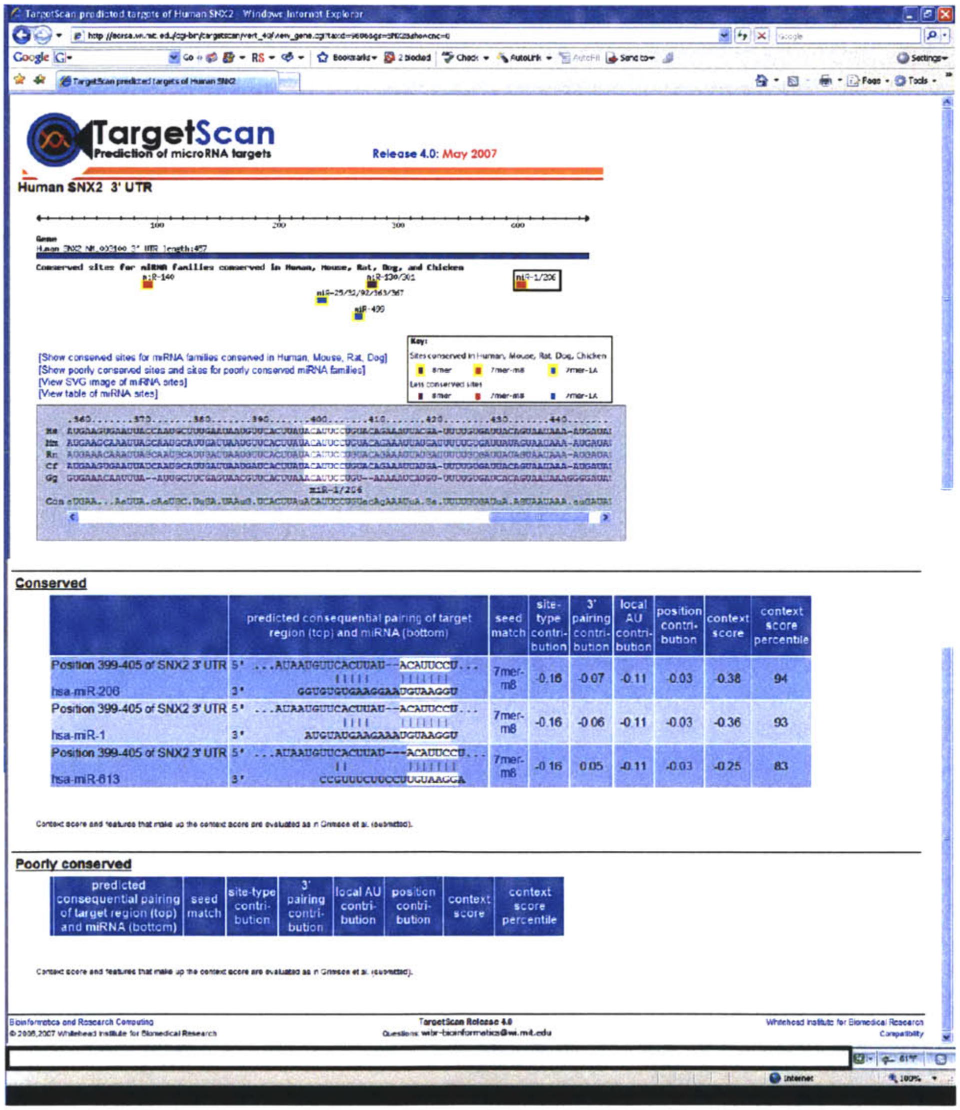
**Supplemental Figure 5.** Evaluation of scoring schemes in an independent dataset. We examined correlations between fold change in mRNA level and 3' pairing score (A), local AU content score (B) and position (C), observed in a dataset not used in development of scoring rubrics (Jackson et al., 2003). For each of the transfection experiments, motif analysis was performed (Farh et al., in preparation) to determine the motif most significantly associated with downregulation. Names and sequences of siRNA sense strands from Jackson et al. (2003) transfection experiments are shown (D); in general, the seed site for one of the strands (indicated as +/-) predominated in accordance with a strand preference for the 5' end with the weaker pairing energy (Schwartz et al., 2003). Results of 18 experiments, each transfecting a duplex for a different siRNA, were consolidated. Messages with a single 7mer-m8 site were scored for 3' pairing, local AU content and position using our previously developed rubrics (Figures 2F, 3B, and 5, respectively) and analyzed as in Figure 6A-C (shown in red). For comparison, the regression lines derived in Figure 6A-C are also shown (grey). The fold changes were larger in the set of siRNA transfections than in the miRNA transfections, and thus the slopes of the regression lines are correspondingly steeper.

For the 3' pairing feature (A), messages with sites scoring >3.5 were markedly downregulated, a result consistent with the miR-155 knockout data (Figure 7E) but in contrast to the linear trend observed for the miRNA transfections (Figure 6A). Perhaps sites with low scores (<3.0), which includes most of the sites (81%), are negligibly affected by 3' pairing, and those with higher scores are affected more than anticipated by fitting all the data to a linear model (red line). To better model this behavior, sites with low scores were excluded (average fold change indicated by dashed line), and the remaining data was modeled by linear regression (blue line). Although more complex, this approach also appears to better reflect the miR-155 in vivo results (Figure 7E). However, because of their rarity, more data will be required to accurately quantify the efficacy of sites with exceptional 3' pairing (scores >4.0), which comprise only ~1% of the population.

## Supplemental Figure S6



D



TargetScan predicted targets of Human SNX2 - Windows Internet Explorer
http://ecrsa.wi.mit.edu/cgi-bin/targetscan/vert_40/view_gene.cgi?taxid=9606&gs=SNX2&showcnc=0
Google
Go
Bookmarks
2 blocked
Check
AutoLink
AutoFill
Send to
Settings
TargetScan predicted targets of Human SNX2
Page
Tools
TargetScan
Prediction of microRNA targets
100
200
300
400
Gene
Human SNX2 NM_003100 3' UTR length:457
miR-140
miR-130/301
miR-25/32/92/363/367
miR-499
miR-1/206
Key:
Sites conserved in Human, Mouse, Rat, Dog, Chicken
8mer
7mer-m8
7mer-1A
Less conserved sites
8mer
7mer-m8
7mer-1A
Internet
100%

Release 4.0: May 2007

**Human SNX2 3' UTR**

Conserved sites for miRNA families conserved in Human, Mouse, Rat, Dog, and Chicken

[Show conserved sites for miRNA families conserved in Human, Mouse, Rat, Dog]
[Show poorly conserved sites and sites for poorly conserved miRNA families]
[View SVG image of miRNA sites]
[View table of miRNA sites]

```
    .360.......370.......380.......390.......400.......410.......420.......430.......440.......
Hs  AUGAAGUGAAUUACCAAUGCUUUGAAUAAUGUUCACUUAUACAUUCCUGUACAGAAAUUACGA-UUUUGUGAUUACAGUAAUAAA-AUGAUAI
Mm  AUGAAGCAAAUUAGCAAUGCAUUGAUUAAUGUUCACUUAUACAUUCCUGUACAGAAAUUAUGAUUUUUGUGAUUAUAGUAAUAAA-AUGAUAI
Rn  AUGAAACAAAUUAGCAAUGCAUUGAUUAAUGUUCACUUAUACAUUCCUGUACAGAAAUUAUGAUUUUUGUGAUUAUAGUAAUAAA-AUGAUAI
Cf  AUGAAGUGAAUUAUCAAUGCAUUGAUUAAUGAUCACUUAUACAUUCCUGUACAGAAAUUAUGA-UUUUGUGAUUACAGUAAUAAA-AUGAUAI
Gg  GUGAAACAAUUUA--AUUGCUUCGAGUAACGUUCACUUAAACAUUCCUGU--AAAAAUCAUGU-UUUUGUGAUCACAGUAAUAAAGGGGAUAI
                                                miR-1/206
Con aUGAA...AaUUA.cAaUGC.UuGA.UAAuG.UCACUUAuACAUUCCUGUacAgAAAUuA.Ga.UUUUGUGAUuA.AGUAAUAAA.auGAUAI
```

**Conserved**

| | predicted consequential pairing of target region (top) and miRNA (bottom) | seed match | site-type contri-bution | 3' pairing contri-bution | local AU contri-bution | position contri-bution | context score | context score percentile |
|---|---|---|---|---|---|---|---|---|
| Position 399-405 of SNX2 3' UTR<br>hsa-miR-206 | 5' ...AUAAUGUUCACUUAU--ACAUUCCU...<br>&#124;&#124;&#124;&#124;&#124; &#124;&#124;&#124;&#124;&#124;&#124;&#124;<br>3' GGUGUGUGAAGGAAUGUAAGGU | 7mer-m8 | -0.16 | -0.07 | -0.11 | -0.03 | -0.38 | 94 |
| Position 399-405 of SNX2 3' UTR<br>hsa-miR-1 | 5' ...AUAAUGUUCACUUAU--ACAUUCCU...<br>&#124;&#124;&#124;&#124; &#124;&#124;&#124;&#124;&#124;&#124;&#124;<br>3' AUGUAUGAAGAAAUGUAAGGU | 7mer-m8 | -0.16 | -0.06 | -0.11 | -0.03 | -0.36 | 93 |
| Position 399-405 of SNX2 3' UTR<br>hsa-miR-613 | 5' ...AUAAUGUUCACUUAU---ACAUUCCU...<br>&#124;&#124; &#124;&#124;&#124;&#124;&#124;&#124;&#124;<br>3' CCGUUUCUUCCUUGUAAGGA | 7mer-m8 | -0.16 | 0.05 | -0.11 | -0.03 | -0.25 | 83 |

Context score and features that make up the context score are evaluated as in Grimson et al. (submitted).

**Poorly conserved**

| predicted consequential pairing of target region (top) and miRNA (bottom) | seed match | site-type contri-bution | 3' pairing contri-bution | local AU contri-bution | position contri-bution | context score | context score percentile |
|---|---|---|---|---|---|---|---|

Context score and features that make up the context score are evaluated as in Grimson et al. (submitted).

Bioinformatics and Research Computing
© 2006,2007 Whitehead Institute for Biomedical Research

TargetScan Release 4.0
Questions: wibr-bioinformatics@wi.mit.edu

Whitehead Institute for Biomedical Research
Compatibility

**Supplemental Figure 6.** Deriving and annotating the context score for a miR-1 target site within the *SNX2* 3'UTR, using the regression parameters of Supplemental Table 6.

(A) Regression relating repression on the array and 3' pairing score, identical to Figure 6A but illustrating how the 3' pairing contribution of the context score was determined for the miR-1 site of *SNX2*. The miR-1 site of *SNX2* had a 3' pairing score of 4 (scored as in Figure 2F). Using the slope of the regression line (–0.031, see Supplemental Table 6 for all parameters used to calculate context scores), y-intercept value (–0.094), and mean fold change for a 7mer-m8 site (–0.161), a score of 4 corresponded to an expected fold change of –0.06 ($\log_2$) over that of an average 7mer-m8 site, calculated as follows: 4(–0.031) – 0.094 + 0.161 = –0.06.

(B) Regression relating expected repression on the array and local AU content score, identical to Figure 6B but illustrating how the local AU content contribution of the context score was determined for the miR-1 site. The miR-1 site of *SNX2* had a local AU content score of 0.76 (scored as in Figure 3B). Using the slope of the regression line (–0.50), y-intercept value (0.108), and mean fold change for a 7mer-m8 site (–0.161), a score of 0.76 corresponded to an expected fold change of –0.11 ($\log_2$) over that of an average 7mer-m8 site, calculated as follows: 0.76(–0.50) + 0.108 + 0.161 = –0.11.

(C) Regression relating expected repression on the array and position within 3'UTR, identical to Figure 6C but illustrating how the position contribution of the context score was determined for the miR-1 site. The miR-1 site of *SNX2* was located 52 nt from the closest UTR terminus. Using the slope of the regression line (0.000091), y-intercept value (–0.198), and mean fold change for a 7mer-m8 site (–0.161), a distance of 52 nt corresponded to an expected fold change of –0.03 ($\log_2$) over that of an average 7mer-m8 site, calculated as follows: 0.000091(52) – 0.198 + 0.161 = –0.03.

(D) TargetScan 4.0 web page illustrating how context scores derived from the newly identified specificity determinants have been annotated for predicted sites in mammals (Targetscan.org). Shown is the context score of the miR-1 target site in *SNX2*, with the contributions from the site type and three context determinants. The combined context score for the miR-1 target site in *SNX2* was equal to the mean repression for the 7mer-m8 site plus the three contributing scores (–0.161 + –0.06 + –0.11 + –0.03 = –0.36). This context score corresponded to a predicted repression in the top 7th percentile of miR-1 sites. To account for the other two context determinants, sites falling within 15 nt of a stop codon are flagged, and sites within optimal distance for cooperative action are displayed.

