Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs


*Genome Res.* 2011 21: 1395-1403 originally published online June 17, 2011
Access the most recent version at doi:10.1101/gr.121210.111

**Supplemental Material**

http://genome.cshlp.org/content/suppl/2011/06/15/gr.121210.111.DC1.html

**References**

This article cites 48 articles, 19 of which can be accessed free at:
http://genome.cshlp.org/content/21/9/1395.full.html#ref-list-1

Article cited in:
http://genome.cshlp.org/content/21/9/1395.full.html#related-urls

**Open Access**

Freely available online through the *Genome Research* Open Access option.

This article is distributed exclusively by Cold Spring Harbor Laboratory Press for the first six months after the full-issue publication date (see http://genome.cshlp.org/site/misc/terms.xhtml). After six months, it is available under a Creative Commons License (Attribution-NonCommercial 3.0 Unported License), as described at http://creativecommons.org/licenses/by-nc/3.0/.

**Email alerting service**

Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

To subscribe to *Genome Research* go to:
http://genome.cshlp.org/subscriptions

© 2011, Published by Cold Spring Harbor Laboratory Press
Unusually effective microRNA targeting within repeat-rich coding regions of mammalian mRNAs

Michael Schnall-Levin,1,2,8 Olivia S. Rissland,3,4,5,8 Wendy K. Johnston,3,4,5 Norbert Perrimon,6,7 David P. Bartel,3,4,5,9 and Bonnie Berger1,2,9

1Department of Mathematics, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; 2Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; 3Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02139, USA; 4Howard Hughes Medical Institute, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; 5Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA; 6Howard Hughes Medical Institute, Harvard Medical School, Boston, Massachusetts 02115, USA; 7Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; 8Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA; 9Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA

MicroRNAs (miRNAs) are ~22-nucleotide RNAs that direct the post-transcriptional repression of target messenger RNAs (mRNAs) (Bartel 2009; Ghildiyal and Zamore 2009). An important component of gene regulation in higher eukaryotes, miRNAs are bound by the effector protein Argonaute to form the mature silencing complex. Watson-Crick base-pairing between the 5¢ effector protein Argonaute to form the mature silencing complex.

MicroRNAs (miRNAs) regulate numerous biological processes by base-pairing with target messenger RNAs (mRNAs), primarily through sites in 3¢ untranslated regions (UTRs), to direct the repression of these targets. Although miRNAs have sometimes been observed to target genes through sites in open reading frames (ORFs), large-scale studies have shown such targeting to be generally less effective than 3¢ UTR targeting. Here, we show that several miRNAs each target significant groups of genes through multiple sites within their coding regions. This ORF targeting, which mediates both predictable and effective repression, arises from highly repeated sequences containing miRNA target sites. We show that such sequence repeats largely arise through evolutionary duplications and occur particularly frequently within families of paralogous C2H2 zinc-finger genes, suggesting the potential for their coordinated regulation. Examples of ORFs targeted by miR-181 include both the well-known tumor suppressor RB1 and RBAP5, encoding a C2H2 zinc-finger protein and transcriptional binding partner of RB1. Our results indicate a function for repeat-rich coding sequences in mediating post-transcriptional regulation and reveal circumstances in which miRNA-mediated repression through ORF sites can be reliably predicted.

[Supplemental material is available for this article.]

MicroRNAs (miRNAs) are ~22-nucleotide RNAs that direct the post-transcriptional repression of target messenger RNAs (mRNAs) (Bartel 2009; Ghildiyal and Zamore 2009). An important component of gene regulation in higher eukaryotes, miRNAs are bound by the effector protein Argonaute to form the mature silencing complex. Watson-Crick base-pairing between the 5¢ end of the miRNA—the so-called “seed” region—and a target message elicits silencing of the target mRNA primarily through miRNA decay (Bartel 2009; Guo et al. 2010). Target sites can be grouped by the extent to which they match the region of the miRNA seed (nucleotides 2–7). The weakest site, with base-pairing to these six nucleotides (6mer site), usually conveys only mild repression and is frequently augmented in sites conferring substantial down-regulation. Those with an adenosine opposite nucleotide 1 (7mer-A1 site) generally confer more repression, followed by those with base-pairing to nucleotide 8 of the miRNA (7mer-m8 site), followed by those with both 8mer site (Lewis et al. 2005; Grimson et al. 2007). Other context factors, such as local AU-content, also influence the repression mediated by an individual site (Grimson et al. 2007; Niesben et al. 2007).

The majority of characterized miRNA target sites are in the 3¢ untranslated regions (UTRs) of mRNAs, and large-scale studies examining the effects of introducing or deleting a miRNA have shown that sites in 3¢ UTRs generally are more effective than those in either 5¢ UTRs or open reading frames (ORFs) (Bartel 2009). The reduced efficacy of 5¢ UTR and ORF sites is attributed to displacement of the miRNA silencing complex within the 5¢ UTR by the scanning machinery, as it passes from the cap to the start codon, and within the coding region by translocating ribosomes (Grimson et al. 2007; Gu et al. 2009). Supporting this model, 3¢ UTR sites that lie within ~15 nucleotides of the stop codon are no more effective than ORF sites, as expected if the silencing complex were displaced by the ribosome leading edge as the stop codon approached the ribosome A site (Grimson et al. 2007).

