

**An evolutionary perspective on the sequence, mechanism, and regulatory function of
animal microRNAs**

Soraya Yekta

*H.B.Sc. Chemistry
University of Toronto, 2001*

SUBMITTED TO THE DEPARTMENT OF BIOLOGY
IN PARTIAL FULFILMENT OF THE REQUIREMENTS
FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

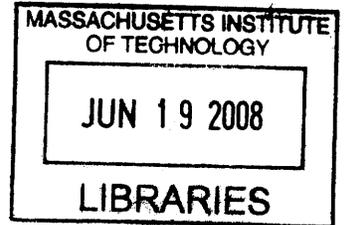
AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

DECEMBER 2007
(June 2008)

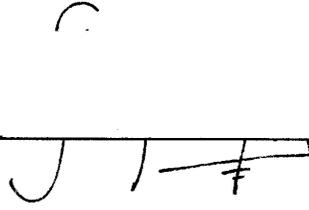
©2007 Soraya Yekta. All Rights Reserved.

The author hereby grants to MIT permission to reproduce
and to distribute publicly paper and electronic copies
of this thesis document in whole or in part.

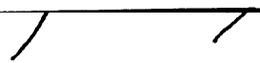


ARCHIVES

Signature of Author:


Soraya Yekta
Department of Biology
(7 December 2007)

Certified by:


David P. Bartel
Professor of Biology
Thesis Advisor

Accepted by:


Stephen P. Bell
Professor of Biology

An evolutionary perspective on the sequence, mechanism, and regulatory function of animal
microRNAs

Soraya Yekta

Submitted to the Department of Biology on date TBD, 2007

In Partial Fulfilment of the Requirements for the Degree of Doctor of Philosophy

ABSTRACT

Noncoding RNAs are encoded by diverse genomes and play many functional roles. MicroRNAs (miRNAs) are endogenous ~22-nucleotide noncoding RNAs, derived from larger hairpin precursors, which act by base-pairing to mRNAs to target these transcripts for destruction or translational repression. miRNA genes have been discovered in viral and multicellular genomes.

The computational procedure MiRscan was developed to identify miRNA genes conserved in more than one genome and applied to the identification of vertebrate miRNAs. Starting with conserved mouse and human sequences with potential for hairpin formation, and subsequent comparison to fish, 15,000 human genomic loci were identified within aligned regions outside protein coding genes, and ranked according to criteria based on shared features of a training set of the first 50 experimentally verified *C. elegans* miRNAs. 188 high-scoring candidates, including 74% of human miRNAs known in 2002, were further examined. Fourteen candidate miRNAs were close paralogues of known miRNAs, and 38 candidates were experimentally validated using cDNA libraries of small RNAs made from zebrafish. Of the 38 verified miRNAs, 21 were sequences identified by random cloning and sequencing of cDNA libraries, while 17 were found by applying a directed PCR approach to the same libraries.

The miR-196 and miR-10 families are transcribed from genomic loci within clusters of Hox transcription factor genes, and in turn mediate the posttranscriptional repression of neighbouring Hox transcripts, with conserved and extensive targeting of Hox genes located in paralogous groups that are 3' but not 5' of each miRNA locus, relative to the direction of transcription within a given cluster. The vertebrate-specific miR-196 family is encoded at three paralogous loci in the mammalian Hox clusters, and has complementarity to messages of several Hox genes, including Hox8 paralogues. RNA fragments diagnostic of miR-196-directed cleavage of *Hoxb8* were detected in mouse embryos. Cell culture experiments demonstrated down-regulation of *Hoxb8*, *Hoxc8*, *Hoxd8*, and *Hoxa7* and supported a cleavage mechanism for miR-196-directed repression of *Hoxb8*. These results point to a miRNA-mediated mechanism for the posttranscriptional restriction of Hox gene expression during vertebrate development and demonstrate that metazoan miRNAs can repress expression of their natural targets through mRNA cleavage in addition to inhibiting a translational step.

Inhibition of the two Hox miRNAs in chick embryos, resulted in axial skeletal patterning defects in domains that overlap considerably with Hox target and miRNA expression, supporting a specialization of miRNAs in Hox gene regulation, and consistent with action by miR-196 and miR-10 to refine posterior boundaries at relative levels of expression for multiple Hox genes. The genomic distribution of target sites and Hox patterns of expression suggest that the miRNAs further act in concert with more posteriorly expressed Hox genes to impose a functional hierarchy over more anterior ones, a molecular mechanism consistent with 'posterior prevalence'. The posttranscriptional downregulation of more 3' and anteriorly expressed Hox genes by miR-196 constitutes an evolutionarily recent regulatory layer of the highly constrained Hox network, one which recapitulates modes of interactions existing at multiple levels of gene expression.

Thesis Advisor: David P. Bartel

Title: Professor

To Banou va Ahmad

ACKNOWLEDGEMENTS

I am grateful to the members of my doctoral defense committee, Phil Sharp, Chris Burge, Hazel Sive, and Cliff Tabin. I thank Cliff Tabin and members of the Cepko and Tabin laboratories at Harvard, who graciously shared their expertise and facilities.

I thank David Bartel, my graduate advisor, for his intellectual and environmental guidance.

I extend my gratitude to former and current members of the Bartel lab, and friends at MIT, who have given me invaluable insight and inspired me in my graduate experience.

Lastly I thank my family and tribe, for their special ways that have shaped me as a scientist.

Table of Contents

Abstract	2
Introduction	6
Summary of thesis	22
Chapter One	33
Published as: Lim L.P., Glasner M.E., Yekta S., Burge C.B., and Bartel D.P. Vertebrate microRNA genes. <i>Science</i> . 299:1540. (2003).	
Chapter Two	55
Published as: Yekta, S., Shih I-H, and Bartel, D.P. MicroRNA-directed cleavage of <i>HOXB8</i> mRNA. <i>Science</i> . 304:594. (2004).	
Chapter Three	73
Regulation by microRNAs contributes to the functional hierarchy among vertebrate Hox genes.	
Future Directions	106
Appendix I	110
Lim, L.P., Lau N.C., Weinstein E.G., Abdelhakim A., Yekta, S., Rhoades M.W., Burge, C. B., and Bartel, D.P. The microRNAs of <i>Caenorhabditis elegans</i> . <i>Genes and Development</i> . 17:991-1008. (2003).	
Appendix II	120
Ohler, U., Yekta, S., Lim, L.P., Bartel, D. P., and Burge, C.B. Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification. <i>RNA</i> . 10:1309-1322. (2004).	
Appendix III	128
Hornstein, E., Mansfield J.H., Yekta, S., Hu, J.K., Harfe, B. , McManus, M., Baskerville, S., Bartel, D.P. and Tabin, C.J. The microRNA miR-196 acts upstream of Hoxb8 and Shh in limb development. <i>Nature</i> . 438:671-674. (2005).	

Introduction

Noncoding RNA genes are ubiquitous elements in genomes of all organisms, implicated in vast aspects of life processes, including the posttranscriptional regulation of protein-coding genes. Many multicellular organisms and several large DNA viruses encode a class of ~22-nucleotide (nt) noncoding RNAs called microRNAs (miRNAs). The first members of this class, *lin-4* and *let-7*, were described in the nematode *C. elegans* and characterised years before microRNAs were termed as such, or found to be as prevalent. *lin-4* and *let-7* were mutants derived from genetic screens for defects of the heterochronic pathway, and control the timing of developmental events. Identified by Ambros, *lin* (cell lineage abnormal) genes act as binary switches that regulate the number of stem-cell divisions, and the timing of cell fate specification (Ambros, 1989; Ambros and Horvitz, 1984; Chalfie et al., 1981). At late larval stages, the corresponding mutant loci cause certain somatic cell lineages, to repeat divisions specific to earlier fates. The *lin-4* loss-of-function mutant, or the *lin-14* gain-of-function mutant reiterate divisions specific to the first larval stage (L1) throughout later stages, and never acquire adult forms. The *lin-4* gene encodes a 22-nt noncoding RNA that represses the translation of *lin-14* by direct and imperfect base-pairing to multiple complementary sequences in the 3' untranslated region (UTR) of *lin-14* mRNA (Wightman et al., 1993). Induction of *lin-4* expression midway through L1 precedes a reduction in the levels of *lin-14* protein, and is required for the succession of certain cellular fates from first to second larval stages. *lin-4* also interacts with *lin-28*, another heterochronic gene required to specify L2-specific fates. In response to the stage dependent rise of *lin-4* RNA, *lin-28* protein output is lowered. This negative regulation of *lin-28* by *lin-4*, less abrupt than that of *lin-14*, is mediated through a single partially complementary 3' UTR site (Moss et al., 1997).