# Chapter III.

The following chapter includes the manuscript, figures, and supplementary data for a manuscript titled "Prediction of MicroRNA targets under selection", which was later revised into a paper submitted for publication by Robin Friedman, Chris Burge, David Bartel, and myself one year later. I have included here the original manuscript, for which I performed all of the computational analyses; the final submitted paper was improved in several ways, including controls that were better matched for dinucleotide content and were more conservative, an increase in the number of genomes included the analysis to 23, and an expansion of the 3' pairing analysis to include microRNAs with the same seed sequence but differing 3' sequences. The sections on relative strength of microRNA seed sites and experimental evidence for the A1 anchor were not included in the submitted manuscript as analogous results were published in Baek *et al*, *Nature* 2008; Lee Lim was listed as an author for the microRNA transfection experiments that supported these results. The last section of the manuscript which uses conservation of the microRNA seed sequence versus conservation of the adjacent UTR sequence to generate a confidence score reflecting the probability that a site is being conserved due to chance was replaced by an alternative method in the submitted manuscript.

# Prediction of Mammalian MicroRNA Target Sites Under Selection

**Kyle Kai-How Farh[1,2,3], Lee P. Lim[4], David P. Bartel[1,2]***

[1]Howard Hughes Medical Institute, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[2]Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, MA 02142, USA

[3]Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139, USA

[4]Rosetta Inpharmatics (wholly owned subsidiary of Merck and Co.), 401 Terry Avenue N, Seattle, WA 98109, USA

*Contact: dbartel@wi.mit.edu (D.P.B.)

## Summary

We have performed a phylogeny-based conservation analysis of mammalian microRNA (miRNA) targets using recently sequenced genomes, thereby increasing fourfold the number of seed sites under selection. We further define the motif involved in miRNA repression by providing experimental evidence for the Adenosine preference across from the unpaired first nucleotide of the miRNA and ranking the relative effectiveness of sites with perfect and imperfect seed pairing, as well as demonstrating the effects of 3' compensatory pairing outside of the seed region. In addition, we show that the confidence with which conserved target sites can be predicted correlates with the depth of conservation of the seed site itself, and anti-correlates with the depth of conservation of the surrounding sequence. This allows individual conserved sites to be assigned a confidence score reflecting the probability that the conservation of the site is due to selection rather than chance.

## Introduction

A central goal for understanding the functions of miRNAs has been to understand how they recognize their target messages. Conserved Watson-Crick pairing to the 5' region of the miRNA, known as the miRNA seed, enables prediction of targets above the background of false-positive predictions, indicating the importance of this region for miRNA target recognition (Lewis et al, Cell 2003; Lewis et al, Cell 2005, Brennecke et al, PLoS Bio 2005; Krek et al, Nat Genetics 2005). Indeed, many messages that either decrease upon miRNA ectopic expression or increase upon miRNA knock-down have matches to the miRNA seed (Lim et al, Nature 2005; Krutzfeldt et al, Nature 2005; Giraldez et al, Science 2006), and more than a third of the human genes appear to have been under selective pressure to maintain their pairing to miRNA seeds (Lewis et al, Cell 2005). With the recent work in sequencing additional mammalian genomes, we have revisited our previous work in order to provide miRNA target predictions with the greatest possible sensitivity and specificity.

## Results and Discussion

### Incorporating Phylogenetic Information in Conservation Analysis

The recent availability of additional sequenced mammalian genomes (rabbit, cow, tenrec, elephant, and opossum) allows us to incorporate phylogenetic information about the ancestral relationships between these species. We empirically derived the average branch lengths for the relationships between the species using the aligned human 3' UTR sequence and DNAML (Felsenstein, Cladistics 1989), a maximum likelihood phylogeny method (Figure 1A.) The conservation for a particular site is measured by the total branch length connecting all species that have conserved the site, thus appropriately weighting species based on the extent of their divergence and shared ancestry and allowing deeply conserved sites to be identified despite the site not being present in all instances due to divergence or incomplete sequencing and alignment. This contrasts with previous analyses, where conservation was typically a binary decision and required that the site was present among all species in the analysis. We performed signal to background analysis at branch length cutoffs of ≥0.6746 and ≥1.139 and plot the number of conserved sites above each threshold versus the number expected by chance (Figure 1B). The cutoffs were chosen based on the minimum branch length required to span a site conserved for human, mouse, rat, and dog (HMRD), 0.6746, and for HMRD + chicken, 1.139, for ease of comparison to previous studies. Increasing the branch length threshold is equivalent to requiring more stringent conservation, and increases specificity at the cost of sensitivity. Employing phylogeny in the prediction of miRNA targets greatly improved the sensitivity of method. In total, we predicted 29598 conserved 7mer or 8mer sites with conserved branch lengths at or above the HMRD+C cutoff compared to only 10356 false positive controls, giving us an estimate of 17944 unique sites under selection (signal above background) and a signal to background ratio of 2.86:1. In contrast, the study by Lewis et al (Cell 2005), required conservation in each of human, mouse, rat, dog, and chicken, and examined 6mer sites, obtaining a signal above background of 7453 sites with a signal to background ratio of 2.4:1.

### Experimental Support for Non-Watson-Crick Recognition of an A at Target Position 1

Introducing a miRNA into cells changes the mRNA expression profile, with messages containing seed matches tending to be down-regulated (Lim et al., 2005). Reasoning that this approach could provide an experimental method of independently determining the types of sites that mediate repression, we considered a set of 11 miRNA transfection experiments: miR-1, miR-7, miR-9, miR-122, miR-124, miR-128, miR-132, miR-133, miR-142-3p, miR-148, miR-181 (Grimson et al 2007, Lee et al 2005). A Chi-Square test on each possible 7mer revealed those motifs associated with 3' UTRs of messages confidently down-regulated (*P* <0.01) on the array (Table

1). Overall, the two 7mers most significantly associated with down-regulated messages both included a perfect Watson-Crick match to the miRNA seed (nucleotides 2-7). One 7mer site, referred to here as the 7mer-m8 site, contained the seed match augmented by a match to nucleotide 8 of the miRNA. The other 7mer, the 7mer-A1 site, contained the seed match augmented by an A at position 1 of the target. The presence of either of these two 7mers could explain much of the direct effect of introducing a miRNA, in that 75% of messages downregulated on the array had at least one of the two 7mers in their 3'UTR.

An important unresolved difference among current target-prediction methods concerns pairing to the first nucleotide of the miRNA. TargetScanS rewards an A across from position 1 (7mer-A1), whereas other algorithms reward a Watson-Crick match across from position 1 (7mer-m1) (Lewis et al, 2005; Brennecke et al, 2005; Krek et al, 2005). The preference for an invariant A across from position 1 of the miRNA is supported by conservation evidence (Lewis et al, 2005), as well as crystallographic studies indicating that the 5' most nucleotide of the miRNA is unpaired to its target message (Ma et al, Nature 2005). The two types of sites were identical for the 10 transfected miRNAs that began with a U, but differed for miR-181, which began with an A. For miR-181, the 7mer-A1 was significantly more strongly associated with down-regulation than was the 7mer-m1 (Table 2, $P = 0.0040$, Chi-square test). To extend these results to additional silencing RNAs that do not begin with a U, we performed a similar analysis on a published dataset of siRNA transfections (Birmingham, 2006). Five siRNA guide strands had a nucleotide other than U at the first position, and in each case the most enriched 7mer starting with a seed match was the 7mer-A1 site, which significantly outperformed the 7mer-m1 sites (Table 2, $P < 10^{-10}$, Chi-square test). Furthermore, 7mer-m1 sites performed no better than comparable 7mers with non-A mismatches at this position, indicating that a Watson-Crick match at position 1 provided little or no benefit.

**Relative Effectiveness of Seed Sites**

Having experimentally defined the canonical sites, we next examined their relative effectiveness. For each of the 11 miRNA transfections, we selected the set of genes that had a single 8mer site in the UTR and no other site with a seed match, and plotted the distribution of expression changes due to transfection of the corresponding miRNA. The same analysis was performed for single 7mer-m8 sites, 7mer-A1 sites, and 6mer seed matches. Expression changes for genes that contained no canonical sites were also plotted for comparison. The 8mer site was most effective, followed by the 7mer-m8, the 7mer-A1, and then the 6mer, which was marginally effective (Figure 1C).

The observed hierarchy of effectiveness was consistent with site conservation (Figure 1B), indicating that our experimental results of mRNA destabilization in cell culture corresponded to selective pressures on the animal during mammalian evolution. For example, the greater effectiveness of the 8mer explained why the number of 8mer sites selectively maintained was comparable to that of the 7mer-m8 and 7mer-A1 sites, even though the 7mer sites, with fewer nucleotides, would each emerge and be maintained more easily than would 8mer sites. Likewise, the better performance of the 7mer-m8 site explained why this 7mer appeared to be selectively maintained more often.

### Sites With Imperfect Seed Pairing

The improved conservation analysis allowed us to re-examine the conservation of sites lacking perfect seed pairing. Among the large number of possibilities for disrupting a seed site, we considered only 8mers with single insertions, mismatches, or GU-wobbles. We found that a mismatch across from position 2 of the miRNA that preserved the 6mer across from nucleotides 3-8 performed nearly as well as the seed site (nucleotides 2-7) in the conservation analysis (Figure 1B), and therefore treated them separately from the remainder of the imperfect site analysis. In contrast, single, non-GU wobble mismatches at nucleotides 3-7 substantially compromised the signal to background ratio, while GU wobbles were better tolerated. Despite the low gains over noise, the large number of sites with imperfect pairing suggests that a substantial number of sites under selection fall into this category.

Next, we measured the repression profiles of messages that lacked perfect seed matches. Although a mismatch at a single seed nucleotide dramatically affected site performance, slight repression was detected, falling below that of a canonical 6mer (Figure 2B; note expanded scale of the X axis compared to Figure 1C). Unlike the conservation analysis, considering only those sites with a single GU wobble did not improve overall efficacy (Figure 2C). As with the conservation analysis, the most tolerated mismatch was at the 5'-most position of the seed (miRNA nucleotide 2). Indeed, efficacy of 6mer sites that paired to nucleotides 3-8 approached that of the canonical 6mer sites (Figure 2D), although a significant difference between the two was detected (P=0.0324, 2-sided KS test.) Overall, the promising results observed for the 3-8 6mer suggest that it should be considered as a weak type of seed site, whereas the marginal efficacy of sites with mismatches was in accord with the weak but detectable preferential conservation of such sites above chance (Lewis, 2005).

One possibility for explaining the small but significant effects of imperfect seed sites in the conservation and expression analyses was that a minority of these sites might contain extensive compensatory 3' pairing. We performed a conservation analysis for sites with increasing amounts of 3' pairing, based on the scoring rubric that awarded sites with Watson-Crick base pairing centered at nucleotides 13-16 (Grimson et al, 2007). For controls, we used the same seed and 3' sequences, but the 3' ends of the miRNAs were assigned to the wrong seeds. Sites containing the 2-7 or 3-8 6mer demonstrated greatly improved signal-to-noise as the amount of conserved 3' compensatory pairing increased (Table 3). In contrast, the imperfect seed sites needed more extensive 3' pairing for the bump in signal-to-noise to be observed, and the increase in signal-to-noise was correspondingly smaller. Therefore, it appears that 3' pairing is principally a supplementary feature of canonical seed sites, and rarely plays a role in salvaging imperfect seed sites. Based on the estimated background, fewer than 10 conserved 3' compensatory sites are expected to exist in mammals.

### Mammalian-Only miRNAs Have Fewer Selectively Maintained Sites

The recent sequencing of several additional mammalian genomes, along with opossum and chicken as outgroups, has allowed us to examine the more recent evolution of miRNAs and their targets. Among 143 mammalian miRNA families (defined as those conserved between human-mouse-rat-dog), 73 were conserved to zebrafish or deeper, leaving 70 more recently evolved mammalian-specific miRNA families. However, in contrast to the vertebrate miRNA families, the mammalian-specific miRNAs did not have appreciable signal above noise (Figure 3A, showing conservation to the HMRD 0.6746 branch length cutoff.) This could be because they are expressed at low levels or in more limited contexts, and hence have fewer regulatory targets, or because the common vertebrate miRNAs, by virtue of being present for a longer period of evolutionary time, have had more time to acquire target sites. To distinguish between these two possibilities, we considered only those sites to vertebrate miRNAs which were absent in marsupial and chicken. Even among these sites, however, the set of 73 deeply conserved vertebrate miRNAs had appreciable signal-to-background (Figure 3B), while the mammalian-only set did not show improvement. This suggests that the common vertebrate miRNA set, besides being more deeply conserved, is also more consequential in the animal, as evidenced by their ability to acquire new target sites following the divergence of the placental mammals from marsupial and chicken.