Although ORF sites generally are less effective, enough ORF sites mediate repression to observe a signal above background in large-scale functional studies (Lim et al. 2005; Grimson et al. 2007; Baek et al. 2008; Selbach et al. 2008), and even more sites appear to bind the silencing complex sufficiently to mediate enrichment of the mRNA (or a cross-linked fragment of the mRNA) during immunoprecipitation of the silencing complex (Easow et al. 2007; Hendrickson et al. 2008; Chi et al. 2009; Hafner et al. 2010; Zisoulis et al. 2010). Supporting the biological function of some of these ORF sites, bioinformatic approaches have shown that many ORF sites are preferentially conserved (Lewis et al. 2005; Stark et al. 2007; Forman et al. 2008; Schnall-Levin et al. 2010). Reporter assays have also confirmed that sites in both 5¢ UTRs and ORFs can mediate repression (Easow et al. 2007; Lytle et al. 2007; Duursma et al. 2008; Forman et al. 2008; Orom et al. 2008; Shen et al. 2008; Tay et al. 2008; Elcheva et al. 2009; Huang et al. 2010; Schnall-Levin et al. 2010).

The efficacy of miRNA-mediated repression increases with the number of sites (Doench and Sharp 2004; Grimson et al. 2007;
Bartel 2009), suggesting that targeting might be substantial if a gene contained many sites in its coding region, as could arise from coding-sequence repeats. Repeats of various types are prevalent within many human genes. Particularly widespread are simple classes of repeats, such as microsatellites, which have been shown to cause important phenotypic effects through action at both the nucleotide and protein levels (Mirkin 2007). Of interest in the present study are repeated sequences of greater complexity, a striking case of which occurs within the C2H2 class of zinc-finger genes. These genes have undergone extensive expansion over the course of vertebrate evolution and constitute the largest group of human transcription factors (Ding et al. 2009). Typically, C2H2 genes contain a significant number of tandemly repeated C2H2 amino acid domains, each of which coordinates a zinc ion and has the potential to bind DNA in a sequence-specific manner (Wolfe et al. 2000). The emergence of highly repeated amino acid domain creates instances of highly repeated short nucleotide sequences in a large fraction of such genes.

Here, we show that the ORFs of many repeat-rich genes contain strikingly large numbers of target sites of particular miRNAs. Moreover, these genes with many sites are frequently strongly repressed. For seven miRNA families, this type of targeting appears to be extensive. In the most notable cases, four miRNA families (miR-23, miR-181, miR-188, and miR-199) have seed sites that match repeated sequences within C2H2 zinc-finger genes. Three others (miR-370, miR-766, and miR-1248) have seed sites that match simpler repeats. Effective targeting of coding-region repeats is highly predictable, and, due to the large number of target sites within a single ORF, down-regulation observed in reporter assays can be stronger than that of many 3′ UTR targets. For the C2H2 class of zinc-finger genes, targeting is shared among paralogous genes, suggesting the potential for their coordinate regulation. miRNAs of both RB1 and its accessory protein, RBK, are targeted by miR-181 primarily through ORF sites, suggesting an unappreciated role in carcinogenesis for miR-181, which has previously been implicated in hematopoiesis and in tumorigenesis through its regulation of NFKB activity (Chen et al. 2004; Filipooulos et al. 2010), and underscoring the potential importance of ORF sites in understanding miRNA biology.

Results

miR-181 represses mRNAs with repeated ORF sites

While re-analyzing previously published microarray data from miR-181a transfection in HeLa cells (Baek et al. 2008), we observed that some of the most strongly down-regulated miRNAs contained miR-181 seed sites within their coding regions. We found this result intriguing because, although miRNA targeting has been observed in ORFs, it usually confers only subtle repression. Further investigation revealed that the enrichment for strong down-regulation was largely confined to a group of miRNAs containing numerous miR-181 sites (Fig. 1A,B). Genes that contained a single 8mer site in their coding region were only slightly, though statistically significantly, down-regulated (mean log2 fold-change: −0.07; p < 2 × 10−4; Mann-Whitney U-test) and significantly less so than those with a single 8mer site in their 3′ UTR (mean log2 fold-change: −0.07 vs. −0.17; p < 0.01; Mann-Whitney U-test). However, those genes with an increasing number of ORF sites showed increasingly strong down-regulation. In particular, the 52 genes with at least four 8mer ORF sites were nearly all repressed and, even when restricted to the subset of 18 of 52 genes with no 3′ UTR sites, showed significantly stronger repression than those genes with a single 8mer 3′ UTR site (mean log2 fold-change: −0.51 vs. −0.17; p < 5 × 10−3; Mann-Whitney U-test). Many of these genes had even more than four 8mers sites as well as equally large numbers of 7mer sites (Supplemental Table 2). Perhaps most surprising, considering the strikingly large number of sites, was that such genes were strongly down-regulated, but that genes should have so many sites to a single miRNA. Indeed, miR-181 was exceptional in this regard (Fig. 1C). Whereas for most miRNAs, few or no genes contained large numbers of 8mer sites, for miR-181 there were many such genes.