The second small noncoding RNA *let-7* (*lethal-7*) to be characterised is another player in the heterochronic pathway. Required for transition of larvae from L4 to adulthood, *let-7* RNA appears in late L3, and remains expressed throughout later stages (Reinhart et al., 2000; Slack et al., 2000). *let-7* negatively regulates *lin-41* through pairing to two partially complementary 3' UTR sites. Mutants of *let-7* fail to downregulate *lin-41* and reiterate L4 cell fates in the adult (Reinhart et al., 2000; Slack et al., 2000).

lin-4 and *let-7* inhibit gene expression in *C. elegans* by translational control and through direct pairing to partially complementary sequences in several 3' UTRs. Both small RNAs are 21-22 nt

in length and derived from larger hairpin precursors (Lee et al., 1993; Reinhart et al., 2000). Before the availability of a large number of sequenced genomes, *let-7* RNA and the temporal nature of its expression, were shown to be conserved from nematodes to vertebrates (Pasquinelli et al., 2000). The finding that the small RNA was not exclusive to *C. elegans* and existed in many metazoans, suggested conserved function. Since then, *let-7* has been implicated in oncogenic potential in mammals (Johnson et al., 2005; Mayr et al., 2007).

While *let-7* and *lin-4* were beginning to be understood in nematodes, other phenomena involving RNAs of similar lengths were uncovered based on unexpected observations of cellular response to foreign nucleic acids, such as viral genetic material or artificially introduced antisense silencing technologies. Double-stranded RNA (dsRNA) triggers a sequence-directed defensive mechanism in many multicellular organisms, originally described and termed posttranscriptional gene silencing (PTGS) in plants, also observed as quelling in fungi, and RNA interference (RNAi) in other animals (Fire et al., 1998; Mello and Conte, 2004). RNAi has likely evolved as an immune response to viral infection from cellular proteins involved in nucleic acid biogenesis. Most viruses, with the possible exception of retroviruses, generate dsRNA in infected eukaryotic cells. In reoviruses, the dsRNA genome—though undetected in its completely uncoated form in infected cells—may be the direct trigger. Viruses with single-stranded RNA (ssRNA) genomes, such as the influenza virus, require the RNA-directed RNA synthesis of replicative intermediate dsRNA. Cells infected with adenovirus, Herpes simplex virus, or some other DNA viruses, accumulate dsRNA as a result of overlapping convergent transcription (reviewed in (Jacobs and Langland, 1996)). Nonviral cellular sources of dsRNA include overlapping sense and antisense transcripts from inverted repeats of transgenes or transposons that anneal to form dsRNA, or more generally convergent promoters that produce overlapping transcripts; structured RNA that forms extended helices; or synthesis driven by RNA-dependent RNA-Polymerases (RdRP) from sense template RNA.

The RNAi trigger starts with processing of dsRNA from one end by the cellular RNase III Dicer into duplexes of short interfering RNAs (siRNAs) at 21-23-nt-long intervals (Zamore et al., 2000) releasing 5' phosphates and 2-nt 3' overhangs bearing 3' hydroxyls. Functional siRNAs guide the RNAi machinery to recognise target antisense RNA, or in organisms that have retained the RDRP gene, act as primers for RdRP-mediated extension of an antisense RNA template,

leading to amplification of dsRNA. Ultimately, RNAi leads to the destruction of mRNAs that share sequence complementarity with siRNAs generated from dsRNA.

The discoveries of long dsRNA-derived siRNAs, and conserved hairpin-derived miRNAs, prompted explorations into endogenous small-sized RNA, historically neglected as the zone of degraded material. Cloning and sequencing of a large variety of small RNAs were first reported in *C. elegans*, *drosophila* and human cells (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001). Cloning procedures were enriched for small RNAs by starting with size-fractionated total RNA. The most successful approach yielded more miRNA clones and less degraded fragments of ribosomal or transfer RNA, as it selected for RNAs with hallmarks of cleavage by the ribonuclease III Dicer: a length of ~22nt, a 5'-terminal monophosphate, and a 3'-terminal hydroxyl. Using radiolabelled size markers as guides, 18-26 nt RNAs were gel purified and sequentially ligated to adaptor molecules. The first ligation to a pre-adenylylated 3' adaptor oligonucleotide was catalysed by T4 RNA ligase in the absence of adenosine triphosphate (ATP) to avoid circularisation of RNA with 5' monophosphate. A standard T4 RNA ligase reaction to a 5' adaptor followed on gel-purified products of the first reaction. Final ligated products were gel-purified, reverse transcribed and amplified using primers that matched adaptor sequences. This cDNA pool was then digested, concatemerised, cloned and sequenced (Lau et al., 2001). Sequences that fell within fold-back structures like those of *lin-4* and *let-7*, were designated as miRNA genes, and further evidence for their expression was provided by Northern blot hybridisation.

Like the founder *lin-4* and *let-7*, other miRNAs are derived from conserved stem-loops with arms that are imperfectly paired, contain G:U wobbles, insertions or bulges, and mismatches. There is a strong bias for the accumulation of an miRNA from one arm of the hairpin but not both, though on occasion clones from the opposing arm are detected (1% in *C. elegans* (Ruby et al., 2006)). The most abundant strand is the miRNA, while the less stable strand is called the miRNA*. There is a substantial preference for a uracil base at the 5' end of the mature miRNA, and the 5'-half of the miRNA is more conserved than the 3'-half.

A vast majority of known miRNAs have been identified by relying on cDNA libraries of small RNAs, generated using the described cloning procedure or comparable approaches. In contrast,

very few miRNAs have been identified through genetic screens. These include *lin-4* and *let-7*, and subsequently, *bantam*, *lys-6*, and miRJAW, in flies, worms and plants, respectively (Hipfner et al., 2002; Johnston and Hobert, 2003; Palatnik et al., 2003). In general due to their small size, miRNAs are more likely to escape random mutagenesis. The redundancy owing to multiple copy families, and perhaps subtle roles in gene regulation, further make their disruption less likely to lead to readily observable and severe phenotypes in model organisms under common experimental conditions. Deletion mutants have been reported for nearly all *C. elegans* miRNAs, and indeed with the exception of *let-7* no single mutant was shown to have a lethal phenotype. The *bantam* locus and the corresponding small-bodied fruit fly were identified in a gain-of-function screen for *drosophila* genes that increased growth (Hipfner et al., 2002; Johnston and Hobert, 2003; Palatnik et al., 2003). The *bantam* miRNA stimulates cell proliferation and inhibits the pro-apoptotic gene *hid* (Brennecke et al., 2003). *lys-6* was identified in a genetic screen for mutants in neuronal left-right asymmetry in *C. elegans*. *lys-6* is only expressed in left taste receptor neurons, and targets the *cog-1* transcription factor involved in repression of left-specific chemoreceptors in right neurons. *lys-6* mutants show higher *cog-1* expression in left neurons, and reversals in the asymmetrical expression profiles of *gcy* chemoreceptors (Johnston and Hobert, 2003).

With the identification of a substantial number of miRNAs in *C. elegans* and other species, common features of the miRNA genes could be detailed. These characteristic features enabled computational methods to identify candidate miRNAs in genomes. Initial studies employed a training set of known miRNA sequences to develop algorithms that find genomic hairpins with similar features, relying on precursor structure, and patterns of phylogenetic conservation, and in some cases additional sequence motifs. A large set of hairpin structures can be found in genomes, scored, and ranked based on relative weights of various criteria. Structural and sequence features used by different algorithms have included hairpin length, loop length, distance of putative miRNA from the loop, overall thermodynamic stability, distribution of base-pairing, size, symmetry, and distribution of bulges, sequence composition, complexity and nucleotide content of the miRNA, identity of the 5' base of the putative miRNA. Conservation features have also been considered. The first prediction program developed to find miRNA genes, MiRscan, took conservation into account by limiting searches to hairpins within conserved sequences (or consensus hairpins), and led to an initial validation of 100 miRNAs in

C. elegans (Lim et al., 2003a; Ohler et al., 2004), and 52 in vertebrates (Lim et al., 2003b), with estimates of 120 and 200-255 total genes respectively in each lineage.

Candidate miRNAs must fit several criteria to be considered valid (Ambros et al., 2003). Sequences similar to existing miRNA representing paralogous loci of a miRNA family are considered valid. Other candidates are validated by detecting expression through Northern blot hybridisation, or random cloning and sequencing of cDNA libraries of small RNAs, or by directed PCR-amplification from similar libraries. Validation by Northern Blot is valuable as it provides qualitative information about tissue or stage specific expression levels, definitively identifies the miRNA length, and the presence of a stable precursor. It is however limited to abundant sequences from available tissue types. Identification by cloning has retrieved mostly abundant miRNA species, these are more represented among the cloned sequences, implying a positive correlation between expression levels and cloning frequency. The conservation-based computational approaches also bias against rare miRNAs, and it is now known that conserved miRNAs are more highly expressed; while rare and often tissue-specific miRNAs are often species-specific and not shared among distant genomes. The discovery of these rare miRNAs owes largely to traditional and high-throughput cloning from specific tissues.