Despite having far fewer predicted sites above background than the common vertebrate miRNA set, individual miRNAs within the mammalian-only miRNA set do have robust signal-to-background ratios, suggesting that a subset of them may be more suitable for conservation-based target prediction. The signal-to-background and signal minus background values for all miRNAs considered in this study are reported in Table 4.

### Confidence of Target Predictions Depends Both on Conservation of the Site and of the Surrounding UTR Context

Conservation has been an invaluable tool for distinguishing biologically relevant sites from identical sequences that would be expected to occur in the genome by chance. To determine the extent to which conserved sites are more likely to function than are nonconserved sites of the same type, we evaluated the performance of 7mer-m8 sites at different branch length cutoffs (Figure 4A), observing that the effectiveness of sites correlated with their depth of conservation. Presumably, sites being selectively maintained in the animal need to have the potential to confer significant downregulation, and would be associated with UTR contexts favorable for efficacy. The more deeply conserved sites were downregulated more extensively, on average, because they were more heavily enriched for sites under selection.

Conserved miRNA target sites would be expected to be informative only in the context of UTRs whose sequence has diverged sufficiently across different species to distinguish conservation due to selection from conservation occurring by chance. When messages containing single 7mer-m8 sites conserved at least to human/mouse/rat/dog were partitioned into equally sized sets based on the fraction of their 3' UTR that was conserved, those conserved sites present in the context of less well-conserved UTRs were downregulated to a greater extent (Figure 4B). In 3' UTRs with a very high density of conservation, any 7- or 8mer is conserved at a high rate, hence conservation of miRNA sites provides little information for distinguishing sites under evolutionary selection from those occurring incidentally (Lewis, 2005). On average, the sites present in less well-conserved UTRs outperformed those in the more well-conserved UTRs, because a larger fraction of the conserved sites in the well-conserved UTRs were likely to be sites that were only being conserved by chance, or were conserved for reasons unrelated to the function of the miRNA.

### Assigning Confidence Scores for Individual Sites

By measuring both the branch length of the site itself and the average branch length of 7mers in the UTR context in which the site occurs, we can populate a 2D-plot for miRNA target sites and their associated controls. Using a nearest neighbor approach, we were able to sample the number of miRNA and control sites within all regions of the plot, thus arriving at empirical signal to background estimates. The overall signal to background estimate for 8mer sites for all miRNAs is shown as a contour map (Figure 5A), and can be used to estimate signal to background values for individual miRNA sites, based on the branch length of the site and the average branch length of the UTR in which the site resides. Unsuprisingly, the sites with the greatest signal to background are those in the upper left quadrant, i.e. highly conserved sites occurring in moderately or poorly conserved UTRs. However, signal-to-background does not simply follow a ratio of site branch length and UTR branch length. In particular, signal-to-background remains respectable among deeply conserved sites present in deeply conserved UTRs (upper right quadrant), consistent with a previous observation that highly conserved UTRs have the greatest density of conserved targets per unit length, but because of the increased conservation of background sequences, these targets are harder to identify (Lewis et al, 2005). We also calculated confidence scores based on the signal to background (S2B) values, where confidence score = (S2B-1)/S2B, and reflects the probability that a site is being conserved by selection rather than chance (confidence score ranges from 0 to 1, and requires that S2B≥1.) A histogram plotting the distribution of confidence scores for individual 8mer sites is shown in Figure 5B.

We turned to the array data to evaluate our confidence scores for individual sites. Confidence scores were obtained for all 7mer-m8 sites on the array using the nearest neighbor approach, and the relationship between confidence score and average repression on the array was plotted (Figure 5C). Not only did repression increase for sites with higher confidence levels, but the two variables followed a linear relationship (Linear regression: $y = 0.1742x – 0.1106$, $r^2 = 0.0285$, $P < 10^{-20}$), as would be expected given a mixing of two populations of sites with intrinsically different mean repression values.

### Summary

While in vitro experimental assays may show the feasibility of a regulatory interaction, this does not necessarily imply that the interaction actually occurs in nature. Thus, the problem of accurately predicting whether the conservation of a miRNA site is due to selection as opposed to chance is an old question familiar to biologists, and cannot be fully addressed merely by assaying the effectiveness of the site and obtaining a fold-repression value. The combination of newly sequenced genomes, functional data on microarray targets, and improved methods of analysis, has vastly improved both the sensitivity and specificity with which we can identify targets under selection. For instance, by considering all 6mer/7mer/8mer sites at the branch length cutoff corresponding to conservation to HMRD + chicken, we maintain a modest signal-to-noise ratio of 2.0:1 while further increasing our signal above background to 28101 sites. This figure quadruples the number of sites above background reported in previous genome-wide studies in mammals, and means that on average, each human gene contains ~2 miRNA binding sites under selection in its 3' UTR. In addition, by taking into account the background level of conservation of the UTR in which the site occurs, we have assigned confidence predictions to all individual miRNA sites and validated the utility of this approach using functional data.

Significant challenges remain. For instance, mammalian-only miRNAs have been conserved, yet their low signal-to-background values render current target prediction methods inadequate, since the few conserved sites they might have are insignificant next to the background conservation level of genomic 3' UTR sequence. Another outstanding question relates to the hundreds of thousands of potential sites that have some degree of imperfect seed pairing. These sites have been shown to be marginal but significant using both conservation methods and functional studies. Given the ease at which a 7 or 8 nt motif with a wildcard matches the genome, the number of imperfect seed sites under selection likely exceeds that of canonical 7mer and 8mer sites. Since the estimated number of sites under selection numbers in the tens of thousands, and even the best imperfect seed sites on the array are only subtly regulated, these marginal results are unlikely to be explained by a small subset of imperfect seed sites that are highly effective but whose voice is diluted by the larger crowd of ineffective sites. Our results suggest that miRNA target sites, despite adhering to variations around a core 8-nt seed motif, come with layer after layer of subtlety and variation. It appears that just as there are sites that have been selected to confer greater downregulation than others, there are also sites that evolution has apparently enjoyed tinkering with to a larger extent than others.

## Experimental Procedures

### MicroRNA and mRNA Sequence Data

MicroRNAs conserved in human, mouse, rat, and dog were clustered into 145 families based on miRNA nucleotides 2-8, and families were classified as either vertebrate or mammalian based on whether zebrafish orthologs were identified. Besides the analysis in Figure 3, only the 73 families conserved to zebrafish were used for the analyses. Human annotated 5' UTR, ORF, and 3' UTR sequences were obtained from RefSeq, and orthologous sequences in mouse, rat, rabbit, dog, cow, elephant, tenrec, opossum, and chicken were derived from the UCSC genome browser multiZ multiple genome alignments (Blanchette, 2004). When multiple RefSeq identifiers mapped to a single Entrez Gene entry, the RefSeq annotation with the longest UTR was used.

### Conservation Analysis

In addition to incorporating phylogeny information in the conservation analysis, we improved upon our controls by selecting control sequencing that were matched to the A/U/C/G nucleotide composition of the miRNA seeds and that had the same frequency of non-conserved sites as real miRNA seeds (non-conserved sites were defined as those having no other conservation outside of human, or zero branch length). The AU-correction was necessary because of the differential conservation of A/U/C/G in the genome, and the improved performance is shown and described (Figure S1, Supplemental Text). The decision to select controls based on the number of strictly nonconserved sites corrects for the fact that miRNAs have signal-above-background even in one genome, and therefore selecting controls just based on the frequency of motif occurrence in one genome underestimates the true number of targets.

### Expression Analysis

For displaying the changes in mRNA abundance (e.g. Figure 1D), the down-regulation values for messages with a cognate site were binned such that each bin corresponded to a range of values of 0.1 on a $\log_2$ scale. Unless indicated otherwise, only messages with a single site to the cognate miRNA were considered. For instance, when evaluating single 8mer sites, only genes with a single 8mer in the UTR and no additional 7mers or 6mers were considered. Likewise, when a single 6mer was being considered, genes were excluded in which that 6mer was part of a canonical 7-8mer. Analyses for multiple sites (Figure 6, Table 3) and additional specificity determinants (Figures 3, 4, and 5) were performed in the same way, requiring the exact numbers of the motif being examined and no other canonical sites, except that additional 6mers were allowed in order to increase the sample sizes, based on the observation that the presence of additional 6mers had marginal effect in the context of stronger sites.

### Motif Analyses

A Chi-Square test with one degree of freedom was performed to test each of the 16,384 possible 7mers for correlation with down-regulation. Genes were first divided into two categories based on whether they were down-regulated with *P* <0.01 on the array. For 3' UTRs of each category, the total number of occurrences of the test 7mer was tabulated, as well as the sum of all 7mers (approximately equal to the total length of all 3' UTRs). The observed values shown in Tables 1 and 2, and represented the total occurrences of the test motif in down-regulated UTRs; the expected values were the number of occurrences anticipated if the motif was uncorrelated with down-regulation, calculated as the sum of all 7mers in down-regulated UTRs multiplied by the fraction of 7mers in all UTRs that were the test 7mer.

### Calculating the Fraction of Genes Downregulated

To determine for each type of site the percentage of genes down-regulated, we compared the cumulative distribution of expression changes for messages with the site versus those with no canonical site, calculating the maximum positive cumulative difference between the two distributions. To correct for bumpiness in the cumulative distributions, we performed 100 permutations in which the down-regulation values for all genes were randomly shuffled, while maintaining the size of the original distributions, and the median maximum positive deviation for these 100 control permutations was subtracted from the value obtained from the real distributions.

### Acknowledgments

We would like to thank Alex Stark and Daehyun Baek for their advice on the phylogenic conservation analysis. Supported by an NIH grant (D.B.) and NIH postdoctoral fellowship (A.G.). D.B. is an HHMI investigator.

## References

Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, Baertsch R, Rosenbloom K, Clawson H, Green ED, Haussler D, Miller W. Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res. 2004 Apr;14(4):708-15.

Brennecke J, Stark A, Russell RB, Cohen SM. Principles of microRNA-target recognition. PLoS Biol. 2005 Mar;3(3):e85.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). Cladistics 5: 164-166

Giraldez AJ, Mishima Y, Rihel J, Grocock RJ, Van Dongen S, Inoue K, Enright AJ, Schier AF. Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs. Science. 2006 Apr 7;312(5770):75-9.

Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lim LP, Bartel DP. MicroRNA targeting specificity in mammals: determinants beyond seed pairing. Accepted.

Krek A, Grün D, Poy MN, Wolf R, Rosenberg L, Epstein EJ, MacMenamin P, da Piedade I, Gunsalus KC, Stoffel M, Rajewsky N. Combinatorial microRNA target predictions. Nat Genet. 2005 May;37(5):495-500.

Krützfeldt J, Rajewsky N, Braich R, Rajeev KG, Tuschl T, Manoharan M, Stoffel M. Silencing of microRNAs in vivo with 'antagomirs'. Nature. 2005 Dec 1;438(7068):685-9.

Lewis BP, Burge CB, Bartel DP. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. Cell 2005 Jan 14;120(1):15-20.

Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. Cell. 2003 Dec 26;115(7):787-98.

Lim LP, Lau NC, Garrett-Engele P, Grimson A, Schelter JM, Castle J, Bartel DP, Linsley PS, Johnson JM. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. Nature 2005 Feb 17;433(7027):769-73.

Ma JB, Yuan YR, Meister G, Pei Y, Tuschl T, Patel DJ. Structural basis for 5'-end-specific recognition of guide RNA by the A. fulgidus Piwi protein. Nature. 2005 Mar 31;434(7033):666-70.

Yekta S, Shih IH, Bartel DP. MicroRNA-directed cleavage of HOXB8 mRNA. Science 2004 Apr 23;304(5670):594-6.