Having observed the impact of miR-181 on a large class of endogenous transcripts, we aimed to confirm the direct, miRNA-mediated repression of a number of specific genes. To do so, we generated reporter proteins in which firefly luciferase was fused in-frame to the C terminus of the protein product of each of five genes: ZNF573, ZFP37, ZNF20, ZNF791, and RBK (Fig. 1D). These genes contained nine, eight, six, five, and nine miR-181 8mer sites and two, seven, two, one, and 10 miR-181 7mer sites, respectively. Repression by miR-181a was evaluated by comparing normalized luciferase values from cells cotransfected with miR-181a to those with miR-23a, a noncognate miRNA. For each gene tested, we observed significant and robust repression by miR-181a (p < 10−6; Mann-Whitney U-test; Fig. 1D). To confirm that the measured signal in each case was coming from the full-length fusion proteins, we disrupted the reading frame register by inserting or deleting one nucleotide in the zinc-finger coding sequence of the miRNA (within a few hundred nucleotides of the start codon), upstream of both the miRNA sites and the firefly luciferase sequence. As expected, these frame shift mutations significantly reduced luciferase activity for each fusion construct (Supplemental Fig. 1).

To confirm that the observed repression was directly mediated by the ORF sites, we generated a ZNF20-luciferase mutant in which each of the six miR-181 8mer and two miR-181 7mer seed sites within the ORF were mutated with two synonymous point substitutions. Compared with this mutated construct, the wild-type ZNF20 reporter was significantly and specifically repressed by miR-181a (p < 10−4; Fig. 1E). The repression attributed to these sites increased from 2.5-fold to 6.3-fold when the fragment containing the sites was incorporated as part of the reporter 3′ UTR (Supplemental Fig. 2A). This difference between the efficacy of ORF sites and 3′ UTR sites was significant (p < 10−4) and consistent with the general observation of ORFs being more refractory to miRNA targeting than are 3′ UTRs (Grimson et al. 2007).

Similarly, we tested whether the repression of RBK, observed in our luciferase assays (Fig. 1D) and in the previous microarray study (Baek et al. 2008), was directly mediated by its miR-181 sites. In addition to its 19 ORF sites (nine 8mers and 10 7mers), the 3′ UTR of RBK contains a conserved 7mer-m8 site and a poorly conserved 7mer-A1 site (Friedman et al. 2009). A panel of constructs was generated, expressing either the RBK ORF with a C-terminal luciferase tag or the RBK 3′ UTR following the luciferase reporter (Fig. 1F). Although the 3′ UTR sites mediated statistically significant repression (1.3-fold; p = 0.0009), the ORF sites gave far stronger repression (3.2-fold repression; p < 10−4). As with ZNF20, the repression mediated by these ORF sites increased significantly when they were incorporated as part of the reporter 3′ UTR (p < 10−4; Supplemental Fig. 2B). When both the ORF and 3′ UTR sites were monitored in combination, 3.3-fold repression was observed (p < 10−6; Supplemental Fig. 3), which was not significantly different from that observed with the ORF sites alone (p = 0.59). Although further experiments will be required to understand why additional repression was not observed in the combined reporter, these results indicate that the repression of RBK
by miR-181a was direct and that the majority of the miR-181a-mediated repression of RBAK was due to targeting through its ORF sites.

Additional miRNA seeds match repeated coding-region motifs
To find other miRNAs that might affect ORF targets similarly, we searched for all 8mers highly repeated within human ORFs, and for each 8mer we counted the number of nonoverlapping occurrences within each ORF. Because four ORF sites mediated robust repression by miR-181, we chose this as a threshold, and for each 8mer recorded the number of genes with four or more instances of that 8mer (Fig. 2A). The majority (77%) of 8mers did not occur four or more times in any coding region, and the vast majority (97%) occurred four or more times in only five or fewer coding regions. At the tail of the distribution, 334 8mers appeared at least four times in at least 25 genes, among which were seven miRNA 8mer seed matches (Table 1). For each of the seven corresponding miRNA seed families, we compiled potential target sets containing genes with at least four 8mer sites (Supplemental Tables 1–7).

Figure 1. miR-181 targets genes with multiple coding-region sites. (A) Response of mRNAs to the introduction of miR-181a into HeLa cells. Plotted are cumulative distributions of fold-changes for mRNAs with the indicated numbers and types of sites. Except for the two categories indicated, categories with ORF sites excluded mRNAs with 3' UTR sites, and those with 3' UTR sites excluded mRNAs with ORF sites. (B) Mean fold-changes for the categories of A. Error bars show standard deviation from bootstrapping. (C) The propensity of miR-181 to have many ORF sites. Plotted are numbers of genes containing at least the indicated number of sites for either miR-181 (green) or the median across all miRNAs (blue). Error bars show the interquartile range. (D) miR-181a-mediated repression of reporters with miR-181 ORF sites. Reporters included the luciferase ORF following the ORF of the indicated mRNA. Fold repression was calculated relative to that of the noncognate miRNA, miR-23a (see Methods). Plotted are the normalized values, with error bars representing the third largest and third smallest values (n = 12; p < 10⁻⁶, except for the control, luciferase-only reporter, for which p = 0.32). (E) Dependence of ZNF20 repression on miR-181 ORF sites. Repression was calculated and depicted as in D, additionally normalizing repression of the reporter with wild-type sites (WT) to that of a reporter in which the eight ORF sites were mutated (n = 12; p < 10⁻⁶). (F) Direct repression of RBAK ORF and 3' UTR mediated by miR-181a, as assayed by luciferase assays. Repression was calculated and depicted as in D, additionally normalizing repression of the reporter with wild-type sites (WT) to that of the mutant reporter in which the 19 ORF or two 3' UTR sites were mutated (n = 12; p = 7 × 10⁻⁷ and p = 0.0009, respectively).
Because of their limited conservation, low expression levels and questionable status as authentic miRNAs (Chiang et al. 2010; Kozomara and Griffiths-Jones 2011), two of these three miRNAs (miR-766 and miR-1248) were not considered further.