Advances in high-throughput sequencing technologies have led to the discovery of additional miRNAs, siRNAs, and novel classes of small RNAs in plants (Lu et al., 2005; Rajagopalan et al., 2006) and animals (Berezikov et al., 2006; Ruby et al., 2006). These technologies, yielding millions of sequence reads directly from cDNA libraries of small RNAs, are likely to dramatically expand our understanding of transcriptomes, and will perhaps replace microarray technologies in profiling the expression of miRNAs. Berezikov et al. (2006) identified numerous repeat-associated siRNAs (rasiRNAs) from Human and Chimp brains, as well as 244 human miRNA candidates, of which the majority are very rare and thus represented by only one sequence read, implying that the sequencing has yet to be saturated. Of these, 50% were conserved in primates, 30% in mammals and 8 % were specific to humans (Berezikov et al., 2006). This study and others concluded that the total number of mammalian miRNAs is larger than initial estimates, perhaps ranging from 500 to thousands. The rare miRNAs likely represent sequences transcribed in a very small subset of cells, rather than sequences with low abundance present in many different cells. Confidently identified miRNA genes have reached 112 genes in

C. elegans, 30% of which are conserved to mammals (Ruby et al., 2006), and 21 conserved families and 44 non-conserved families in *Arabidopsis* (Rajagopalan et al., 2006). One-half of mammalian miRNAs arise from gene clusters and are transcribed as polycistronic primary transcripts. Many miRNAs originate from sense introns of protein-coding genes, in the sense orientation, estimates of the number of intronic miRNAs vary widely in the literature. Host transcript and the intronic miRNA generally correlate in expression (Baskerville and Bartel, 2005).

Biogenesis of animal miRNAs

Most microRNA genes that have been examined are transcribed by the RNA polymerase II (Pol II) as large primary transcripts (pri-miRNA) with the Pol II-characteristic 5' 7-methylguanosine cap (m7G), 3' poly A tail, and α -amanitin sensitivity (Lee et al., 2004). Pol II also localises to miRNA promoters (Cai et al., 2004), which are diverse and varied in regulation by transcription factors producing wide-ranging and tissue-specific expression patterns. Within the larger pri-miRNA lie one or more ~70-nt stem-loop structures that are typically excised at the base by the endonucleolytic cleavage of the ~160 kDa nuclear RNase III, Drosha, and its cofactor, the DiGeorge syndrome Critical Region gene 8 (DGCR8) in humans, or Pasha in *drosophila* and *C. elegans*, together forming the 650 kDa microprocessor complex. Knockdown of Drosha leads to accumulation of pri-miRNA in cultured cells (Lee et al., 2003). DGCR8 has two dsRNA-binding domains (dsRBD) and assists in substrate recognition and in positioning of the catalytic Drosha, by binding to the junction at the base of the pri-miRNA stem, and to flanking 5' and 3' ssRNA arms, thus defining the position of the scissile phosphates, ~11 bp or one helical turn up from the junction. DGCR8 distinguishes this junction from the loop-to-stem junction by preferential binding to the more flexible flanking ssRNA, relative to more constrained loop structures (Denli et al., 2004; Gregory et al., 2004; Han et al., 2004; Han et al., 2006; Landthaler et al., 2004; Lee et al., 2003). The two cleavage events produce a transient precursor (pre-miRNA) that bears a 5' phosphate and a 2-nt 3' overhang necessary for recognition by the Ran-dependent nuclear transport receptor, exportin-5 (Exp5), which mediates export to the cytoplasm. Knockdown of Exp5 does not lead to the accumulation of pre-miRNA (Lund et al., 2004; Yi et al., 2003).

Once in the cytoplasm, pre-miRNAs undergo a final round of catalytic processing similar to the fate of cytoplasmic dsRNAs, under the action of the cytoplasmic RNase III Dicer (Bernstein et al., 2001; Grishok et al., 2001; Hutvagner et al., 2001), which cuts twice, two helical turns away from the pre-miRNA or dsRNA termini to yield ~22-nt duplexes (Lee et al., 2003). Dicer is a ~200 kDa protein with dsRBD, helicase, PAZ and two RNase III domains. In their purified form, the unique human Dicer, or the *drosophila* Dcr-2 are sufficient to catalyse dsRNA cleavage. Like Drosha, Dicer associates with a cofactor, the human HIV-1 TAR RNA-binding protein (TRBP), which contains three dsRBDs, or with its redundant paralogue, PACT (Lee et al., 2006). There are two Dicers in *drosophila*, Dicer-1 (Dcr-1) and its cofactor R2D2 operate on pre-miRNAs to produce miRNA duplexes, while Dicer-2 (Dcr-2) and the R2D2 paralogue Loquacious (Loqs), cleave long dsRNAs into siRNA duplexes (Forstemann et al., 2005; Saito et al., 2005). Depletions of either Loquacious or Dcr-1 result in pre-miRNA accumulation in *drosophila* S2 cells (Forstemann et al., 2005; Saito et al., 2005).

The small RNA duplexes have 5' phosphates and 2-nt 3' overhangs, and one strand is destined to be the final functional miRNA or the siRNA guide strand, while the other, miRNA* or passenger strand for siRNAs is rapidly degraded. Mature miRNAs are incorporated into multiprotein complexes called miRNP or RNA-induced Silencing Complex (RISC). Assembly of RISC initiates unwinding of the miRNA duplex, and inclusion of the miRNA strand with the relatively more unstable 5' end. Thermodynamic asymmetry is also a determinant of strand fate in siRNA duplexes (Khvorova et al., 2003; Schwarz et al., 2003), and in turn, if both ends of the duplex have symmetrical stability, the two strands tend to accumulate with equal frequency.

Dcr-2 and R2D2 initiate loading of small RNAs onto the RISC (Liu et al., 2003; Tomari et al., 2004). R2D2 binds to the 5' end of small RNA duplex, with preference for a 5' phosphate rather than a 5' OH, likely involving one or both of its dsRBDs. Dcr-2 binds to the less stable 5' end of the presumed guide strand. The R2D2/Dcr-2 heterodimer orients the RNA duplex for unwinding of siRNAs in the presence of Ago2, resulting in retention of the guide strand and replacement of Dcr2/R2D2 by Ago2 on the guide strand (Liu et al., 2003; Tomari et al., 2004).

Recruitment of miRNAs to Ago2 in mammals also requires Dicer and TRBP (Chendrimada et al., 2005; Gregory et al., 2005). The ~ 500 kDa complex Dicer-TRBP-Ago2 represents a

minimal holo RISC and can process pre-miRNA or long dsRNA into a functional form that cleaves a target mRNA. A number of other factors associated with RISC have been identified. These proteins include Vasa intronic gene (VIG), Tudor-SN, Fragile X-related, putative RNA helicase Dmp68, Gemin3 (Caudy et al., 2002; Ishizuka et al., 2002; Mourelatos et al., 2002; Valencia-Sanchez et al., 2006), the RNA helicase MOV10 (the human homologue of *drosophila* Armitage), nearly all peptides of the 60S ribosomal particle, and, eIF6, a ribosomal inhibitor protein that prevents assembly of the translationally competent (80S) ribosome (Caudy et al., 2002; Hammond et al., 2000; Ishizuka et al., 2002; Martinez et al., 2002; Mourelatos et al., 2002; Valencia-Sanchez et al., 2006).

Loss of Dicer in mice leads to early lethality and roles for Dicer and the RNAi machinery have been suggested in the maintenance of stem-cell populations (Bernstein et al., 2003). A conditional knockout of Dicer in the mouse limb mesoderm results in the abolition of miRNA processing, and formation of a smaller limb due to massive cell death (Harfe et al., 2005).

In most eukaryotes, the components of RNAi, Dicers, Argonautes, RdRPs and others, exist in expanded families that reflect the advantages of pathway diversification (Vaucheret, 2006; Xie et al., 2004). Single-stranded miRNAs or siRNAs are very tightly bound (Martinez and Tuschl, 2004) to the ubiquitously expressed Agos, the major proteins implicated in RISC function, and one of two subfamily members of the Argonaute family. In *drosophila*, Ago1 and Ago2 belong to the Ago subfamily of Argonaute proteins and bind to miRNAs and siRNAs. Other *drosophila* Argonautes, Piwi, Aubergine, and Ago3, belong to the Piwi subfamily of Argonautes, are associated with repeat-associated siRNAs (rasiRNAs) or Piwi-interacting RNAs (piRNAs) and are involved in transposon silencing, heterochromatin formation, male germ-cell maturation and renewal, and other unknown functions (Seto et al., 2007). The classes of piRNAs, rasiRNAs, and other small RNAs, have distinct properties from miRNAs and siRNAs, often reflecting their differential biogenesis. In mammals, 26-31-nt-long piRNAs are derived from one strand of the genome within few clustered loci, and along their counterpart PIWI proteins are expressed in the male germline, where they are presumably involved in sperm development (Aravin et al., 2007; Girard et al., 2006; Lau et al., 2006).