**Table 1.** 7mer motifs associated with 3' UTRs of messages confidently down-regulated upon transfection of miR-142 (5'-uGUAGUGuuuccuacuuuaugga, seed indicated in capitals). Messages were considered down-regulated using significance cutoff of $P$ <0.01 on the array. The 7mer-m8 (*) and 7mer-A1 (+) sites are indicated. Other over-represented motifs included matches to nucleotides 3-8 preceded by an A or U matches to nucleotides 2-7 or 1-6 followed by an A or U.

| Motif | | Expected | Observed | *P* ~ |
|---|---|---|---|---|
| n**ACACUAC**nn | * | 7 | 85 | $10^{-122}$ |
| Nn**CACUACA**n | + | 10 | 79 | $10^{-78}$ |
| A**ACACUA**nnn | | 16 | 66 | $10^{-28}$ |
| Nn**CACUAC**Un | | 9 | 46 | $10^{-26}$ |
| **UACACUA**nnn | | 9 | 45 | $10^{-25}$ |
| NG**CACUAC**nn | | 5 | 29 | $10^{-20}$ |
| UUUUAAA | | 195 | 322 | $10^{-17}$ |
| nnn**ACUACA**U | | 11 | 43 | $10^{-17}$ |
| UUUAAAA | | 172 | 280 | $10^{-14}$ |
| nnn**ACUACA**A | | 10 | 35 | $10^{-13}$ |

**Table 2.** 7mer motifs beginning with a seed match that were associated with 3' UTRs of messages confidently down-regulated after transfection of the indicated miRNA or siRNA. Analysis was as in Table 1. The 7mer-A1 is listed first in each series, and the 7mer-m1 alternative is indicated (*).

| Motif | Expected | Observed | *P* ~ |
|---|---|---|---|
| miR-181 5'-aACAUUCaacgcugucggugagu | | | |
| GAATGTA | 18 | 88 | $10^{-51}$ |
| GAATGTC | 13 | 21 | 0.04 |
| GAATGTG | 24 | 35 | 0.04 |
| GAATGTT* | 24 | 42 | $10^{-3}$ |
| map2k1 siRNA 5'-aGAACCTccatccatgtgc | | | |
| AGGTTCA | 36 | 63 | $10^{-3}$ |
| AGGTTCC | 31 | 43 | 0.05 |
| AGGTTCG | 4 | 8 | 0.11 |
| AGGTTCT* | 47 | 69 | 0.01 |
| map2k3 siRNA 5'-aCTTGATccagagaacctc | | | |
| ATCAAGA | 54 | 82 | $10^{-2}$ |
| ATCAAGC | 30 | 35 | 0.49 |
| ATCAAGG | 41 | 47 | 0.39 |
| ATCAAGT* | 44 | 52 | 0.28 |
| map2k4 siRNA 5'-aGTTGCTtcaaatctgctc | | | |
| AGCAACA | 69 | 135 | $10^{-11}$ |
| AGCAACC | 41 | 56 | 0.04 |
| AGCAACG | 11 | 21 | 0.01 |
| AGCAACT* | 55 | 83 | 0.01 |
| gapdh2 siRNA 5'-cTTGAGGctgttgtcatac | | | |
| CCTCAAA | 45 | 95 | $10^{-11}$ |
| CCTCAAC | 25 | 47 | $10^{-3}$ |
| CCTCAAG* | 38 | 52 | 0.04 |
| CCTCAAT | 22 | 43 | $10^{-3}$ |
| ppib5 siRNA 5'-cTCTCCTgtagctaaggcc | | | |
| AGGAGAA | 77 | 105 | 0.01 |
| AGGAGAC | 51 | 58 | 0.41 |
| AGGAGAG* | 90 | 75 | 0.14 |
| AGGAGAT | 49 | 63 | 0.06 |

**Table 3.** Signal to noise analysis for 3' pairing. Sites were scored for 3' pairing as in Grimson et al. The 3' pairing score is approximately the number of contiguous Watson-Crick base pairs situated at positions 13-16 of the miRNA. Scores of 2+ in the table are greater than or equal to 2, but less than 3. The lone imperfect seed site with a 8+ pairing score is the miR-196 and HOXB8 interaction (Yekta et al, 2004.)

**Imperfect Seed Sites**

| 3' Pairing Score | Conserved miRNA Sites | Conserved Controls | Signal-to-Background Ratio |
|---|---|---|---|
| **less than 2** | 47139 | 46555.6 | 1.01 |
| 2 + | 23883 | 24303.4 | 0.98 |
| 3 + | 4115 | 4139.2 | 0.99 |
| 4 + | 588 | 596.4 | 0.99 |
| 5 + | 58 | 49.7 | 1.17 |
| 6 + | 3 | 2.57 | 1.53 (for scores of 6 or greater) |
| 7 + | 0 | 0.04 | |
| **8 or greater** | 1* | 0.00 | |

**Seed Sites Containing Either the 3-8 or the 2-7 6mer Motif**

| 3' Pairing Score | Conserved miRNA Sites | Conserved Controls | Signal-to-Background Ratio |
|---|---|---|---|
| **less than 2** | 58421 | 58565.4 | 1.00 |
| 2 + | 29193 | 29467.2 | 0.99 |
| 3 + | 5004 | 4803.0 | 1.04 |
| 4 + | 741 | 566.6 | 1.31 |
| 5 + | 80 | 46.3 | 1.73 |
| 6 + | 7 | 2.71 | 2.66 (for scores of 6 or greater) |
| 7 + | 2 | 0.27 | |
| **8 or greater** | 0 | 0.40 | |

**Table 4.** Signal to background values for all miRNAs used in this study. S = signal, B = background. Signal to background ratio (S / B, the prediction accuracy), Signal above background (S – B, the total number of sites under selection). MicroRNAs names with "as" appended indicate the passenger strand of the annotated miRNA. Signal to background values are reporter for both 7mer-m8 and 7mer-A1 sites.

**MicroRNA families conserved to Zebrafish**

| miRNA Family | 7mer-m8 S / B | 7mer-m8 S - B | 7mer-m8 stdev(S – B) | 7mer-A1 S / B | 7mer-A1 S - B | 7mer-A1 stdev(S – B) |
|---|---|---|---|---|---|---|
| let-7 | 3.858782 | 216.3285 | 35.01552 | 6.073227 | 312.4182 | 27.10211 |
| 1 | 2.780159 | 174.1639 | 35.64953 | 1.977432 | 174.9799 | 63.13112 |
| 7 | 1.84231 | 94.18385 | 53.13088 | 1.515632 | 114.3103 | 158.6225 |
| 9 | 4.895774 | 373.2031 | 34.28266 | 2.588554 | 298.8641 | 75.41533 |
| 10 | 0.723649 | -52.7001 | 78.70392 | 0.871605 | -12.6685 | 32.12693 |
| 15 | 3.406921 | 612.518 | 134.3426 | 2.747763 | 108.7676 | 23.0869 |
| 17 | 2.161499 | 325.639 | 164.9144 | 1.079008 | 15.08398 | 104.6776 |
| 18 | 1.498863 | 49.92412 | 43.43594 | 0.816559 | -18.1968 | 37.70397 |
| 19 | 5.785943 | 407.7935 | 33.0918 | 2.393887 | 168.8581 | 35.93005 |
| 21 | 2.433044 | 91.88278 | 25.2266 | 2.288487 | 65.31148 | 17.02412 |
| 22 | 2.321052 | 204.8979 | 61.79893 | 1.614042 | 38.04373 | 22.71605 |
| 23 | 1.480103 | 117.7468 | 73.92074 | 1.19495 | 65.74732 | 130.5662 |
| 24 | 1.97177 | 131.5887 | 57.44237 | 1.145624 | 32.15956 | 120.8913 |
| 26 | 1.923324 | 103.2144 | 45.7887 | 1.913254 | 234.3691 | 78.04031 |
| 27 | 2.614271 | 256.256 | 66.42815 | 2.106342 | 274.7023 | 116.4388 |
| 29 | 4.171165 | 458.4361 | 48.63963 | 5.672752 | 167.2149 | 11.49605 |
| 30 | 2.413022 | 230.1337 | 79.8416 | 4.431989 | 302.7778 | 26.44083 |
| 31 | 1.372462 | 50.4771 | 54.05802 | 1.447325 | 60.88683 | 48.75131 |
| 33 | 1.325883 | 29.24846 | 30.774 | 1.761103 | 154.7183 | 66.19596 |
| 34 | 2.775467 | 209.182 | 51.58825 | 3.383474 | 165.5448 | 34.25391 |
| 92 | 2.648981 | 163.0941 | 38.84243 | 3.938169 | 303.6525 | 34.27415 |
| 93 | 2.024235 | 265.1368 | 154.5408 | 2.807594 | 104.2993 | 27.59367 |
| 96 | 4.047095 | 187.4744 | 24.20359 | 4.184781 | 482.4986 | 62.98081 |
| 99 | 2.08503 | 9.887421 | 4.456314 | 2.769269 | 9.583407 | 3.997079 |
| 101 | 1.773131 | 92.43747 | 38.07414 | 2.719272 | 319.9208 | 61.05164 |
| 103 | 2.908601 | 200.1386 | 39.13992 | 1.952979 | 174.2024 | 75.0638 |
| 122 | 1.568487 | 32.61987 | 25.99673 | 0.655494 | -63.068 | 112.0692 |
| 124 | 6.755224 | 644.0866 | 38.81808 | 4.567957 | 417.0987 | 37.04341 |
| 125 | 2.087269 | 158.876 | 49.23735 | 2.163538 | 149.5067 | 53.29371 |
| 126 | 1.88228 | 0.937459 | 0.959726 | 4.131111 | 20.46423 | 3.565494 |
| 128 | 1.338325 | 90.50135 | 113.8877 | 3.332473 | 345.7617 | 60.94862 |
| 129 | 1.183575 | 38.77555 | 80.99375 | 0.140529 | -966.323 | 619.2212 |
| 130 | 4.152315 | 413.7479 | 51.51293 | 2.92749 | 113.2466 | 26.40585 |
| 132 | 1.482175 | 69.29226 | 44.52528 | 1.410746 | 68.42143 | 54.49507 |
| 133 | 1.895424 | 112.4345 | 86.24827 | 3.209326 | 167.9716 | 34.6211 |
| 135 | 2.811392 | 240.3255 | 48.97862 | 2.110566 | 82.08619 | 33.07349 |
| 137 | 0.961974 | -7.55003 | 108.6165 | 4.305317 | 290.9693 | 38.42738 |
| 138 | 1.948136 | 105.6115 | 43.07222 | 2.328998 | 136.9514 | 37.24225 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 139 | 1.336742 | 40.55791 | 32.22793 | 1.133704 | 21.46425 | 57.45762 |
| 140 | 1.674205 | 70.87551 | 43.77679 | 1.300012 | 39.69355 | 54.56925 |
| 141 | 1.644999 | 176.0515 | 118.6299 | 1.439503 | 53.73554 | 40.91397 |
| 142-3p | 2.070982 | 52.23086 | 23.38034 | 1.462921 | 38.92167 | 33.62845 |
| 143 | 0.809937 | -39.4236 | 124.2543 | 1.268663 | 40.87134 | 80.28071 |
| 144 | 1.427225 | 78.42699 | 63.00469 | 3.046445 | 354.6832 | 69.32381 |
| 145 | 1.533495 | 89.06111 | 69.28612 | 1.194731 | 50.36447 | 102.0314 |
| 146 | 0.549684 | -77.0074 | 81.97337 | 0.864566 | -28.5102 | 108.0261 |
| 148 | 2.341961 | 218.8888 | 69.48697 | 2.125558 | 110.1433 | 28.75978 |
| 150 | 0.405407 | -221.465 | 148.6101 | 1.033204 | 6.491617 | 93.64578 |
| 153 | 3.174518 | 135.6283 | 25.73851 | 3.948695 | 290.4864 | 32.15977 |
| 181 | 2.139153 | 317.9176 | 99.37829 | 2.028578 | 148.0568 | 44.46087 |
| 182 | 2.3413 | 171.8661 | 65.24219 | 4.557933 | 441.0403 | 52.0472 |
| 183 | 1.543029 | 76.36751 | 83.57143 | 1.527946 | 60.12165 | 51.15182 |
| 184 | 1.036408 | 0.878217 | 15.03291 | 1.185283 | 1.875835 | 6.515008 |
| 187 | 1.429094 | 5.104348 | 5.693307 | 0.549145 | -4.10506 | 3.826817 |
| 192 | 1.406852 | 26.89499 | 24.7754 | 1.085458 | 8.975236 | 34.92599 |
| 193 | 1.13899 | 17.2061 | 46.31417 | 1.275538 | 20.08958 | 27.28854 |
| 194 | 1.135403 | 15.62251 | 36.27742 | 1.218639 | 41.44429 | 56.99499 |
| 196 | 3.041865 | 118.1408 | 25.1401 | 1.09387 | 5.749575 | 18.73072 |
| 199 | 1.394927 | 54.92459 | 62.05826 | 1.310288 | 55.17649 | 69.76176 |
| 200b | 3.337734 | 441.9496 | 61.01623 | 1.473405 | 76.79073 | 55.73867 |
| 203 | 1.319762 | 128.1702 | 162.714 | 1.084243 | 37.91629 | 201.4883 |
| 204 | 1.177284 | 42.76683 | 98.86241 | 0.697109 | -96.4582 | 129.4021 |
| 205 | 1.127931 | 19.16809 | 53.60332 | 1.105675 | 30.20174 | 128.3903 |
| 210 | 0.929277 | -0.98938 | 7.42973 | 3.40011 | 8.470703 | 2.224783 |
| 214 | 1.88762 | 216.3069 | 96.45392 | 2.172719 | 184.5936 | 77.0064 |
| 216 | 1.033578 | 7.17965 | 90.29178 | 0.777808 | -20.8535 | 30.82662 |
| 217 | 1.402772 | 58.28656 | 42.6867 | 1.467463 | 70.71852 | 55.16731 |
| 218 | 2.717274 | 240.154 | 56.33563 | 2.938986 | 183.4095 | 41.926 |
| 219 | 4.425791 | 91.3381 | 9.845708 | 2.223088 | 100.6821 | 24.37504 |
| 221 | 2.023791 | 65.76412 | 24.37025 | 2.132829 | 130.1291 | 32.13519 |
| 223 | 1.69504 | 49.61527 | 25.80342 | 1.269165 | 36.90198 | 53.1445 |
| 338 | 1.089282 | 13.93396 | 59.28098 | 1.417906 | 84.29415 | 116.5424 |
| 375 | 0.787554 | -0.80926 | 1.896236 | 1.563424 | 145.9532 | 141.515 |