The most common form of the repeated C2H2 amino acid domain is XCX[2]CX[12]HX[3]H (X represents any amino acid) (Emerson and Thomas 2009). Typically, C2H2 zinc-finger genes contain many tandem repeats of the finger motif (8.5 on average in humans), which are connected by a specific linker sequence most commonly of the form TGEKPY (Emerson and Thomas 2009). The 8mer sites for the four miRNA families each occurred within specific amino acid realizations at specific locations in this motif (Fig. 2C). Because of the large numbers of paralogous motifs within a single gene, the number of seed sites can be large. A typical example is the gene RBAK, which contains 14 C2H2 domains, eight miR-181 8mers and 10 miR-181 7mers (Fig. 2D).

Predicted ORF targets are subject to miRNA-mediated repression

From the list of predicted miR-23 targets (Supplemental Table 1), we chose three genes (ZNF225, ZNF486, and ZNF85) for experimental follow-up. Fusion constructs with a C-terminal luciferase tag were made as before, and disruption of the reading frame by a nucleotide insertion substantially decreased luciferase expression, confirming that the majority of signal came from the full-length protein (Supplemental Fig. 1). The constructs were transfected in the presence of miR-23a or a non-cognate miRNA, miR-181a (Fig. 3A). For all three targets, normalized luciferase values were significantly reduced in the presence of miR-23a when compared with those with the noncognate miRNA ($p < 10^{-6}$). The magnitudes of repression were particularly notable because miR-23 has high target abundance and thus generally weaker targeting efficacy (Arvey et al. 2010).

We extended this analysis to predicted targets of either miR-199 or miR-370. ZNF20 and ZNF791, experimentally supported targets of miR-181 (Fig. 1), also contain multiple seed matches to miR-199 and were examined for their response to miR-199a (Fig. 3B). ZNF791, which contains six miR-199 8mer sites and one 7mer site, was significantly repressed ($p < 10^{-3}$). Similarly, in the case of miR-370, we probed the response of two targets, IVL and HDAC5, both of which were significantly repressed ($p < 10^{-2}$ and $p = 0.009$, respectively; Fig. 3C). When we generated a mutant IVL construct, wherein all miR-370 sites were disrupted, we observed significant, direct repression by miR-370 ($p < 10^{-4}$; Fig. 3D). In contrast, the luciferase control was either unaffected by these miRNAs or, in the case of miR-199a, significantly less repressed than the ZNF791-luciferase fusion ($p = 0.0003$; Supplemental Fig. 4). Taken together, although individual ORF sites are less effective than 3’ UTR sites (Grimson et al. 2007), these results indicate that highly repetitive

### Table 1. Annotated miRNAs with repeated coding-region sites in many genes

<table>
<thead>
<tr>
<th>Family</th>
<th>miRNAs</th>
<th>8mer site</th>
<th>miRNA conservation</th>
<th>Genes with ≥4 sites</th>
<th>Target family</th>
</tr>
</thead>
<tbody>
<tr>
<td>miR-23</td>
<td>miR-23a/b</td>
<td>AATGTGAA</td>
<td>Vertebrates</td>
<td>43</td>
<td>C2H2 zinc fingers</td>
</tr>
<tr>
<td>miR-181</td>
<td>miR-181a/b/c/d</td>
<td>TGAATGTA</td>
<td>Vertebrates</td>
<td>75</td>
<td>C2H2 zinc fingers</td>
</tr>
<tr>
<td>miR-188</td>
<td>miR-188-3p</td>
<td>TGCGGAAA</td>
<td>Mammals</td>
<td>210</td>
<td>C2H2 zinc fingers</td>
</tr>
<tr>
<td>miR-199</td>
<td>miR-199a/b-5p</td>
<td>ACCTGGA</td>
<td>Vertebrates</td>
<td>82</td>
<td>C2H2 zinc fingers</td>
</tr>
<tr>
<td>miR-370</td>
<td>miR-370</td>
<td>CCGCAGA</td>
<td>Mammals</td>
<td>25</td>
<td>Varied</td>
</tr>
<tr>
<td>miR-766</td>
<td>miR-766</td>
<td>GCCTGAGA</td>
<td>Human</td>
<td>27</td>
<td>Varied</td>
</tr>
<tr>
<td>miR-1248</td>
<td>miR-1248</td>
<td>AAGACAGA</td>
<td>Human</td>
<td>34</td>
<td>Varied</td>
</tr>
</tbody>
</table>
ORFs containing many miRNA sites can generally be subject to significant and, in some cases, substantial repression by the cognate miRNA.