Ago proteins possess an amino-terminal 110-residue Paz domain that contains a single-stranded RNA binding motif and might be involved in the recognition of 3' single-stranded overhangs and miRNA/siRNA binding (Lingel et al., 2003; Ma et al., 2004; Song et al., 2003; Yan et al., 2003). Agos also have a carboxy-terminal PIWI domain, more conserved than PAZ, which has been shown to contain a subdomain that is a structural homologue of RNase H, an endonuclease that cleaves RNA-DNA hybrids (Liu et al., 2004). Though not present in all Agos, two conserved aspartic acids and a third histidine/aspartic acid form a catalytic triad that provides PIWI with Mg^{2+} -dependent catalytic or 'slicer' activity, an endonucleolytic cleavage target mRNA paired to a small RNA by some Ago-associated RISCs at a position ten base pairs away from the 5' of the guide RNA (Hammond et al., 2000; Liu et al., 2004). PIWI also contains another subdomain—the most conserved region of Argonaute—that folds into a highly basic pocket where three amino acid side chains together with a divalent metal ion coordinate to the 5' phosphate of the guide RNA, and a fourth aromatic residue stacks with the first unpaired 5' base—usually a uracil. PIWI makes other contacts with the phosphate backbone to position the small RNA for proper binding to mRNA (Parker et al., 2005). In *drosophila*, silencing of an mRNA with complementary sites to small RNAs requires Ago2-RISC, while repression of a target with imperfect sites requires Ago1-RISC (Forstemann et al., 2007). Ago1 and Ago2 are functionally distinct, and both compete for loading of miRNAs. Ago2 requires Dcr-2/R2D2 for loading, suggesting that miRNA/miRNA* must dissociate from Dcr1/Loqs before loading onto RISC. Both Agos are capable of endonucleolytic cleavage, however the initial rate of target mRNA cleavage by Ago-2 is 12-fold faster than that of Ago-1. Furthermore while both Agos can bind target mRNAs with similar efficiencies, cleavage by Ago1 does not exhibit multiple turnover. Ago1 is likely to slow to silence targets by cleavage *in vivo*, and may have only retained its endonucleolytic ability to degrade the miRNA* to enable Ago1-RISC loading (Forstemann et al., 2007).

In humans, Ago2 is capable of slicer activity and bears the necessary functional catalytic triad (Rivas et al., 2005; Song et al., 2004). Ago3 also has the required catalytic DDH sequence, but is catalytically inactive, while Ago1 and Ago4 do not have functional catalytic triads. In human cells, miRNAs appear to be equally distributed within the four human Agos. (Liu et al., 2004; Meister et al., 2004). Ago2-deficient mice die as embryos, and by mid-gestation display several

non-specific abnormalities such as failure to close the neural tube, mispatterning of the forebrain and an enlarged heart (Liu et al., 2004).

Once programmed into RISC, small RNAs direct the complex to mRNAs through recognition of complementary sequences and trigger silencing. If the small RNA is perfectly complementary to the mRNA, as in an artificially introduced siRNA, the slicer activity of RISC causes cleavage of the mRNA, with the scissile phosphate located 10 nucleotides away from the base paired to the 5' end of the small RNA. This cleavage can happen regardless of where the complementary target site is located. If the miRNA is partially complementary to its target, protein expression is reduced. Although miRNAs accumulate and appear to be highly stable little is understood about their decay process.

Mechanisms of repression by animal miRNAs

Most miRNAs in animals direct RISC to endogenous target mRNAs through the recognition of sites with partial complementarity in 3' UTRs, resulting in the direct repression of translation and to a lesser extent in the destabilization of mRNAs. If the miRNA site is perfectly complementary, Ago2-mediated cleavage ensues, and the destabilized cleavage products are subject to nuclease-mediated decay. In most endogenous targets, the mRNA degradation is not caused by the site-specific slicer cleavage, and it does not fully account for the reduction in protein synthesis. Profiles of mRNA from human cells 12 hours after transfection with miR-124 or miR-1 were examined by microarrays. In each case, about 100 messages bearing 3' UTRs with a significant propensity to pair to the 5' region of the miRNA were downregulated (Lim et al., 2005). In worms, mRNA levels of a LacZ transgene fused to the 3' UTR of the *let-7* target *lin-41* decrease between L2 and L3 stages, coinciding with the onset of *let-7* expression. This decrease depends on the presence of two closely-spaced *let-7*-complementary sites with central mismatches, and occurs at the posttranscriptional level (Bagga et al., 2005). These studies suggest that targeting by miRNAs lowers mRNA stability. However, for some messages the majority of the initial silencing is due to a reduction in protein levels. Further evidence has shown that targeting by miRNAs in RISC destines mRNAs for degradation at Processing bodies (P-body), which are dynamic cytoplasmic aggregates of RNA and protein, associated with factors involved in mRNA decay and translational repression, including the 5'-3' exonuclease

Xrn1 (Pillai et al., 2007). P-body foci lack ribosomes and translation initiation factors except the eukaryotic Initiation Factor 4E (eIF4E), and in fact polysome association is inversely correlated with P-body accumulation of mRNAs consistent with P-bodies as storage and degradation sites for translationally repressed mRNAs (Pillai et al., 2007). The storage of mRNAs may also be temporary and reversible. In yeast, mRNAs can cycle in and out of P-bodies (Brenques et al., 2005), and mRNAs within P-bodies can be released upon induced conditions of stress. Agos, miRNAs and their target mRNAs, have been shown to colocalise and co-immunoprecipitate in P-bodies (Pillai et al., 2007).

Mechanistic details of miRNA-mediated translational repression are not entirely understood. A subdomain within the middle domain of Ago2, has homology to the cap-binding domain of eIF4E, a motif that binds the m7G cap at 5' ends of eukaryotic mRNAs. In one study (Kiriakidou et al., 2007), Ago2 bound specifically to a cap analogue resin, and binding was competed off with the addition of free m7GpppG cap analogue. Mutations at crucial and conserved Phenylalanine residues perturbed cap binding, and disrupted translation of a reporter in an Ago-tethering assay, (where Ago2 alone tethered to elements in a 3' UTR can mediate repression in the absence of miRNAs). Mutations also disrupted the ability to inhibit translational repression as assayed by polysome profiles. However, substitutions of these critical residues did not impair *in vitro* cleavage of a perfectly complementary target. These results suggest a model where Ago2 competes with eIF4E for binding to the 5' cap, and blocks initiation of translation (Kiriakidou et al., 2007).

Further evidence for this mechanism of repression was obtained from another study involving incubation of reporter mRNAs and *in vitro* translation extracts (Mathonnet et al., 2007). The reporters contained six let-7-complementary sites with central mismatches in their 3' UTR. A repression of the reporter protein was observed, along with a reduction in the association of reporter mRNAs with the 80S ribosomal complex. This Inhibition did not occur when reporter mRNAs contained viral internal ribosomal entry sites; when the 5' cap was modified; or upon the addition of excess eIF4F. These results confirm suggestions that the initial step in miRNA-mediated translational repression involves inhibition of translational initiation through interferences with the recognition of the 5' cap (Mathonnet et al., 2007). This conclusion is however not in agreement with that of others, which suggest that inhibition affects elongation

and causes premature termination (Petersen et al., 2006). It is possible that different aspects of translation, the stability of the nascent peptide and the targeted mRNA are all affected with different kinetics, contributing to the overall silencing effects of miRNAs.

Target recognition

The original observation of base-pairing between *lin-4* and *let-7* and target mRNAs paved the way for the use of computational approaches to identify targets of all miRNAs. These approaches are complicated by the fact that the base-pairing is partial, flexible, and mismatches are tolerated, leading to the inclusion of many false positives in predictions.