**MicroRNA families conserved only among mammals**

| miRNA | 7mer-m8 | 7mer-m8 | 7mer-m8 | 7mer-A1 | 7mer-A1 | 7mer-A1 |
|---|---|---|---|---|---|---|
| Family | S / B | S - B | stdev(S – B) | S / B | S - B | stdev(S – B) |
| 28 | 1.207373 | 3.60687 | 10.05985 | 1.220157 | 35.90628 | 96.25466 |
| 30 | 0.887843 | -17.3066 | 84.90591 | 1.078712 | 26.26855 | 144.2471 |
| 34b | 1.9749 | 20.7331 | 9.751901 | 2.016587 | 168.3737 | 75.65442 |
| 127 | 0.588549 | -1.39819 | 2.703318 | 2.579434 | 9.797089 | 3.656563 |
| 134 | 0.666367 | -13.5182 | 13.45691 | 0.926627 | -8.5518 | 38.80786 |
| 136 | 0.689813 | -25.1814 | 34.70527 | 0.729936 | -50.3178 | 68.15075 |
| 142-5p | 1.334815 | 28.34406 | 35.82429 | 1.319112 | 94.83031 | 101.6508 |
| 151 | 1.429709 | 8.415592 | 7.324613 | 1.045157 | 2.28991 | 18.95551 |
| 185 | 1.256185 | 13.6639 | 29.96233 | 0.922843 | -13.9626 | 84.5778 |
| 186 | 0.931753 | -9.74172 | 68.13355 | 0.767934 | -136.894 | 316.1696 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 188 | 1.134147 | 4.021523 | 24.81956 | 0.850562 | -22.4888 | 79.87208 |
| 189 | 1.226067 | 3.134525 | 5.447203 | 0.874168 | -8.06094 | 22.61192 |
| 190 | 2.000844 | 25.01055 | 12.23806 | 0.847702 | -9.88132 | 30.78857 |
| 191 | 6.437839 | 8.446684 | 1.132621 | 3.648586 | 31.94053 | 5.281131 |
| 208 | 2.786146 | 4.487569 | 2.751404 | 0.389955 | -4.6932 | 4.874959 |
| 224 | 1.128774 | 4.335147 | 14.56071 | 1.302985 | 61.85338 | 76.13555 |
| 299 | 2.018827 | 24.72849 | 14.12564 | 1.464285 | 22.82925 | 20.37623 |
| 320 | 2.835917 | 73.80137 | 19.00904 | 1.537626 | 145.1035 | 130.2656 |
| 323 | 2.499499 | 74.39005 | 20.35463 | 1.214407 | 39.19468 | 70.15636 |
| 324 | 1.166263 | 4.847054 | 13.11364 | 1.128108 | 21.57645 | 58.29893 |
| 324as | 0.550056 | -16.3599 | 31.04497 | 0.81003 | -17.5892 | 59.87255 |
| 325 | 0.551765 | -8.93601 | 9.526843 | 0.698675 | -12.5072 | 16.27932 |
| 326 | 1.543886 | 28.18269 | 28.71259 | 1.254235 | 57.5671 | 102.3746 |
| 328 | 0.934291 | -1.96926 | 11.54491 | 0.974026 | -3.73333 | 42.64561 |
| 329 | 1.419212 | 12.11073 | 17.05173 | 0.73368 | -52.9969 | 79.23315 |
| 330 | 2.287859 | 87.81399 | 36.80597 | 2.034871 | 284.7982 | 124.864 |
| 331 | 1.068612 | 2.760884 | 13.90739 | 0.832375 | -28.5962 | 43.38722 |
| 335 | 1.961985 | 23.53498 | 13.34953 | 1.172104 | 23.78697 | 56.22757 |
| 339 | 0.823772 | -7.48749 | 16.46009 | 0.595934 | -66.4477 | 58.00753 |
| 340 | 0.243822 | -9.30408 | 11.01812 | 0.334433 | -11.9408 | 8.519861 |
| 342 | 0.642011 | -16.7282 | 24.98139 | 0.641143 | -70.5241 | 87.04481 |
| 346 | 0.599475 | -18.0394 | 21.16469 | 0.676568 | -27.2487 | 33.21389 |
| 361 | 1.591422 | 24.15604 | 30.11998 | 1.009118 | 1.201702 | 66.22131 |
| 362 | 0.728937 | -8.18092 | 21.78834 | 0.966514 | -4.46931 | 70.94128 |
| 363 | 3.211449 | 4.131684 | 1.996321 | 2.659793 | 10.60853 | 4.807691 |
| 365 | 2.208595 | 21.34172 | 8.250366 | 1.361685 | 42.76415 | 43.11352 |
| 369 | 1.491246 | 41.83633 | 29.23861 | 1.022708 | 5.617521 | 70.09888 |
| 370 | 1.106058 | 6.999861 | 27.0252 | 1.006794 | 1.552185 | 81.38959 |
| 376 | 0.937027 | -3.56189 | 19.48646 | 0.754192 | -25.7478 | 25.73944 |
| 377 | 1.365415 | 33.18509 | 43.36181 | 1.052067 | 14.35203 | 118.5421 |
| 378 | 0.991611 | -0.63453 | 44.44187 | 0.565814 | -168.821 | 156.5515 |
| 379 | 1.010828 | 0.21424 | 11.23191 | 0.921788 | -3.98785 | 20.07679 |
| 380as | 1.024463 | 0.35818 | 6.187144 | 0.606126 | -15.5957 | 15.90185 |
| 381 | 1.530916 | 38.84119 | 24.68669 | 1.217454 | 48.76157 | 70.95493 |
| 382 | 0.896596 | -3.22924 | 28.87611 | 1.128471 | 20.03678 | 90.78311 |
| 383 | 1.846949 | 20.63549 | 11.7632 | 0.960888 | -3.33776 | 35.88878 |
| 384 | 0.715452 | -19.0904 | 28.61347 | 0.712092 | -37.6011 | 42.93263 |
| 409 | 0.84186 | -8.82878 | 22.68341 | 1.620843 | 83.88515 | 56.24362 |
| 409as | 1.658425 | 10.32247 | 6.676508 | 1.027725 | 0.89023 | 13.82001 |
| 410 | 0.605117 | -31.9761 | 26.42415 | 0.722741 | -56.3923 | 58.94433 |
| 423 | 2.233948 | 2.76181 | 1.577606 | 1.381155 | 2.483714 | 3.648045 |
| 431 | 1.08339 | 2.155206 | 9.318698 | 0.658931 | -37.2679 | 35.1268 |
| 433 | 1.523432 | 25.42546 | 21.15427 | 1.367604 | 44.08222 | 43.40362 |
| 448 | 3.817645 | 112.9229 | 32.10091 | 1.337718 | 54.531 | 81.54711 |
| 450 | 2.496269 | 1.798206 | 0.948385 | 0.566819 | -2.2927 | 2.648492 |
| 451 | 1.673375 | 1.207216 | 1.405598 | 2.083019 | 6.759058 | 2.857589 |
| 485 | 1.952661 | 60.49692 | 49.4361 | 1.316979 | 40.67605 | 62.97121 |
| 485as | 1.263134 | 12.70744 | 24.51362 | 1.155575 | 39.31194 | 96.41782 |
| 487as | 2.249755 | 1.666521 | 1.508161 | 2.691645 | 8.798719 | 3.91008 |
| 488 | 0.754737 | -9.74893 | 14.40312 | 0.952448 | -4.89276 | 26.90241 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 490 | 0.965238 | -0.64826 | 7.81248 | 0.727792 | -45.2562 | 57.78719 |
| 494 | 0.851054 | -19.6015 | 59.86998 | 0.942886 | -16.476 | 126.0441 |
| 495 | 0.925233 | -7.11118 | 42.87663 | 0.627485 | -241.027 | 312.0291 |
| 496 | 1.516907 | 52.81838 | 39.56634 | 1.007484 | 1.455893 | 63.29368 |
| 499 | 1.110699 | 5.97996 | 37.56199 | 0.787525 | -26.4406 | 58.93389 |
| 503 | 8.633664 | 13.26262 | 1.584616 | 1.697132 | 13.9662 | 13.06149 |
| 504 | 1.105138 | 3.329739 | 12.29496 | 0.748723 | -47.6563 | 63.41167 |
| 505 | 1.18359 | 10.39258 | 26.51655 | 1.179383 | 24.79217 | 52.40033 |
| 539 | 0.713304 | -49.0351 | 92.72268 | 0.793274 | -74.0099 | 189.397 |

## Figure Legends

**Figure 1.** Hierarchy of miRNA complementary sites

(A) Unrooted phylogenetic tree displaying the branch lengths calculated by maximum likelihood for all 7mer sequences in 3' UTRs. Only branch lengths of at least 0.01 were displayed.

(B) Signal to background analysis of conserved sites. MicroRNA sites with conserved branch lengths of at least 0.6746 (the minimum required to span HMRD, left panel) or 1.1390 (the minimum required to span HMRD + chicken, right panel) are shown as blue bars; conserved matches to 50 sets of control sequences are shown at white bars, with the error bars indicating one standard deviation. The bars contain non-overlapping data, so 8mer sites are excluded from the 7mer-m8 and 7mer-A1 analyses.

(C) Effectiveness of single canonical sites. Changes in abundance of mRNAs following miRNA transfection were monitored with microarrays. Distributions of changes (0.1 unit bins) for messages containing the indicated single sites in their UTRs are shown (left), together with the cumulative distributions (right). Results of eleven experiments, each performed in duplicate and each transfecting a duplex for a different miRNA, were consolidated.

**Figure 2.** Efficacy of sites with imperfect seed pairing

(A) Signal to background analysis of conserved imperfect sites. Imperfect miRNA sites with conserved branch lengths of at least 0.6746 (HMRD) are shown as blue bars; conserved matches to 50 sets of control sequences are shown at white bars, with the error bars indicating one standard deviation. The leftmost pair of bars defines imperfect seeds in the most flexible manner and includes insertions, mismatches, and GU-wobbles, while the bars displaying the mismatched sites and GU-wobble sites are mutually exclusive.

(B) Cumulative distribution of changes in abundance of mRNAs possessing 8mer sites with single mismatches to the seed following miRNA transfection. Changes were monitored with microarrays and are displayed with the cumulative distribution as in Figure 1D, including for reference the results of canonical 6mer and 8mer sites.

(C) As in (B) except the sites had single G:U wobble mismatches to the seed.

(D) As in (B) except the sites were offset 6mers remaining after removing messages with canonical 7mers.

**Figure 3**. Target analysis for vertebrate and mammalian-only miRNA families

(A) Signal to background analysis of conserved 7mer and 8mer sites. MicroRNA sites with conserved branch lengths of at least 0.6746 (HMRD) are shown as blue bars; conserved matches to 50 sets of control sequences are shown at white bars, with the error bars indicating one standard deviation.

(B) As in (A), except miRNA sites conserved to chicken or opossum have been excluded from the analysis, and only genes with 3' UTR sequences aligned to chicken and opossum are considered.

**Figure 4.** Site effectiveness is a function of conservation of the site and its 3' UTR sequence

(A) Improved efficacy of conserved sites. 7mer-m8 sites were partitioned by branch length into those with no conservation (branch length = 0), weak conservation (0 < branch length < 0.6746), conservation among HMRD (0.6746 ≤ branch length < 1.1390), or conservation among HMRD+chicken (branch length ≥ 1.1390). Cumulative distributions of changes in mRNA levels following miRNA transfection were plotted for the sites in each of these categories.

(B) Conserved sites are more likely to be effective in the context of weak background 3' UTR conservation. The cumulative distributions of changes in mRNA levels was plotted for sites falling short of the criteria of conservation to human, mouse, rat, and dog. Sites meeting the conservation criteria were divided into two evenly sized groups based on the percentage of conserved 7mers in the UTR, again requiring conservation to human, mouse, rat, and dog.) Conserved sites within less well-conserved UTRs were significantly more effective than conserved sites within more well-conserved UTRs.