miRNAs target multiple paralogous families of C2H2 zinc-finger genes

Having observed the extent and, in some cases, surprising magnitude of targeting arising from ORF repeats, we considered the evolutionary processes giving rise to this phenomenon. Our focus was on C2H2 zinc-finger genes because these formed the most dramatic and frequent instances of this phenomenon. Among the predicted targets, most C2H2 genes (>80%) contained the general transcriptional repressor KRAB domain (Fig. 4A). KRAB-containing C2H2 genes display particularly interesting patterns of evolution, having high rates of gene duplication and loss as well as a dramatic expansion over the course of vertebrate and mammalian evolution (Huntley et al. 2006). To understand the role that these duplications played in forming miRNA target sets, we collected sequences of all KRAB domains annotated in human and used these to create a multiple alignment of KRAB domains. From this alignment, the inferred phylogeny of KRAB-containing genes provided a context for considering the four miRNA target sets (Fig. 4B; Supplemental Fig. 5). For this analysis, we again defined each target set as those genes containing at least four 8mer sites for the given miRNA.

The analysis revealed that the four miRNAs target multiple clades, each of which has emerged from significant expansions through gene duplication. Combined over the four miRNA families, the target sets covered the majority of KRAB-containing C2H2 genes with varying amounts of overlap between each set (Fig. 4C). Whereas the miR-188 target set spanned nearly the entire phylogeny, the other three miRNAs each targeted more specific subfamilies of genes. For instance, the miR-23 targets were nearly all members of one clade: the recently duplicated ZNF91 subfamily, which has undergone a significant expansion in the primate lineage (Hamilton et al. 2006).

Sequence analysis suggested that two duplication processes contributed to the creation of such extensive repeat-containing subfamilies. In one process, individual C2H2 domains were duplicated multiple times within a single zinc-finger gene. In a second process, genes were duplicated to form an extensive gene subfamily with high nucleotide similarity among many members. Due to the intragenic duplication process, the sequences of individual zinc fingers within a gene were far more similar to each other than expected by chance. To verify the importance of this effect, we implemented a randomization procedure for nucleotide sequences within C2H2 domains that preserved amino acid sequences and average codon usage over all domain instances. Even when fixing the observed amino acid sequences, real instances of C2H2 domains from within the same gene showed significantly higher nucleotide similarity than expected by chance (Supplemental Fig. 6). When C2H2 domains with randomized sequences were mapped back to genes, far fewer of these genes contained large numbers of miRNA sites than did the real genes (Fig. 5A). It was tempting to speculate that the similarity in nucleotide sequences across C2H2 domains within a gene represented an additional selective pressure to maintain miRNA seed sites in these genes. We observed marginal evidence for this model but could not detect such an effect at high confidence (data not shown).

The initial intragenic duplication of individual C2H2 domains allowed for the founding nucleotide-sequence choice to be amplified. After repeated duplication, a gene would ultimately include either many sites or no sites at all, depending on the presence or absence of a target site within a founding C2H2 domain. Such duplication contributed to the modularity of miRNA target sets; even for subfamilies of genes with C2H2 domains containing similar amino acid sequences, genes from one subfamily often contain many target sites, whereas those from another contain almost none. For example, although miR-188 sites are prevalent throughout the set of KRAB genes, they are almost entirely absent from the ZNF91 subfamily of genes, despite these genes encoding comparable numbers of instances of the amino acid triplet (CGK) within which the miR-188 seed site appears.

The presence of large numbers of miRNA sites within coding regions also provided an opportunity to gain miRNA sites in the 3’ UTR through the acquisition of nonsense mutations. For all four miRNAs, we observed clear evidence of this process. For each miRNA, the set of predicted target genes was far more likely to contain 3’ UTR target sites than was the overall set of genes (p < 10^{-7} for all comparisons except miR-199 7mers; binomial test; Fig. 5B). Moreover, regions flanking 3’ UTR sites of predicted ORF targets had high similarity to the regions flanking the corresponding ORF sites, which suggested that many of these 3’ UTR sites resided in the remnants of zinc-finger domains that had been lost to the 3’ UTR (Fig. 5C). Although the average number of 3’ UTR sites in these genes was modest compared with the number of ORF sites (mean total number of 7mer and 8mer sites: miR-23, 2.7; miR-181,
expression for these genes is identical to that for most 3' mRNA. It therefore appears likely that the mechanism of repression of genes. While some of the target sites we have identified show the potential for supplemental 3' repressed ORF targets exist and that such targets can be easily identified by the presence of large numbers of sites to the same miRNA. Recently, a similar observation of ORF targeting was identified by the presence of large numbers of sites to the indicated miRNA. (Huang et al. 2010). Our work, going significantly further, provides a systematic examination of this phenomenon and gives the first de novo observed to occur in the case of a single miRNA (Huang et al. 2010). Our work, going significantly further, provides a systematic examination of this phenomenon and gives the first bioinformatic evidence and experimental verification of the widespread nature of this type of targeting. Indeed, we have found that this targeting involves multiple miRNAs and, likely, hundreds of genes. While some of the target sites we have identified show the potential for supplemental 3' base-pairing to the miRNA, none exhibited the extensive complementarity required for cleavage of the miRNA. It therefore appears likely that the mechanism of repression for these genes is identical to that for most 3' UTR targets.