Many miRNA target prediction algorithms have been applied on metazoan genomes. The targetsScan algorithm (Lewis et al., 2005; Lewis et al., 2003) concurs with biochemical, genetic, and phylogenetic evidence that pairing to the 5' end of the miRNA is more critical than pairing to the 3' end (Doench and Sharp, 2004; Lai, 2002; Mallory et al., 2004). In particular uninterrupted pairing to miRNA nucleotides 2-8, termed the 'seed' is sometimes sufficient to confer repression, while additional pairing to miRNA nucleotides 13-16, increases the probability of suppression (Grimson et al., 2007; Lewis et al., 2005; Lewis et al., 2003). Other determinants of specificity have also been found. The mRNA sequence paired to the miRNA seed is referred to as the 'seed match', and is often flanked by an adenosine at the 3' end, and AU-rich sequences at both ends, consistent with preference for unstructured surrounding RNA. The proximity of a site to other matches to coexpressed miRNAs enables cooperative action, and increases the likelihood of repression; as do positioning of the seed match at least 15 nt from the stop codon within the 3' UTR, and away the central region of longer UTRs (Grimson et al., 2007; Nielsen et al., 2007). In general conserved target sites, that is seed matches falling within regions of high alignment are predicted with more confidence, but it is likely that many species-specific sites have also evolved even for conserved miRNAs. An mRNA may have evolved target sites against miRNAs expressed in tissues other than the endogenous mRNA as a mechanism to prevent effects of aberrant or low-level transcription. Messages are also targeted by miRNAs in cases (usually involving cellular differentiation) where mRNA levels must be rapidly cleared where miRNAs act as switches, or in cases where mRNA expression levels need to be brought to appropriate levels, or be fine-tuned and dampened.

Plant miRNAs

The miRNA pathways of plants and metazoans likely evolved independently built upon common ancestral protein coding genes, with numerous differences between *Arabidopsis* miRNAs and animal miRNAs (Reinhart et al., 2002; Vaucheret, 2006). Conserved plant miRNAs exist as expanded gene families and are mostly transcribed as independent transcription units. The pre-miRNAs lie within hairpin structures that are both longer and more variable in length relative to animal miRNAs. They are processed by nuclear Dicers, and their biogenesis requires DCL1, which likely possesses both Drosha and Dicer functions (Kurihara and Watanabe, 2004; Papp et al., 2003; Park et al., 2002; Reinhart et al., 2002), and its dsRBD-containing cofactor *HYL1* (Hiraguri et al., 2005; Vazquez et al., 2004). A few miRNAs appear to be processed by DCL4 (Rajagopalan et al., 2006), a nuclear Dicer that also excises trans-acting siRNAs (tasiRNAs) from long noncoding precursor dsRNAs. These initiate as ssRNA primary transcripts (TAS) whose ends are specified by miRNA-directed cleavage. They require an RdRP (RDR6) with cofactor SGS3 to form dsRNA, which is then processed into 21-nt tasiRNAs by DCL4 (reviewed in (Vaucheret, 2006)). Like miRNAs, tasiRNAs direct the cleavage of target mRNAs or their own TAS transcript precursors. The 3' ends of mature miRNAs, siRNAs and tasiRNAs are methylated by the dsRNA methylase HEN1, a modification likely to protect small RNAs from degradation and polyuridylation (Li et al., 2005; Yu et al., 2005). They are exported into the nucleus by the importin- β family of nucleocytoplasmic transporters, *HST*, a plant orthologue of Exportin-5 (Park et al., 2005). Plants miRNA target sites occur within coding or untranslated regions of mRNAs (Reinhart et al., 2002; Rhoades et al., 2002). Most display four or less mismatches to the 3' region of miRNAs. This degree of conserved complementarity confers confidence to the validity of plant miRNA targets computational prediction strategies (Rhoades et al., 2002). Plant miRNAs direct mRNA cleavage through the slicer activity of AGO1 as a primary silencing mechanism (Fagard et al., 2000), with detection of validation of cleaved 3' fragments by 5' RACE (Llave et al., 2002; Tang et al., 2003). The mRNA cleavage products accumulate and are more stable than their rare animal counterparts. Plant miRNAs target less than 1% of protein coding genes with bias for transcription factors (Jones-Rhoades and Bartel, 2004; Rhoades et al., 2002). Finally it may be that miRNAs, like plant siRNAs, may be mobile along long distances through vascular tissue, and may spread both locally and systemically

(Vaucheret, 2006). In animals cell-to-cell spreading of an RNAi trigger has only been observed in *C. elegans*, and appears to be mediated by the membrane channel *sid-1* (Winston et al., 2002).

Diversity of miRNA function

Similar to other classes of regulatory genes, miRNAs play roles in diverse biological processes of multi-cellular life. Several interesting cases are considered below.

The muscle-specific and highly conserved miR-1, when deleted in *drosophila*, results in embryonic lethality with defects in muscle differentiation and maintenance (Kwon et al., 2005), or in larval lethality, with compromised movement and deformed musculature prior to death (Sokol and Ambros, 2005). Overexpression of the miR-1 in the cardiac mesoderm leads to a lowered number of cardiac progenitor cells (Kwon et al., 2005). Overexpression of this miRNA in mouse cardiac progenitors reduces proliferation, and in myoblast cultures causes skeletal muscle differentiation and reduced proliferation (Chen et al., 2006; Zhao et al., 2005). Misexpression of miR-1 in *Xenopus laevis* embryos reduced anterior structures, and anterior-posterior axis length; disrupted segmentation and somite formation; and resulted in a complete absence of cardiac tissue (Chen et al., 2006). Humans with coronary artery disease have elevated miR-1 expression, as do rats whose hearts have been subjected to induced myocardial infarction (Yang et al., 2007). Antisense inhibition of miR-1 function in infarcted rat hearts relieves arrhythmia, while delivery of miR-1 into healthy hearts induces arrhythmia (Yang et al., 2007). Knockouts of one the two miR-1 loci in mouse cause abnormal heart morphogenesis and electrophysiology, lowered heart rate, thickening of the walls of the heart and early lethality. miR-1 targets Hand2, a transcription factor involved in myocyte expansion. Protein levels of the target Hand2 are upregulated and myocytes overproliferate in miR-1-2-null embryos (Zhao et al., 2007).

Viruses that encode miRNAs can target host mRNAs, or regulate their viral mRNAs to promote latency of virulence, thus ironically using parts of a pathway evolved originally as a viral immune response. The Simian virus 40 (SV40) is a polyomavirus that infects humans, monkeys and some other mammals with persistent latency and oncogenic potential. Its dsDNA genome encodes a pre-miRNA within a late viral transcript that gives rise to stable miRNAs from both arms of the hairpin, expressed late in infection (Sullivan et al., 2005). These also overlap in the

antisense with the 3' UTR region of the early transcript of the large T antigen, and thus are perfectly complementary to two UTR sites. miRNA-directed cleavage fragments of T antigen can be detected in monkey kidney cells infected with SV40. Downregulation of T antigen lowers the cytotoxic T lymphocyte response to viral infections, triggering less cytokine production, and further lowers another cellular response by limiting interferon- γ release. In the case of SV40, autoregulation of a viral mRNA by viral miRNAs leads to lowered susceptibility to a cellular response to viral infection (Sullivan et al., 2005) and thus increases latency of infection.

Dominant mutations in two related *Arabidopsis* HD-ZIP III transcription factors *PHABULOTA* and *PHAVOLUTA* occur within a binding site for miR-165/166 and disrupt the potential for pairing (Rhoades et al., 2002). The mutations cause a change in mature leaf polarity, where upper and lower surfaces of the leaf are switched, due to abaxial to adaxial transformations in the fates of leaf primordia (McConnell and Barton, 1998). Synonymous substitutions within the miR165/166 site of *REVOLUTA*, another member of the HD-ZIP III family and a homologue of *PHABULOTA* and *PHAVOLUTA*, also lead to the leaf phenotype. The miRNA precursors have a biased expression in the abaxial domain of leaf primordia where they are responsible for clearing of transcription factors. miR165/166 direct the *in vitro* cleavage of the *PHABULOTA* and *PHAVOLUTA* messages, but are less efficient at cleaving the dominant mutant form, which bears a 33-nt insertion at the scissile phosphate (Tang et al., 2003). Adding a single silent mutation in the miRNA site of *PHABULOTA* at 5-nt from the 5' end of miR165/166 causes a dramatic reduction in cleavage efficiency both *in vivo* and *in vitro*, and result in plants that phenocopy the dominant *PHABULOTA* leaf phenotype (Mallory et al., 2004). Thus miRNA regulation of HD-ZIP III transcription factors is required for the correct cell fate specification within leaf primordia, and its disruption leads to transformation of organ polarity.

Summary of thesis

The work presented here starts with an early chronicle of miRNA sequences. The first chapter reports the prediction and validation of conserved vertebrate miRNA genes, a collaborative effort with Lee Lim, a former joint member of Chris Burge's group, who developed and applied the MiRscan algorithm to multiple vertebrate genomes, and Margy Glasner who prepared and cloned cDNA libraries of small RNAs from zebrafish maintained by Hazel Sive's laboratory. My contribution consisted of the development of a directed PCR-based approach and its use to validate predicted targets (Lim et al., 2003b).