**Figure 5.** Sites with higher confidence of being under selection are more efficacious

(A) 2-D contour map displaying the signal to background values for miRNA 8mer sites as a function of a site's conserved branch length and the average conserved branch length of its UTR. Each contour line is labeled with the signal to background value at that threshold.

(B) Histogram showing the bimodal distribution of confidence scores for miRNA 8mer sites.

(C) Linear regression, using confidence scores as a predictor of site efficacy. Site efficacy and confidence scores follow a linear relationship, r = -0.1689, $P < 10^{-20}$, Pearson correlation. Sites with non-zero confidence scores were placed into 20 equally sized bins based on their confidence score, with individual points on the plot representing the average repression and confidence score at each bin. Sites with confidence scores of 0 were binned together and their average repression plotted as the red data point.

## Figure 1

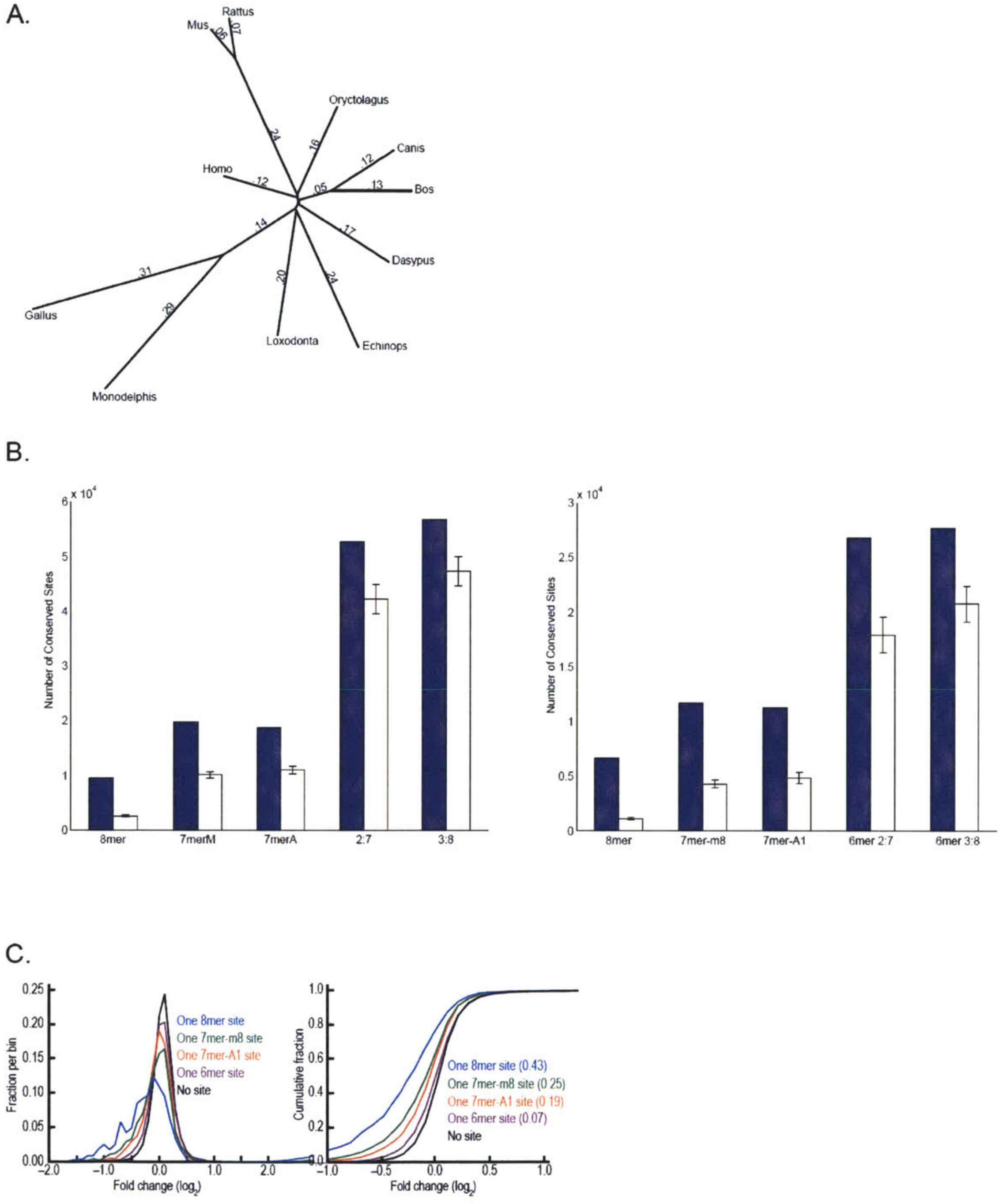

A.
Rattus
Mus
.06
.07
.24
Oryctolagus
.16
Canis
.12
Homo
.12
.05
.13
Bos
.14
.17
Dasypus
.31
Gallus
.29
.20
.24
Loxodonta
Echinops
Monodelphis
B.
x 10^4
Number of Conserved Sites
0
1
2
3
4
5
6
8mer
7merM
7merA
2:7
3:8
x 10^4
Number of Conserved Sites
0
0.5
1
1.5
2
2.5
3
8mer
7mer-m8
7mer-A1
6mer 2:7
6mer 3:8
C.
Fraction per bin
0.00
0.05
0.10
0.15
0.20
0.25
−2.0
−1.0
0.0
1.0
2.0
Fold change (log2)
One 8mer site
One 7mer-m8 site
One 7mer-A1 site
One 6mer site
No site
Cumulative fraction
0.0
0.2
0.4
0.6
0.8
1.0
−1.0
−0.5
0.0
0.5
1.0
Fold change (log2)
One 8mer site (0.43)
One 7mer-m8 site (0.25)
One 7mer-A1 site (0.19)
One 6mer site (0.07)
No site

Figure 2

**Figure 3**

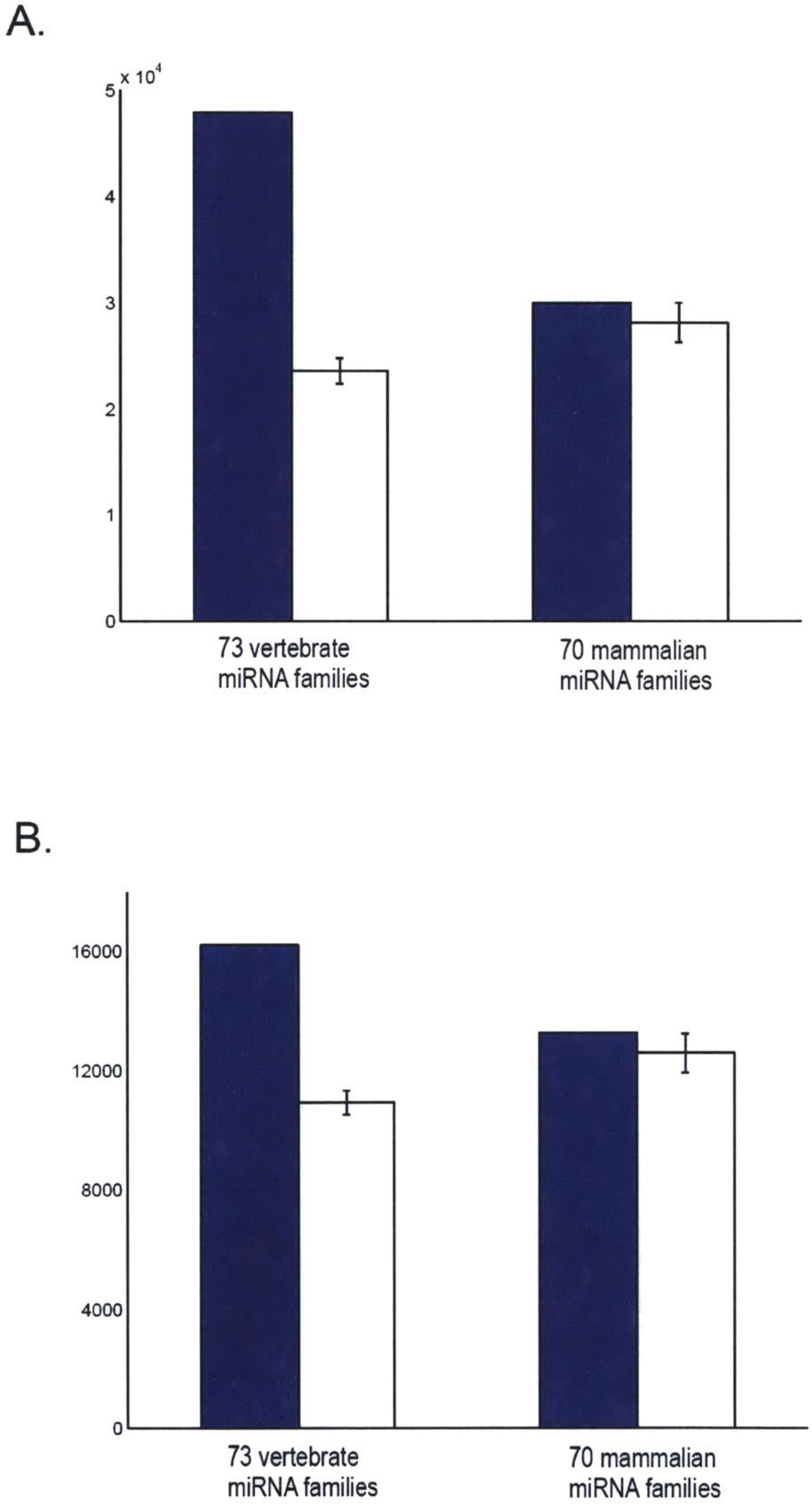
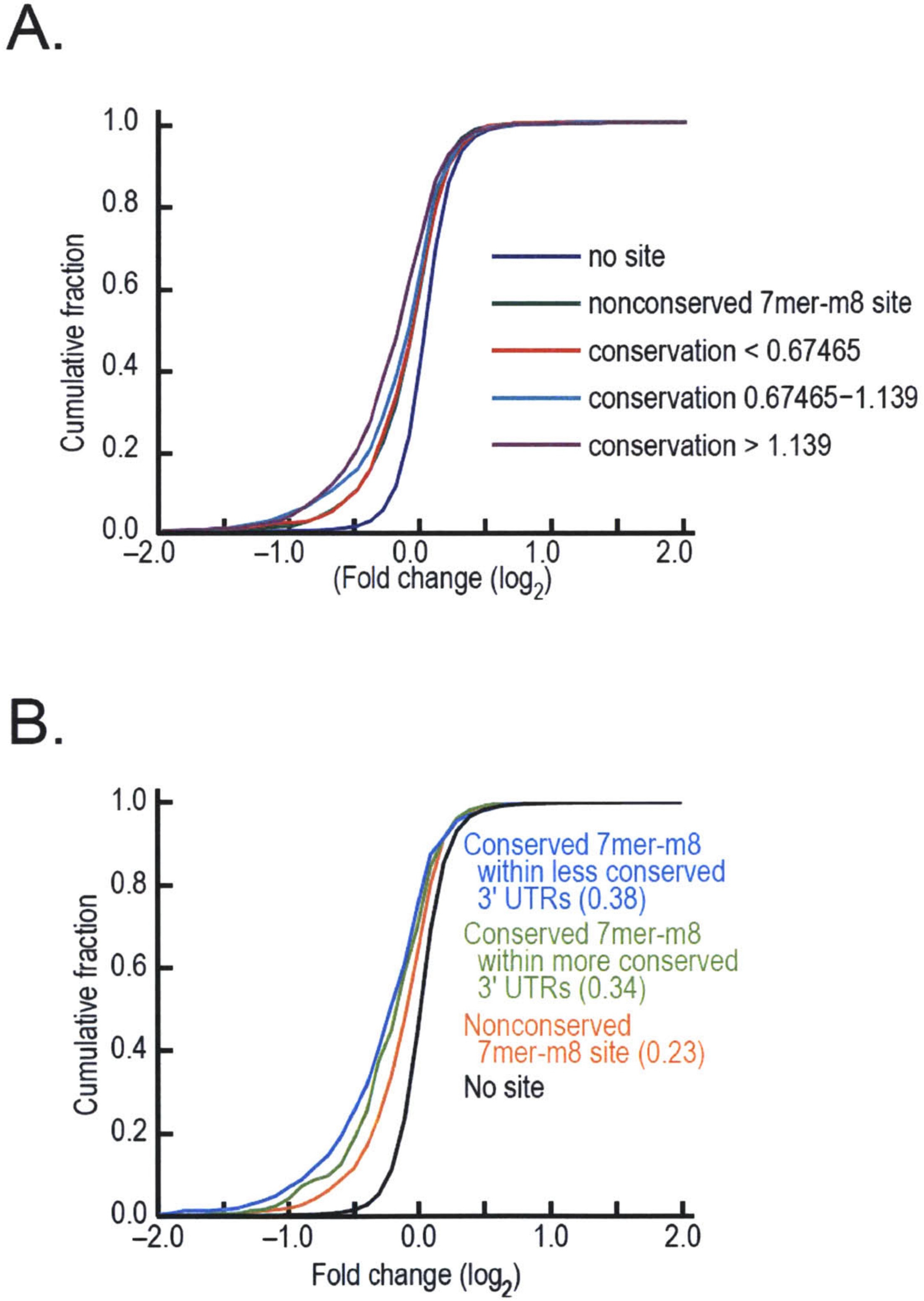
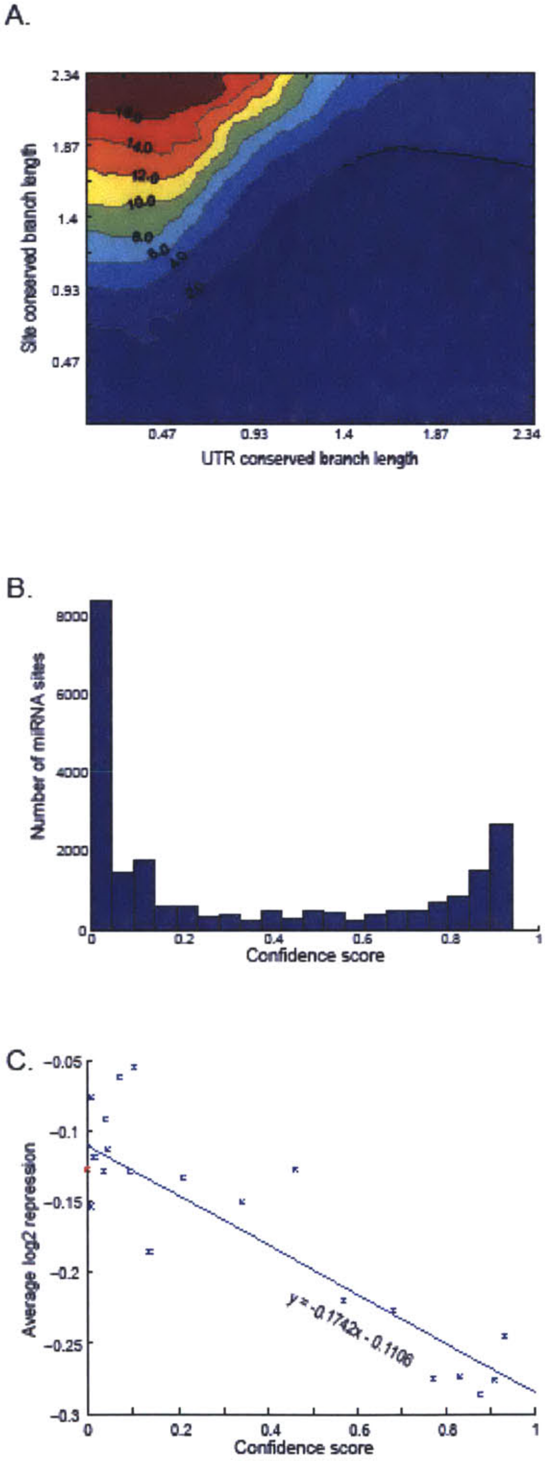