Interestingly, although there are examples of both 5' and 3' UTRs containing repeats, none of the highly repeated motifs matched known miRNA seeds, which indicates that repeat-mediated targeting is largely ORF specific. Moreover, although repeats in coding regions are prevalent in numerous animal clades, the presence of miRNA sites within these repeats appeared to be vertebrate-specific. Here, we focused on those miRNAs with at least four sites in large sets of genes, but there were many other miRNAs with at least this many sites in a handful of genes (Fig. 1). In addition, many genes contained multiple ORF sites to multiple miRNAs. The eventual determination of expression patterns for these miRNAs at cellular resolution should enable prediction of additional targets in which ORF sites combine to achieve substantial repression.

The most striking cases of repeat-rich targeting occurred within the KRAB-domain C2H2 zinc-finger genes, which constitute the largest collection of human transcription factors. Our results indicate that a single miRNA can target entire families of these genes, thereby simultaneously regulating large numbers of evolutionarily related transcription factors. The targeting of such a large number of transcription factors by a miRNA has the potential to cause significant and widespread downstream effects. Through phylogenetic analysis we have shown how repeat-mediated targeting arises in these genes under an extensive evolutionary duplication process. Analysis of KRAB-domain-containing families indicates positive selection toward diversification of DNA-binding residues (Emerson and Thomas 2009). Hence, following duplication, individual genes frequently gain new downstream regulatory roles, while most members of the family retain the potential for upstream miRNA-mediated regulation. With a few exceptions, the functions of KRAB-domain genes remain unknown (Huntley et al. 2006). While a few show tissue-specific expression, most are widely expressed (Vaquerizas et al. 2009), and most predicted target genes appear to have expression patterns overlapping that of the corresponding miRNA (data not shown), suggesting that the targeting described here is likely to affect their in vivo expression.

One ORF target of miR-181, RBAK, is reported to function as an RB1-associated transcriptional repressor (Skapek et al. 2000). Suggestive of coordinate regulation, RB1, which encodes the well-characterized tumor-suppressor protein (Hanhana and Weinberg 2000; Classon and Harlow 2002), also contains sites to miR-181 in both its ORF (two 8mers and one 7mer) and 3' UTR (one non-conserved 7mer). Although no repression was observed for the 3' UTR site, significant repression by miR-181a was observed for the RB1 ORF-luciferase fusion (1.4-fold; $p < 10^{-6}$; Supplemental Fig. 7). This newly recognized miR-181 targeting of RB1 and RBAK had not been appreciated from previous analyses focusing only on 3' UTR

![Figure 4](https://example.com/figure4.png)

**Figure 4.** MicroRNA targeting of paralogous C2H2 genes. (A) Diagram of domain structure of C2H2 zinc-finger genes. (B) The relationship between shared ancestry of KRAB-containing C2H2 genes and shared miRNA sites. The phylogeny inferred from alignment of the KRAB domains is shown, marking at the perimeter those with at least four 8mer sites to the indicated miRNA. (C) Overlap between the predicted targets of the indicated miRNAs.

Discussion

In animals, miRNAs target many genes through sites in their 3' UTRs but cause only modest repression of most of these targets. Compared with this generally modest targeting within 3' UTRs, most targeting within ORFs is substantially weaker still. Nonetheless, we have shown that significant numbers of substantially repressed ORF targets exist and that such targets can be easily identified by the presence of large numbers of sites to the same miRNA. Recently, a similar observation of ORF targeting was independently observed to occur in the case of a single miRNA (Huang et al. 2010). Our work, going significantly further, provides a systematic examination of this phenomenon and gives the first bioinformatic evidence and experimental verification of the widespread nature of this type of targeting. Indeed, we have found that this targeting involves multiple miRNAs and, likely, hundreds of genes. While some of the target sites we have identified show the potential for supplemental 3' base-pairing to the miRNA, none exhibit the extensive complementarity required for cleavage of the miRNA. It therefore appears likely that the mechanism of repression for these genes is identical to that for most 3' UTR targets.

Interestingly, although there are examples of both 5' and 3' UTRs containing repeats, none of the highly repeated motifs matched known miRNA seeds, which indicates that repeat-mediated targeting is largely ORF specific. Moreover, although repeats in coding regions are prevalent in numerous animal clades, the presence of miRNA sites within these repeats appeared to be vertebrate-specific. Here, we focused on those miRNAs with at least four sites in large sets of genes, but there were many other miRNAs with at least this many sites in a handful of genes (Fig. 1). In addition, many genes contained multiple ORF sites to multiple miRNAs. The eventual determination of expression patterns for these miRNAs at cellular resolution should enable prediction of additional targets in which ORF sites combine to achieve substantial repression.

The most striking cases of repeat-rich targeting occurred within the KRAB-domain C2H2 zinc-finger genes, which constitute the largest collection of human transcription factors. Our results indicate that a single miRNA can target entire families of these genes, thereby simultaneously regulating large numbers of evolutionarily related transcription factors. The targeting of such a large number of transcription factors by a miRNA has the potential to cause significant and widespread downstream effects. Through phylogenetic analysis we have shown how repeat-mediated targeting arises in these genes under an extensive evolutionary duplication process. Analysis of KRAB-domain-containing families indicates positive selection toward diversification of DNA-binding residues (Emerson and Thomas 2009). Hence, following duplication, individual genes frequently gain new downstream regulatory roles, while most members of the family retain the potential for upstream miRNA-mediated regulation. With a few exceptions, the functions of KRAB-domain genes remain unknown (Huntley et al. 2006). While a few show tissue-specific expression, most are widely expressed (Vaquerizas et al. 2009), and most predicted target genes appear to have expression patterns overlapping that of the corresponding miRNA (data not shown), suggesting that the targeting described here is likely to affect their in vivo expression.