The second chapter focuses on the identification of miRNA targets, and mechanisms of silencing by a family of vertebrate-specific miRNAs. In collaboration with I-Hung Shih, we reported the regulation of Hox mRNAs by miR-196, a miRNA family located within Hox intergenic regions (Yekta et al., 2004). This miRNA had a conserved target site with near-perfect complementarity in the 3' UTR of *Hoxb8*. Very few animal miRNAs use the slicer activity of Ago2 to destroy their targets, and Ago2 catalytic function likely mostly mediates siRNA-directed cleavage events. Most target mRNAs have short seed matches and operate through silencing mechanisms not requiring mRNA cleavage, and involving translational repression. The *Hoxb8* mRNA appeared as an exception, and was targeted for *in vitro* and *in vivo* cleavage by miR-196, implying that the RNAi pathway intersects with miRNA pathways at least in a few cases to regulate endogenous transcripts. Another example of miRNA-mediated cleavage, involves multiple miRNAs that are expressed from a maternally imprinted locus, antisense to the paternally imprinted *Rtl1* (Davis et al., 2005; Seitz et al., 2003). These miRNAs form perfectly complementary target sites to the message coded in the antisense of their locus, and differ from the regulation of *Hoxb8* in that they derive from the locus they target, and in the nature of the target RNA. The target transcript, *Rtl1*, is a retrotransposon-like gene specific to eutherian mammals, which contains a predicted 4 KB open reading frame, with no evidence for the production of a protein (Davis et al., 2005).

Chapter three expands the targeting potential of miR-196 and another miRNA located in the Hox cluster, miR-10. A more complete analysis of Hox targets of these miRNAs revealed a genomic distribution that significantly favours targeting of Hox genes that are anterior in expression

relative to the miRNA loci. The functional roles of miR-196 and miR-10 were explored in developing chick embryos in collaboration with Jennifer Mansfield, a former member Cliff Tabin's laboratory, and Eddy McGlenn, a present member. Introduction of an antisense reagent designed to functionally block miR-196 and miR-10 led to skeletal defects consistent with their roles as Hox regulators. I propose that gene networks evolve at multiple layers of gene expression by repeating similar modes of regulation and thus reinforcing a final functional outcome.

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.
- Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K., and Hannon, G. J. (2007). Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* 316, 744-747.
- 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.
- 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., 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.
- Bernstein, E., Caudy, A. A., Hammond, S. M., and Hannon, G. J. (2001). Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409, 295-296.
- Bernstein, E., Kim, S. Y., Carmell, M. A., Murchison, E. P., Alcorn, H., Li, M. Z., Mills, A. A., Elledge, S. J., Anderson, K. V., and Hannon, G. J. (2003). Dicer is essential for mouse development. *Nat Genet* 35, 215-217.
- Bregues, M., Teixeira, D., and Parker, R. (2005). Movement of eukaryotic mRNAs between polysomes and cytoplasmic processing bodies. *Science* 310, 486-489.
- 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.
- Cai, X., Hagedorn, C. H., and Cullen, B. R. (2004). Human microRNAs are processed from capped, polyadenylated transcripts that can also function as mRNAs. *Rna* 10, 1957-1966.
- Caudy, A. A., Myers, M., Hannon, G. J., and Hammond, S. M. (2002). Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes Dev* 16, 2491-2496.

- Chalfie, M., Horvitz, H. R., and Sulston, J. E. (1981). Mutations that lead to reiterations in the cell lineages of *C. elegans*. *Cell* 24, p59-69.
- Chen, J. F., Mandel, E. M., Thomson, J. M., Wu, Q., Callis, T. E., Hammond, S. M., Conlon, F. L., and Wang, D. Z. (2006). The role of microRNA-1 and microRNA-133 in skeletal muscle proliferation and differentiation. *Nat Genet* 38, 228-233.
- Chendrimada, T. P., Gregory, R. I., Kumaraswamy, E., Norman, J., Cooch, N., Nishikura, K., and Shiekhattar, R. (2005). TRBP recruits the Dicer complex to Ago2 for microRNA processing and gene silencing. *Nature* 436, 740-744.
- 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.
- Denli, A. M., Tops, B. B., Plasterk, R. H., Ketting, R. F., and Hannon, G. J. (2004). Processing of primary microRNAs by the Microprocessor complex. *Nature* 432, 231-235.
- Doench, J. G., and Sharp, P. A. (2004). Specificity of microRNA target selection in translational repression. *Genes Dev* 18, 504-511.
- Fagard, M., Boutet, S., Morel, J. B., Bellini, C., and Vaucheret, H. (2000). AGO1, QDE-2, and RDE-1 are related proteins required for post-transcriptional gene silencing in plants, quelling in fungi, and RNA interference in animals. *Proc Natl Acad Sci U S A* 97, 11650-11654.
- 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.
- Forstemann, K., Horwich, M. D., Wee, L., Tomari, Y., and Zamore, P. D. (2007). *Drosophila* microRNAs are sorted into functionally distinct argonaute complexes after production by dicer-1. *Cell* 130, 287-297.
- Forstemann, K., Tomari, Y., Du, T., Vagin, V. V., Denli, A. M., Bratu, D. P., Klattenhoff, C., Theurkauf, W. E., and Zamore, P. D. (2005). Normal microRNA maturation and germ-line stem cell maintenance requires Loquacious, a double-stranded RNA-binding domain protein. *PLoS Biol* 3, e236.
- Girard, A., Sachidanandam, R., Hannon, G. J., and Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature* 442, 199-202.
- Gregory, R. I., Chendrimada, T. P., Cooch, N., and Shiekhattar, R. (2005). Human RISC couples microRNA biogenesis and posttranscriptional gene silencing. *Cell* 123, 631-640.
- Gregory, R. I., Yan, K. P., Amuthan, G., Chendrimada, T., Doratotaj, B., Cooch, N., and Shiekhattar, R. (2004). The Microprocessor complex mediates the genesis of microRNAs. *Nature* 432, 235-240.

- 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.
- Grishok, A., Pasquinelli, A. E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D. L., Fire, A., Ruvkun, G., and Mello, C. C. (2001). Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* 106, 23-34.
- Hammond, S. C., Bernstein, E., Beach, D., and Hannon, G. J. (2000). An RNA-directed nuclease mediates posttranscriptional gene silencing in *Drosophila* cells. *Nature* 404, 293-296.
- Han, J., Lee, Y., Yeom, K. H., Kim, Y. K., Jin, H., and Kim, V. N. (2004). The Drosha-DGCR8 complex in primary microRNA processing. *Genes Dev* 18, 3016-3027.
- 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.
- Harfe, B. D., McManus, M. T., Mansfield, J. H., Hornstein, E., and Tabin, C. J. (2005). The RNaseIII enzyme Dicer is required for morphogenesis but not patterning of the vertebrate limb. *Proc Natl Acad Sci U S A* 102, 10898-10903.
- Hipfner, D. R., Weigmann, K., and Cohen, S. M. (2002). The bantam gene regulates *Drosophila* growth. *Genetics* 161, p1527-1537.
- Hiraguri, A., Itoh, R., Kondo, N., Nomura, Y., Aizawa, D., Murai, Y., Koiwa, H., Seki, M., Shinozaki, K., and Fukuhara, T. (2005). Specific interactions between Dicer-like proteins and HYL1/DRB-family dsRNA-binding proteins in *Arabidopsis thaliana*. *Plant Mol Biol* 57, 173-188.
- 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.
- Ishizuka, A., Siomi, M. C., and Siomi, H. (2002). A *Drosophila* fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes Dev* 16, 2497-2508.
- Jacobs, B. L., and Langland, J. O. (1996). When two strands are better than one: the mediators and modulators of the cellular responses to double-stranded RNA. *Virology* 219, 339-349.
- Johnson, S. M., Grosshans, H., Shingara, J., Byrom, M., Jarvis, R., Cheng, A., Labourier, E., Reinert, K. L., Brown, D., and Slack, F. J. (2005). RAS is regulated by the *let-7* microRNA family. *Cell* 120, 635-647.
- Johnston, R. J., and Hobert, O. (2003). A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* 426, 845-849.