A.
x 10^4
5
4
3
2
1
0
73 vertebrate miRNA families
70 mammalian miRNA families
B.
16000
12000
8000
4000
0
73 vertebrate miRNA families
70 mammalian miRNA families

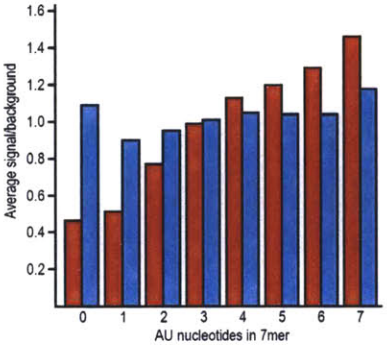
**Figure 4**


A.
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative fraction
–2.0
–1.0
0.0
1.0
2.0
(Fold change (log$_2$)
no site
nonconserved 7mer-m8 site
conservation < 0.67465
conservation 0.67465−1.139
conservation > 1.139
B.
1.0
0.8
0.6
0.4
0.2
0.0
Cumulative fraction
–2.0
–1.0
0.0
1.0
2.0
Fold change (log$_2$)
Conserved 7mer-m8 within less conserved 3' UTRs (0.38)
Conserved 7mer-m8 within more conserved 3' UTRs (0.34)
Nonconserved 7mer-m8 site (0.23)
No site

# Figure 5


A.
2.34
1.87
1.4
0.93
0.47
Site conserved branch length
16.0
14.0
12.0
10.0
8.0
6.0
4.0
2.0
0.47
0.93
1.4
1.87
2.34
UTR conserved branch length
B.
8000
6000
4000
2000
0
Number of miRNA sites
0
0.2
0.4
0.6
0.8
1
Confidence score
C.
-0.05
-0.1
-0.15
-0.2
-0.25
-0.3
Average log2 repression
y = -0.1742x - 0.1106
0
0.2
0.4
0.6
0.8
1
Confidence score

# Supplementary Figure S1



**Supplemental Figure 1.** The importance of considering nucleotide composition when selecting control sequences. All 16,384 possible 7mers were binned according to their AU content. Preferential conservation was analyzed as for Figure 1C, using either observed 7mer counts for selecting control sequences (red) or our current method (blue). Results for all 7mers with the same AU content were combined, and the average signal-to-background ratio reported. The controls from the current method gave signal/background values approaching 1.0 for all compositions, whereas the other method gave artifactually high ratios for AU-rich 7mers and artifactually low ratios for GC-rich 7mers.

# Future Directions

## Improvements to Targeting Algorithms Based on mRNA Structure

Several groups have proposed mRNA secondary structure and effects on site accessibility as important determinants of microRNA targeting (Long et al., 2007; Robins et al., 2005; Zhao et al., 2005), although when implementing these algorithms, we found them to perform no better than simply tallying the local AU nucleotide composition around the seed site, weighted by the inverse of the distance from the site (Grimson et al., 2007). As described before, this strong preference for AU rich composition around functional microRNA sites may be due to a nucleotide preference of the RISC itself, or more likely, the weaker secondary structure and consequential greater site-accessibility associated with higher AU composition. Decreased site accessibility due to the interference of the translating ribosome appears to explain why microRNA sites in the ORF tend to be less functional than equivalent sites in the 3' UTR. This effect extends into the first ~15 nucleotides of the 3' UTR, a figure that matches well with the biochemical and structural understanding of the translating ribosome (Grimson et al., 2007; Takyar et al., 2005; Yusupova et al., 2001).

Although AU composition was by far the strongest determinant we identified outside of seed motif type, it can likely be improved by a more accurate representation of mRNA structure and site accessibility. The lack of success thus far in predicting effective microRNA targets using RNA secondary structure folding algorithms may be explained in one of three ways: the algorithms are insufficient because they do not capture tertiary mRNA structure; or there may be a lack of ordered RNA structure in the relatively random mRNA sequences; or the mRNA structures may be significantly altered by the effect of RNA-binding proteins. RNA-binding proteins could potentially affect microRNA function either directly via protein-protein interactions, or indirectly by altering mRNA structure.

The majority of mRNA 3' UTR sequence is thought to be under neutral selection, and could be expected to evolve in a largely random fashion. Over the course of this random walk, a given mRNA may readily sample a variety of secondary structures (Schuster et al., 1994), potentially leading to certain regions of mRNA becoming more open and accessible to the action of RNA-binding proteins or microRNAs, while other regions become less accessible. Such structural patterns might have initially occurred out of happenstance, as functional and non-functional microRNA sites accumulated in the 3' UTR, but over time selective pressure would maintain structural configurations that favored accessibility to conserved sites and denied accessibility to potentially deleterious seed sites that had accumulated in less accessible regions. The existence of such highly structured 3' UTRs is suggested by the recent discovery of a hammerhead ribozyme in the Clec2 3' UTR, whose domains are separated by hundreds of base pairs, but appear to be brought together during in vivo and in vitro cleavage events (Martick et al., 2008). Structural constraints supporting conserved microRNA target sites may also partially explain the islands of conservation seen in the comparative genome analyses, where conservation extends beyond the 7-8 nt microRNA seed sites to adjacent contiguous stretches of 3' UTR (Lewis et al., 2005). For a consistently structured mRNA, different regions of the mRNA should be more or less responsive to microRNAs sites within those regions. One way to test this hypothesis would be to examine large quantities of microRNA and siRNA transfection microarray data to map out each gene's 3' UTR and determine regions of mRNA sequence that contain clusters of sites with high or low efficacy (Grimson et al., 2007; Jackson et al., 2003; Lim et al., 2005). These maps could then be evaluated to determine if these results provide any additional information for predicting microRNA site efficacy beyond local AU composition and predicted secondary structure. Because the distance of local AU effects appears to be no more than about 30 nucleotides upstream and downstream around the seed site and the probability of matching either microRNA 7mer is roughly 1/8192 (depending largely on dinucleotide composition), over a hundred transfections might be required to obtain sufficient resolution. Other clues to mRNA structure could also be revealed using large-scale sequencing and mRNA degradome analysis, a method that has been successfully employed for identification of plant microRNA cleavage targets (Addo-Quaye et al., 2008; German et al., 2008); perhaps the rate of mRNA degradation could be affected in the vicinity of structural motifs or RNA-binding proteins, leading to an enrichment of reads near these regions. Besides potentially improving microRNA predictions, investigations along these lines could also provide a glimpse into the relatively uncharacterized behavior of mRNAs in vivo. Alternatively, other mRNAs, being largely composed of random sequence, may rarely hold to stable structures and instead prefer to explore multiple competing conformations (Schultes et al., 2005). Should this be the case, mapping out the accessibility of different regions of a highly fluid mRNA might be an exercise in futility. Different mRNAs might fall at different places in the spectrum between having a well-defined structure and having fluid competing configurations, depending on UTR length, GC content, RNA-binding proteins, the presence of other structural motifs, and the effects of selection.

## Interactions with Other Modes of Posttranscriptional Regulation

Although there have been studies suggesting that mRNA-binding proteins interact with microRNAs via AU-rich elements to modulate transcript stability (Jing et al., 2005), there has been no evidence that this is a systematic effect. Furthermore, there were no other motifs that were consistently found to be associated with downregulation on the array in the vicinity of the microRNA seed motif (Grimson et al., 2007). Out of the hundreds of known RNA-binding proteins, relatively few have characterized motifs, and of these, their motifs vary greatly in specificity. One of the more specific motifs appears to be the UGUAAAUA motif recognized by the PUMILIO family of proteins, which appears to be sufficient for PUMILIO-mediated posttranscriptional downregulation of a target mRNA (Gerber et al., 2006). Aside from the target sites of microRNAs themselves, motifs of this length and quality have not appeared in analyses searching for motifs that are systematically associated with upregulation or downregulation in microRNA transfection experiments (Grimson et al., 2007; Sood et al., 2006). While motifs such as PUMILIO may generally be unrelated to microRNA targeting, smaller, less specific motifs, might have been missed by these analyses. Compared to PUMILIO, where much of the specificity could be accounted for by the motif sequence itself, smaller less specific motifs would presumably rely on greater context dependence. One such motif, the AU-rich element, is known to have roles in posttranscriptional mRNA stability, despite relatively poor characterization of the binding motif, and may cross-talk directly with the AGO2 silencing system (Vasudevan and Steitz, 2007). Intriguingly, the study by Sood *et al.* identified smaller and less-specific motifs similar to the AU-rich element (AUUUA) that were associated with decreased repression on microRNA transfection of a couple of microRNAs (Sood et al., 2006); however, these results did not appear to generalize in a larger dataset of microRNA transfections (Grimson et al., 2007). Another class of RNA binding motifs whose role has yet to be fully explored includes proteins which may recognize their targets primarily on the basis of secondary structure rather than primary sequence. The recognition motif of the Hu family (HuF) of RNA-binding proteins, which have known roles in the posttranscriptional stabilization of cell cycle genes, appears to consist of a stem loop with an enrichment for uridine residues, rather than for any specific primary sequence (Lopez de Silanes et al., 2004). The interactions between different mRNA-binding proteins and microRNA regulation clearly await further investigation.