One ORF target of miR-181, RBAK, is reported to function as an RB1-associated transcriptional repressor (Skapek et al. 2000). Suggestive of coordinate regulation, RB1, which encodes the well-characterized tumor-suppressor protein (Hanhana and Weinberg 2000; Classon and Harlow 2002), also contains sites to miR-181 in both its ORF (two 8mers and one 7mer) and 3' UTR (one non-conserved 7mer). Although no repression was observed for the 3' UTR site, significant repression by miR-181a was observed for the RB1 ORF-luciferase fusion (1.4-fold; $p < 10^{-6}$; Supplemental Fig. 7). This newly recognized miR-181 targeting of RB1 and RBAK had not been appreciated from previous analyses focusing only on 3' UTR
sites and is intriguing when considering that miR-181 is up-regulated in some cancers and is important for maintaining cancer stem cells in hepatocellular carcinoma (Schetter et al. 2008; Cervigne et al. 2009; Ji et al. 2009). Even transient induction of miR-181b is sufficient to mediate an epigenetic switch to cancer, and inhibition of miR-181b reduces colony formation in several cancer cell lines, observations proposed to result from direct targeting of the tumor-suppressor CYLD (Iliopoulos et al. 2010). Because both RB1 and RBAK repress activation of E2F-dependent promoters and decrease DNA synthesis (Skapek et al. 2000), the ability of miR-181 to repress RB1 and RBAK might provide an additional mechanism by which this miRNA mediates transformation.

Our work also suggests a general role of coding-sequence repeats in post-transcriptional regulation. For transcription regulation, the accumulation and clustering of multiple transcription factor binding motifs has been used for some time to predict functional regulatory relationships (Hannenhalli 2008). Given the large number of post-transcriptional regulatory processes that exist beyond miRNAs (Garneau et al. 2007; Parker and Sheth 2007) and the vast extent of sequence repeats within many protein-coding genes, miRNAs might not be the only regulatory process utilizing this phenomenon.

Methods
Luciferase assays
HEK293 cells (ATCC) were plated in 24-well plates and transfected 24 h later using Lipofectamine 2000 (Invitrogen) and Opti-MEM (Sigma) with 50 ng of Renilla luciferase control reporter plasmid pS1 (Grimson et al. 2007), 400 ng of firefly luciferase reporter plasmid, and 25 nM miRNA duplex (Supplemental Table 9) per well. After 12 h, transfection media was replaced with DMEM containing 10% fetal bovine serum and penicillin-streptomycin. Cells were harvested 48 h post-transfection. Luciferase activities were measured using dual-luciferase assays (Promega), as described by the manufacturer. Four to five biological replicates, each with three technical replicates, were performed. Firefly activity was first normalized to Renilla activity to control for transfection efficiency, and then normalized values were analyzed as described previously (Grimson et al. 2007). If a mutant construct was generated, Renilla-normalized firefly values were normalized to the geometric mean of values from the reporter in which the miRNA sites were mutated. To control for variability in plasmid preparation, the value plotted for each construct was the geometric mean of normalized firefly values from transfections with the cognate miRNA divided

Figure 5. Impact of evolutionary processes on microRNA targeting of C2H2 genes. (A) Increased targeting due to intragenic C2H2 domain similarity. Shown are ratios giving the numbers of genes containing the indicated minimum number of 8mers for codon-randomized genes divided by the number of genes containing that many 8mers for real gene sequences. Values shown are the mean ratios across 50 codon-randomization trials. Error bars show standard deviation across the trials. All but three values were statistically significant (genes with four sites for miR-23, p < 0.01; genes with four or five sites for miR-188, p > 0.05; genes with four sites for miR-199, p < 0.05; all other sets, p < 10^-7; paired Student’s t-test with Bonferroni correction). (B) Propensity of ORF target genes to also contain 3’ UTR target sites. Shown are fractions of genes containing either 8mer or 7mer sites for the indicated miRNA within their 3’ UTRs, comparing the ORF target sets with the background of all genes. Error bars show standard deviations from 100 bootstrapping trials (p < 10^-7 for all comparisons except miR-199 7mers; binomial test). (C) Evidence that 3’ UTR sites derived from ORF sites. Shown are nucleotide compositions flanking miRNA sites in both ORFs and 3’ UTRs of miRNAs with ORF sites. Letter size indicates nucleotide enrichment, visualized using WebLogo.
by that from transfections with the noncognate miRNA. To combine
replicate values from independent experiments for which no
mutant construct was generated, each Renilla-normalized firefly
value was normalized to the geometric mean of values from trans-
fections with the noncognate miRNA. Statistical significance was
determined using the Mann-Whitney U-test. Plasmids (deposited at
Addgene) were constructed as described (see Supplemental
Methods).