- Jones-Rhoades, M. W., and Bartel, D. P. (2004). Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol Cell* *14*, 787-799.
- Khvorova, A., Reynolds, A., and Jayasena, S. D. (2003). Functional siRNAs and miRNAs exhibit strand bias. *Cell* *115*, 209-216.
- Kiriakidou, M., Tan, G. S., Lamprinaki, S., De Planell-Saguer, M., Nelson, P. T., and Mourelatos, Z. (2007). An mRNA m7G cap binding-like motif within human Ago2 represses translation. *Cell* *129*, 1141-1151.
- Kurihara, Y., and Watanabe, Y. (2004). Arabidopsis micro-RNA biogenesis through Dicer-like 1 protein functions. *Proc Natl Acad Sci U S A* *101*, 12753-12758.
- Kwon, C., Han, Z., Olson, E. N., and Srivastava, D. (2005). MicroRNA1 influences cardiac differentiation in *Drosophila* and regulates Notch signaling. *Proc Natl Acad Sci U S A* *102*, 18986-18991.
- Lagos-Quintana, M., Rauhut, R., Lendeckel, W., and Tuschl, T. (2001). Identification of novel genes coding for small expressed RNAs. *Science* *294*, 853-858.
- Lai, E. C. (2002). MicroRNAs are complementary to 3'UTR motifs that mediate negative post-transcriptional regulation. *Nature Genetics* *30*, 363-364.
- Landthaler, M., Yalcin, A., and Tuschl, T. (2004). The human DiGeorge syndrome critical region gene 8 and its *D. melanogaster* homolog are required for miRNA biogenesis. *Curr Biol* *14*, 2162-2167.
- 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.
- Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., and Kingston, R. E. (2006). Characterization of the piRNA complex from rat testes. *Science* *313*, 363-367.
- 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., Hur, I., Park, S. Y., Kim, Y. K., Suh, M. R., and Kim, V. N. (2006). The role of PACT in the RNA silencing pathway. *Embo J* *25*, 522-532.

- 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., Jones-Rhoades, M. W., Bartel, D. P., and Burge, C. B. (2003). Prediction of mammalian microRNA targets. *Cell* 115, 787-798.
- Li, J., Yang, Z., Yu, B., Liu, J., and Chen, X. (2005). Methylation protects miRNAs and siRNAs from a 3'-end uridylation activity in Arabidopsis. *Curr Biol* 15, 1501-1507.
- Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B., and Bartel, D. P. (2003b). 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. (2003a). The microRNAs of *Caenorhabditis elegans*. *Genes Dev* 17, 991-1008.
- Lingel, A., Simon, B., Izaurralde, E., and Sattler, M. (2003). Structure and nucleic-acid binding of the Drosophila Argonaute 2 PAZ domain. *Nature* 426, 465-469.
- 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, Q., Rand, T. A., Kalidas, S., Du, F., Kim, H. E., Smith, D. P., and Wang, X. (2003). R2D2, a bridge between the initiation and effector steps of the Drosophila RNAi pathway. *Science* 301, 1921-1925.
- 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.
- Lu, C., Tej, S. S., Luo, S., Haudenschild, C. D., Meyers, B. C., and Green, P. J. (2005). Elucidation of the small RNA component of the transcriptome. *Science* 309, 1567-1569.
- Lund, E., Güttinger, S., Calado, A., Dahlberg, J. E., and Kutay, U. (2004). Nuclear export of microRNA precursors. *Science* 303, 95-98 Published online on November 20, 2003, 2010.1126/science.1090599.
- Ma, J. B., Ye, K., and Patel, D. J. (2004). Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. *Nature* 429, 318-322.

- Mallory, A. C., Reinhart, B. J., Jones-Rhoades, M. W., Tang, G., Zamore, P. D., Barton, M. K., and Bartel, D. P. (2004). MicroRNA control of PHABULOSA in leaf development: importance of pairing to the microRNA 5' region. *Embo J* 23, 3356-3364.
- Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. (2002). Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* 110, 563-574.
- Martinez, J., and Tuschl, T. (2004). RISC is a 5' phosphomonoester-producing RNA endonuclease. *Genes Dev* 18, 975-980.
- Mathonnet, G., Fabian, M. R., Svitkin, Y. V., Parsyan, A., Huck, L., Murata, T., Biffo, S., Merrick, W. C., Darzynkiewicz, E., Pillai, R. S., *et al.* (2007). MicroRNA Inhibition of Translation Initiation in Vitro by Targeting the Cap-Binding Complex eIF4F. *Science*.
- Mayr, C., Hemann, M. T., and Bartel, D. P. (2007). Disrupting the pairing between let-7 and Hmga2 enhances oncogenic transformation. *Science* 315, 1576-1579.
- McConnell, J. R., and Barton, M. K. (1998). Leaf polarity and meristem formation in *Arabidopsis*. *Development* 125, 2935-2942.
- Meister, G., Landthaler, M., Patkaniowska, A., Dorsett, Y., Teng, G., and Tuschl, T. (2004). Human Argonaute2 mediates RNA cleavage targeted by miRNAs and siRNAs. *Mol Cell* 15, 185-197.
- Mello, C. C., and Conte, D., Jr. (2004). Revealing the world of RNA interference. *Nature* 431, 338-342.
- 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.
- Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev* 16, 720-728.
- Nielsen, C. B., Shomron, N., Sandberg, R., Hornstein, E., Kitzman, J., and Burge, C. B. (2007). Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *Rna*.
- 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*, in press.
- Palatnik, J. F., Allen, E., Wu, X., Schommer, C., Schwab, R., Carrington, J. C., and Weigel, D. (2003). Control of leaf morphogenesis by microRNAs. *Nature* 425, 257-263.
- Papp, I., Mette, M. F., Aufsatz, W., Daxinger, L., Schauer, S. E., Ray, A., van der Winden, J., Matzke, M., and Matzke, A. J. (2003). Evidence for nuclear processing of plant micro RNA and short interfering RNA precursors. *Plant Physiol* 132, 1382-1390.

- Park, M. Y., Wu, G., Gonzalez-Sulser, A., Vaucheret, H., and Poethig, R. S. (2005). Nuclear processing and export of microRNAs in Arabidopsis. *Proc Natl Acad Sci U S A* *102*, 3691-3696.
- Park, W., Li, J., Song, R., Messing, J., and Chen, X. (2002). CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in Arabidopsis thaliana. *Curr Biol* *12*, 1484-1495.
- Parker, J. S., Roe, S. M., and Barford, D. (2005). Structural insights into mRNA recognition from a PIWI domain-siRNA guide complex. *Nature* *434*, 663-666.
- Pasquinelli, A. E., Reinhart, B. J., Slack, F., Martindale, M. Q., Kuroda, M., Maller, B., Srinivasan, A., Fishman, M., Hayward, D., Ball, E., *et al.* (2000). Conservation across animal phylogeny of the sequence and temporal regulation of the 21 nucleotide *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., and Filipowicz, W. (2007). Repression of protein synthesis by miRNAs: how many mechanisms? *Trends Cell Biol* *17*, 118-126.
- Rajagopalan, R., Vaucheret, H., Trejo, J., and Bartel, D. P. (2006). A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana. *Genes Dev* *20*, 3407-3425.
- Reinhart, B. J., Slack, F. J., Basson, M., Bettinger, J. C., Pasquinelli, A. E., 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.
- Rivas, F. V., Tolia, N. H., Song, J. J., Aragon, J. P., Liu, J., Hannon, G. J., and Joshua-Tor, L. (2005). Purified Argonaute2 and an siRNA form recombinant human RISC. *Nat Struct Mol Biol* *12*, 340-349.
- 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.
- Saito, K., Ishizuka, A., Siomi, H., and Siomi, M. C. (2005). Processing of pre-microRNAs by the Dicer-1-Loquacious complex in *Drosophila* cells. *PLoS Biol* *3*, e235.
- 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., Youngson, N., Lin, S. P., Dalbert, S., Paulsen, M., Bachellerie, J. P., Ferguson-Smith, A. C., and Cavaille, J. (2003). Imprinted microRNA genes transcribed antisense to a reciprocally imprinted retrotransposon-like gene. *Nat Genet* *34*, 261-262.
- Seto, A. G., Kingston, R. E., and Lau, N. C. (2007). The coming of age for Piwi proteins. *Mol Cell* *26*, 603-609.
- Slack, F. J., Basson, M., Liu, Z., Ambros, V., Horvitz, H. R., and Ruvkun, G. (2000). The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. *Molecular Cell* *5*, 659-669.
- Sokol, N. S., and Ambros, V. (2005). Mesodermally expressed *Drosophila* microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes Dev* *19*, 2343-2354.
- Song, J. J., Liu, J., Tolia, N. H., Schneiderman, J., Smith, S. K., Martienssen, R. A., Hannon, G. J., and Joshua-Tor, L. (2003). The crystal structure of the Argonaute2 PAZ domain reveals an RNA binding motif in RNAi effector complexes. *Nat Struct Biol* *10*, 1026-1032.
- Song, J. J., Smith, S. K., Hannon, G. J., and Joshua-Tor, L. (2004). Crystal structure of Argonaute and its implications for RISC slicer activity. *Science* *305*, 1434-1437.
- Sullivan, C. S., Grundhoff, A. T., Tevethia, S., Pipas, J. M., and Ganem, D. (2005). SV40-encoded microRNAs regulate viral gene expression and reduce susceptibility to cytotoxic T cells. *Nature* *435*, 682-686.
- 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.
- Tomari, Y., Matranga, C., Haley, B., Martinez, N., and Zamore, P. D. (2004). A protein sensor for siRNA asymmetry. *Science* *306*, 1377-1380.
- 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.
- Vaucheret, H. (2006). Post-transcriptional small RNA pathways in plants: mechanisms and regulations. *Genes Dev* *20*, 759-771.
- Vazquez, F., Gascioli, V., Crete, P., and Vaucheret, H. (2004). The nuclear dsRNA binding protein HYL1 is required for microRNA accumulation and plant development, but not posttranscriptional transgene silencing. *Curr Biol* *14*, 346-351.
- 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.
- Winston, W. M., Molodowitch, C., and Hunter, C. P. (2002). Systemic RNAi in *C. elegans* requires the putative transmembrane protein SID-1. *Science* *295*, 2456-2459.