## MicroRNA Regulation and Gene Networks

With the recognition that microRNAs affected a large fraction of the human genome, there has been great interest in investigating the role of microRNAs in the greater context of gene regulatory networks and how they interact with transcription factors to control cell processes and differentiation. In a study by Marson *et al*, key ES cell transcription factors were found to be associated with the promoters of a large number of microRNAs, both to a set of ES cell expressed microRNAs, as well as to a set of ES cell non-expressed microRNAs (Marson et al., 2008). However, the promoters of these silent microRNAs were co-occupied by Polycomb group proteins, preventing their transcription during the ES cell state. With differentiation, these formerly silent microRNAs were then expressed in a tissue-specific manner. Conversely, more effort will also need to be directed towards understanding the effects of microRNAs on transcriptional programs (Tsang et al., 2007). The onset of expression of a microRNA, whether endogenously or through transfection, produces a large number of changes in gene expression (Giraldez et al., 2006; Lim et al., 2005). While the downregulated genes tend to be highly enriched in microRNA seed sites, the majority of changing genes do not appear to be directly regulated by the microRNA. These secondary targets are presumably due to effects of the microRNA on transcription factor targets and other cellular regulators, and the foldchanges associated with these secondary targets often dwarfs the modest 20-30% change in expression for a typical direct microRNA target. For instance, a study on the mouse knockout of miR-223 focused on the Mef2c transcription factor as the major target of miR-223 (Johnnidis et al., 2008). Out of the hundreds of conserved miR-223 targets, Mef2c was one of the most upregulated, increasing its expression twofold in the miR-223 null animal. In comparison, the c-Jun transcription factor, which contained no target sites to miR-223, experienced one of the most dramatic changes on the array, with its mRNA levels more than quadrupling in the miR-223 knockout compared to wild type. While several studies have shown that the Mef2c positively regulates c-Jun expression, it remains unclear how much of the resulting phenotype could be attributed to just one of the hundreds of miR-223 conserved and nonconserved targets that were affected by knocking out the microRNA (Schuler et al., 2008; Wei et al., 2003). This underscores the difficulty in sorting through the secondary effects of microRNA targeting and the challenges in understanding gene regulation networks in general.

### Few Targets, Many Targets

Compared with just a few short ago, when dozens of new microRNAs were being discovered and only a handful of targets were known, the success of computational and genome-scale methods for finding microRNA targets has reversed the state of the field. Biologists are now confronted with the problem of having too many microRNA targets to choose from. One approach has been to disregard to some extent the large number of number of targets per microRNA, and just focus on a single target selected by a combination of the quality and conservation of the site and intuition (Johnnidis et al., 2008; Zhao et al., 2005). However, the microRNAs used in these studies (miR-1, miR-223) are deeply conserved vertebrate microRNAs, each with over a hundred conserved targets above chance. It seems unlikely that a knockout phenotype could be explained away by a single microRNA-target interaction.

One area where reducing the number of sites might prove to be a particularly effective strategy may be with mammalian-only microRNAs. The set of microRNAs conserved only among mammals has a very poor signal-to-noise ratio, indicating that they only have a handful of conserved targets compared with the more deeply conserved vertebrate microRNA set (Friedman et al., 2008). As a consequence of the low signal-to-background ratio of mammalian-only microRNAs, conserved targets of these microRNAs are difficult to distinguish from conservation due to chance, even when considering 8mer sites. These less well-conserved microRNAs were also cloned at lower frequency than the more highly conserved vertebrate microRNAs, suggesting that they were either expressed at lower levels or in a smaller population of cells (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Furthermore, compared to the more deeply conserved vertebrate microRNAs, mammalian-only microRNAs have acquired fewer conserved targets per microRNA since the divergence of mammals from other vertebrates (Friedman et al., 2008). For these weakly expressed microRNAs, different rules might apply. For instance, in *C. elegans*, the *lsy-6* microRNA, which is only expressed in a handful of neurons, controls left and right neuronal patterning through its target *cog-1*, which maintains two *lsy-6* seed sites, both with 3' pairing, spaced 34 nucleotides apart (Johnston and Hobert, 2003). Extensive experimentation with *lsy-6* and *cog-1* has shown that typical 7mer seed sites to *lsy-6* are not downregulated effectively, and that *cog-1* can still be regulated by *lsy-6* despite mismatches in its seed sequences (Didiano and Hobert, 2006). The weakly conserved and weakly expressed mammalian microRNAs might follow similar rules, with each microRNA targeting a handful of genes with the most optimal target sites in a few cell types, and having negligible effect on a typical 7mer. It will be very interesting to compare the phenotypes of mammalian-specific microRNA knockouts with the phenotypes of more deeply conserved microRNAs, and to determine if the mammalian-specific microRNA knockout phenotypes can be explained by the derepression of only one or two genes.

For the more deeply conserved microRNAs, it has been challenging to reconcile the large numbers and high reliability of microRNA targeting with the relatively modest repression typically observed in reporter assays and on microarrays. The fact that each vertebrate microRNA averages hundreds of conserved sites over chance – the majority of which can be expected to confer 30% repression or less – is a testament to the pervasiveness and subtlety of selection (Grimson et al., 2007). Although the majority of mammalian protein coding genes are microRNA targets, many of these target sites are relatively weak 6mer and offset 6mer sites with low signal-to-background and low efficacy when measured by microarray and quantitative mass spectrometry (Baek et al., 2008; Friedman et al., 2008; Grimson et al., 2007; Selbach et al., 2008). Adding in the hundreds of genes actively avoiding sites to that microRNA (Farh et al., 2005), many of these interactions and the reasons why nature originally selected for them and still maintains them today seem mysterious and seem likely to remain so because of their subtlety. Our greater understanding of microRNA targeting has also yielded insights into how the complex web of gene regulation has evolved and functions in living organisms. Although the seed sequences of many animal microRNAs have remained nearly constant across worms, flies, fish, and mammals, their targets have almost completely diverged (Chen and Rajewsky, 2006). Among the deeply conserved vertebrate microRNAs present in both pufferfish and human, the sequences of the microRNAs themselves have changed minimally or not at all, yet only a quarter of the original TargetScan microRNA targets remain conserved (Lewis et al., 2003). The recent characterization of microRNA sequences and protein machinery in basal metazoan lineages has shown that proteins such as Drosha and Dicer were present prior to the divergence of *Amphimedon*, the furthest diverged extant animal lineage (Grimson et al., 2008). However, *Amphimedon* shares no microRNAs in common with higher invertebrates, and has pre-microRNA lengths vastly different from the tightly regulated ~60-70 nt pre-microRNAs of fly, worm, and human. *Nematostella*, in comparison, had one microRNA, miR-100, in common with higher invertebrates and vertebrates, although it appeared to be frame-shifted one nucleotide. Assuming that microRNAs in Nematostella also recognize their targets via nucleotides 2-8, this 1-nt shift would significantly alter its seed-match targets, although it is plausible that in this primitive metazoan, the seed-match register might also be shifted. In contrast, despite the divergence of worm, fly, and vertebrate microRNA targets, the sequences of the conserved microRNAs and the principles by which they recognize their targets are largely constant, implying that the functions of the core microRNA machinery has become more fixed. Charting the course of microRNA evolution, selection appears to have operated in a step-wise fashion, with new mechanisms of regulation proliferating and fluidly evolving even as older mechanisms become fixed.

## References

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., and Axtell, M. J. (2008). Endogenous siRNA and miRNA targets identified by sequencing of the Arabidopsis degradome. Curr Biol *18*, 758-762.

Baek, D., Villen, J., Shin, C., Camargo, F. D., Gygi, S. P., and Bartel, D. P. (2008). The impact of microRNAs on protein output. Nature *455*, 64-71.

Chen, K., and Rajewsky, N. (2006). Deep conservation of microRNA-target relationships and 3'UTR motifs in vertebrates, flies, and nematodes. Cold Spring Harb Symp Quant Biol *71*, 149-156.

Didiano, D., and Hobert, O. (2006). Perfect seed pairing is not a generally reliable predictor for miRNA-target interactions. Nat Struct Mol Biol *13*, 849-851.

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.

Friedman, R. C., Farh, K. K., Burge, C. B., and Bartel, D. P. (2008). Most mammalian mRNAs are conserved targets of microRNAs. Genome Res.

Gerber, A. P., Luschnig, S., Krasnow, M. A., Brown, P. O., and Herschlag, D. (2006). Genome-wide identification of mRNAs associated with the translational regulator PUMILIO in Drosophila melanogaster. Proc Natl Acad Sci U S A *103*, 4487-4492.

German, M. A., Pillay, M., Jeong, D. H., Hetawal, A., Luo, S., Janardhanan, P., Kannan, V., Rymarquis, L. A., Nobuta, K., German, R., *et al.* (2008). Global identification of microRNA-target RNA pairs by parallel analysis of RNA ends. Nat Biotechnol *26*, 941-946.

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.

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

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., Degnan, B. M., Rokhsar, D. S., and Bartel, D. P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. Nature *455*, 1193-1197.

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.

Jing, Q., Huang, S., Guth, S., Zarubin, T., Motoyama, A., Chen, J., Di Padova, F., Lin, S. C., Gram, H., and Han, J. (2005). Involvement of microRNA in AU-rich element-mediated mRNA instability. Cell *120*, 623-634.

Johnnidis, J. B., Harris, M. H., Wheeler, R. T., Stehling-Sun, S., Lam, M. H., Kirak, O., Brummelkamp, T. R., Fleming, M. D., and Camargo, F. D. (2008). Regulation of progenitor cell proliferation and granulocyte function by microRNA-223. Nature *451*, 1125-1129.

Johnston, R. J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. Nature *426*, 845-849.

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

Lau, N. C., Lim, L. P., Weinstein, E. G., and Bartel, D. P. (2001). An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science *294*, 858-862.

Lee, R. C., and Ambros, V. (2001). An extensive class of small RNAs in *Caenorhabditis elegans*. Science *294*, 862-864.

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.

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.

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.

Long, D., Lee, R., Williams, P., Chan, C. Y., Ambros, V., and Ding, Y. (2007). Potent effect of target structure on microRNA function. Nat Struct Mol Biol *14*, 287-294.

Lopez de Silanes, I., Zhan, M., Lal, A., Yang, X., and Gorospe, M. (2004). Identification of a target RNA motif for RNA-binding protein HuR. Proc Natl Acad Sci U S A *101*, 2987-2992.

Marson, A., Levine, S. S., Cole, M. F., Frampton, G. M., Brambrink, T., Johnstone, S., Guenther, M. G., Johnston, W. K., Wernig, M., Newman, J., *et al.* (2008). Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells. Cell *134*, 521-533.

Martick, M., Horan, L. H., Noller, H. F., and Scott, W. G. (2008). A discontinuous hammerhead ribozyme embedded in a mammalian messenger RNA. Nature *454*, 899-902.

Robins, H., Li, Y., and Padgett, R. W. (2005). Incorporating structure to predict microRNA targets. Proc Natl Acad Sci U S A *102*, 4006-4009.

Schuler, A., Schwieger, M., Engelmann, A., Weber, K., Horn, S., Muller, U., Arnold, M. A., Olson, E. N., and Stocking, C. (2008). The MADS transcription factor Mef2c is a pivotal modulator of myeloid cell fate. Blood *111*, 4532-4541.

Schultes, E. A., Spasic, A., Mohanty, U., and Bartel, D. P. (2005). Compact and ordered collapse of randomly generated RNA sequences. Nat Struct Mol Biol *12*, 1130-1136.

Schuster, P., Fontana, W., Stadler, P. F., and Hofacker, I. L. (1994). From sequences to shapes and back: a case study in RNA secondary structures. Proc Biol Sci *255*, 279-284.

Selbach, M., Schwanhausser, B., Thierfelder, N., Fang, Z., Khanin, R., and Rajewsky, N. (2008). Widespread changes in protein synthesis induced by microRNAs. Nature *455*, 58-63.

Sood, P., Krek, A., Zavolan, M., Macino, G., and Rajewsky, N. (2006). Cell-type-specific signatures of microRNAs on target mRNA expression. Proc Natl Acad Sci U S A *103*, 2746-2751.

Takyar, S., Hickerson, R. P., and Noller, H. F. (2005). mRNA helicase activity of the ribosome. Cell *120*, 49-58.

Tsang, J., Zhu, J., and van Oudenaarden, A. (2007). MicroRNA-mediated feedback and feedforward loops are recurrent network motifs in mammals. Mol Cell *26*, 753-767.

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

Wei, X., Sun, W., Fan, R., Hahn, J., Joetham, A., Li, G., Webb, S., Garrington, T., Dakhama, A., Lucas, J., *et al.* (2003). MEF2C regulates c-Jun but not TNF-alpha gene expression in stimulated mast cells. Eur J Immunol *33*, 2903-2909.

Yusupova, G. Z., Yusupov, M. M., Cate, J. H., and Noller, H. F. (2001). The path of messenger RNA through the ribosome. Cell *106*, 233-241.

Zhao, Y., Samal, E., and Srivastava, D. (2005). Serum response factor regulates a muscle-specific microRNA that targets *Hand2* during cardiogenesis. Nature *436*, 214-220.