Gene sequences
ReSeq sequences for human ORFs were downloaded from the
UCSC Genome Browser (www.genome.ucsc.edu), version hg19,
Feb. 2009. In cases of multiple transcript variants for a single gene,
one variant was chosen at random as a representative. 3' UTR se-
quences were downloaded from the TargetScan website (www.
targetscan.org), version 5.1. For analysis of k-mers and generation
of target sets, only nonoverlapping k-mer instances were consid-
ered, and all transcripts with coding region length >10 kilobases
were excluded. Note that there was a single nucleotide difference
between the RBAK clone used in luciferase assays and the ReSeq
resulting in an additional 8mer in the clone sequence. In
Figure 2 and Supplemental Table 2, we referred to the counts from
the ReSeq sequence.

Microarray data
The microarray data examining the response of introducing miR-
181a into HeLa cells (GSM302995) (Baek et al. 2008) were analyzed
using Agilent Feature Extraction software. Log2 fold-change values
for genes were obtained by taking the median value of log2 fold-
change for all probes against that gene (excluding any probes not
possible among amino acids at each position). Empirical codon usage
frequencies combined across all C2H2 domains were calculated. For
each possible amino acid within this position, empirical codon usage
frequencies were obtained by taking the median value of log2 fold-
change values estimated using 100 bootstrap trials.

Phylogenetic reconstruction
The sequences of all KRAB-A domain instances, as well as the Uni-
Prot identifier for the protein in which each domain occurred, were
downloaded from the Pfam website (pfam.sanger.ac.uk), version 24,
Oct. 2009. UniProt identifiers were mapped to the correspond-
ing genes using conversion files from the HUGO Gene Nomencla-
ture Committee (www.genenames.org). Of the resulting 311 genes
containing KRAB-A domains, all but four (ZNF862, ZNFS60, ZNF333,
and ZFP28) contained a single copy of the domain. For those four
genes, a single instance of the domain was chosen at random to use
in the phylogenetic analysis. A multiple alignment was inferred on the amino acid sequences of the 311 KRAB-A do-
 mains, based on which a phylogenetic tree of the corresponding
genes was reconstructed, using ClustalX software (version 2.0.12)
under default parameter settings. Visualization of the tree and overlap with miRNA target sets was done using the Interactive
Tree of Life software tool (itol.embl.de).

Randomization of C2H2-domain sequences
Instances of a C2H2 domain of the form XCX26CX12HXX6H, as
well as the six residues N-terminal of this motif (to capture the
linker sequence) were recorded from within all KRAB-containing
genes. For each position in this 28-amino-acid motif, and each
possible amino acid within this position, empirical codon usage
frequencies combined across all C2H2 domains were calculated.
For each instance of a C2H2 domain, randomized nucleotide versions of
this C2H2 domain were generated by maintaining the amino acid
sequence of each motif instance but randomly sampling with re-
placement from the empirical codon frequencies for each amino
acid at each position in the domain. To generate C2H2 genes with
randomized nucleotide sequences, the above randomization pro-
cEDURE was repeated for all of the C2H2 domains in the gene, and the
randomized nucleotide sequences for each C2H2 domain were used
to replace the true nucleotide sequences.

Acknowledgments
We thank L. Mirny, C. Jan, and P. Schmid for helpful discussions.
This work was supported by grants from the NIH (GM067031 to
D.P.B. and 1R01GM081871 to B.B.). M.S.I. was supported by the
Hertz Foundation and NDSEG Fellowships. O.S.R. was supported
by a Ruth L. Kirschstein National Research Service Award (GM088872).
D.P.B. and N.P. are investigators of the Howard Hughes Medical
Institute.

References
Arvey A, Larsson E, Sander C, Leslie CS, Marks DS. 2010. Target miRNA
Cervigne NK, Reis PP, Machado J, Sadikovic B, Bradley G, Galloni NN,
Identification of a microRNA signature associated with progression of
Chen CZ, Li L, Lodish HF, Bartel DP. 2004. MicroRNAs modulate
Chiang HR, Schoenfeld JW, Ruby JG, Auyeung VC, Spies N, Baek D,
microRNAs: Experimental evaluation of novel and previously annotated
Classon M, Harlow E. 2002. The retinoblastoma tumor suppressor in
nar/gkn782.
Duursma AM, Kedde M, Schrier M, le Sage C, Agami R. 2008. miR-148 targets
by miRNP immunopurification. RNA 13: 1198–1204.
Elcheva I, Goswami S, Noubissi FK, Spiegelman VS. 2009. CRD-BP protects
the coding region of betaTcP1 mRNA from miR-183-mediated
Emerson RO, Thomas JH. 2009. Adaptive evolution in zinc finger
sequence transcription factors. PLoS Genet 5: e1000325. doi: 10.1371/
journal.pgen.1000325.
sequences in coding regions reveals that the let-7 microRNA targets
Dicer within its coding sequence. Proc Natl Acad Sci USA 105: 14879–14884.
Forman BC, Farh KK, Burge CB, Bartel DP. 2009. Most mammalian miRNAs
Garneau NL, Wilusz J, Wilusz CJ. 2007. The highways and byways of miRNA
Grimson A, Farh KK, Johnston WK, Garrett-Engele P, Lin LP, Bartel DP.
2007. MicroRNA targeting specificity in mammals: Determinants
of microRNA targets to the 3' untranslated region in mammalian
Guo H, Ingolia NT, Weissman JS, Bartel DP. 2010. Mammalian microRNAs
MicroRNA targeting of coding repeats


Received January 21, 2011; accepted in revised form June 8, 2011.