Xie, Z., Johansen, L. K., Gustafson, A. M., Kasschau, K. D., Lellis, A. D., Zilberman, D., Jacobsen, S. E., and Carrington, J. C. (2004). Genetic and functional diversification of small RNA pathways in plants. *PLoS Biol* 2, E104.

Yan, K. S., Yan, S., Farooq, A., Han, A., Zeng, L., and Zhou, M. M. (2003). Structure and conserved RNA binding of the PAZ domain. *Nature* 426, 468-474.

Yang, B., Lin, H., Xiao, J., Lu, Y., Luo, X., Li, B., Zhang, Y., Xu, C., Bai, Y., Wang, H., *et al.* (2007). The muscle-specific microRNA miR-1 regulates cardiac arrhythmogenic potential by targeting GJA1 and KCNJ2. *Nat Med* 13, 486-491.

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.

Yu, B., Yang, Z., Li, J., Minakhina, S., Yang, M., Padgett, R. W., Steward, R., and Chen, X. (2005). Methylation as a crucial step in plant microRNA biogenesis. *Science* 307, 932-935.

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 one

Vertebrate microRNA genes

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

ABSTRACT

Much of the mammalian genome is transcribed into RNA that does not encode protein. One class of noncoding RNA (ncRNA) transcripts gives rise to microRNAs (miRNAs), small ncRNAs that can act by base-pairing to mRNAs to target these transcripts for destruction or translational repression. The extent to which miRNAs are involved in gene regulatory networks is unknown, in part because many of the miRNAs and the genes that encode them have not been identified, despite extensive cloning efforts. We have developed a computational procedure to identify miRNA genes conserved in more than one genome. Here, we apply this program, known as MiRscan, to the identification of vertebrate miRNA genes. Starting with an alignment of the mouse and human genomes, with subsequent comparison to fish sequences, MiRscan identified 188 vertebrate miRNA gene candidates, including three quarters of the genes previously identified through cloning of miRNA cDNAs. Many of the newly identified candidates were validated using libraries of miRNA cDNAs from zebrafish. Our analysis appears to have detected the majority of vertebrate miRNA genes and indicates that no more than 40 miRNA genes remain to be identified in mammals. In humans and other vertebrates, nearly one percent of the genes code for miRNAs—a fraction similar to that seen for other very large gene families with regulatory roles, such as those encoding transcription factor proteins.

MicroRNAs (miRNAs) are an abundant class of ~22-nucleotide (nt) noncoding RNAs, some of which are known to control the expression of other genes at the posttranscriptional level (Lagos-Quintana et al., 2001; Lau et al., 2001; Lee and Ambros, 2001; Moss and Poethig, 2002). We developed a computational procedure (MiRscan) to identify miRNA genes (Lim et al., 2003) and apply it here to identify most of the miRNA genes in vertebrates. MiRscan relies on the observation that the known miRNAs derive from phylogenetically conserved stem-loop precursor RNAs with characteristic features. MiRscan evaluates conserved stem-loops as miRNA precursors by passing a 21-nt window along each conserved stem-loop, assigning a log-likelihood score to each window that measures how well its attributes resemble those of the first 50 experimentally verified *C. elegans* miRNAs with *C. briggsae* homologues (Lau et al., 2001; Lee and Ambros, 2001; Lim et al., 2003).

Folding of aligned regions of the human and mouse genomes, with subsequent comparison to the pufferfish *Fugu rubripes* genome, identified ~15,000 human genomic segments that fell outside of predicted protein coding genes, were predicted to form stem-loops, and were at least loosely conserved among the three vertebrate species (Supplemental material describing methods and sequences of the predicted miRNA loci and their validation in zebrafish). MiRscan evaluation revealed a high-scoring set of 188 human loci, using a natural cutoff score of 10, defined by a dip in the distribution at this point (Fig. 1). This set included 81 of the 109 members of a reference set of known human miRNA loci, for a sensitivity of 0.74. The fact that a procedure developed and trained solely using nematode miRNAs could also identify most of the vertebrate miRNAs shows that the generic features of the miRNAs and their precursors are conserved broadly among diverse animals, even though the sequences of most miRNAs are not as broadly conserved.

Our analysis can be used to calculate an upper bound on the number of human miRNA genes. If all 188 candidates were authentic miRNA genes and these represented 74% of the total miRNA genes, then there are no more than 255 miRNA genes in the genome. Note that this calculation assumes that rare miRNAs—those expressed at low levels or in a limited set of conditions or cell types, which would be underrepresented in our reference set of cloned miRNAs—will have a distribution of scores and degree of conservation similar to the cloned miRNAs. This assumption is supported by our finding that in nematodes, there is no correlation between the number of times an miRNA was cloned and its MiRscan score (Lim et al., 2003). Furthermore, a tissue such as mouse brain, which might be expected to have miRNAs unique to mammals, is not a particularly rich source of miRNAs without *Fugu* homologues (Lagos-Quintana et al., 2002).

The estimate of 255 human genes is an upper bound implying that no more than 40 miRNA genes remain to be identified in mammals [$\sim 40 = \sim 255 - (109 \text{ known genes} + 107 \text{ new candidates})$]. The estimates for both the gene total and genes remaining to be identified would be lower if some of the 107 newly identified gene candidates were false positives. To evaluate this possibility, we sought to verify these new candidates. Of the 107 new candidates, 14 were close paralogues of loci in the reference set or represented cloned human miRNAs for which loci had not been previously reported. Another 38 were detected in zebrafish cDNA libraries constructed specifically to contain miRNA and siRNA sequences (Supplemental Material). Zebrafish was chosen for this analysis to facilitate examination of a diverse range of tissues and developmental stages. This leaves 55 of the 188 candidates as either false positives or authentic miRNAs expressed at levels too low to be detected. Even if all 55 were false positives, the specificity of our computational procedure would be $133/188 (= 0.71)$, at a score cutoff that identifies 74% of

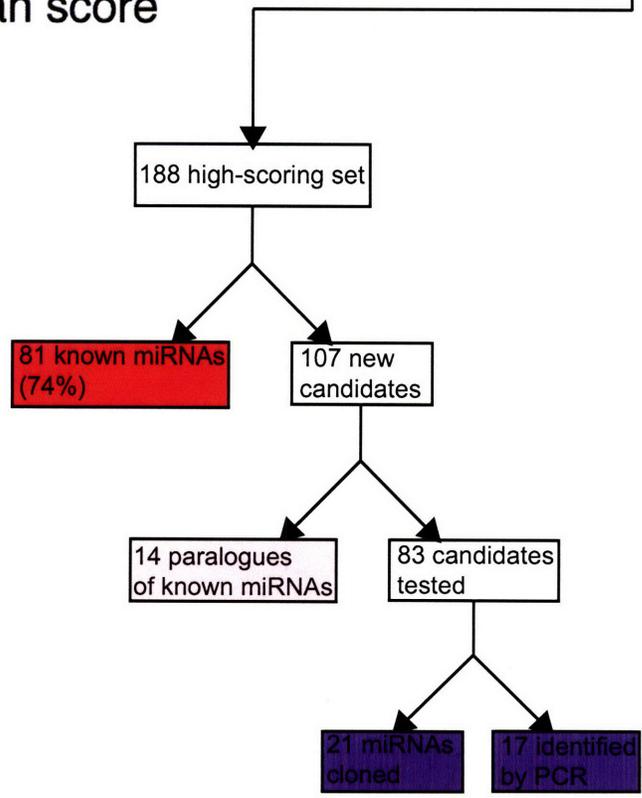
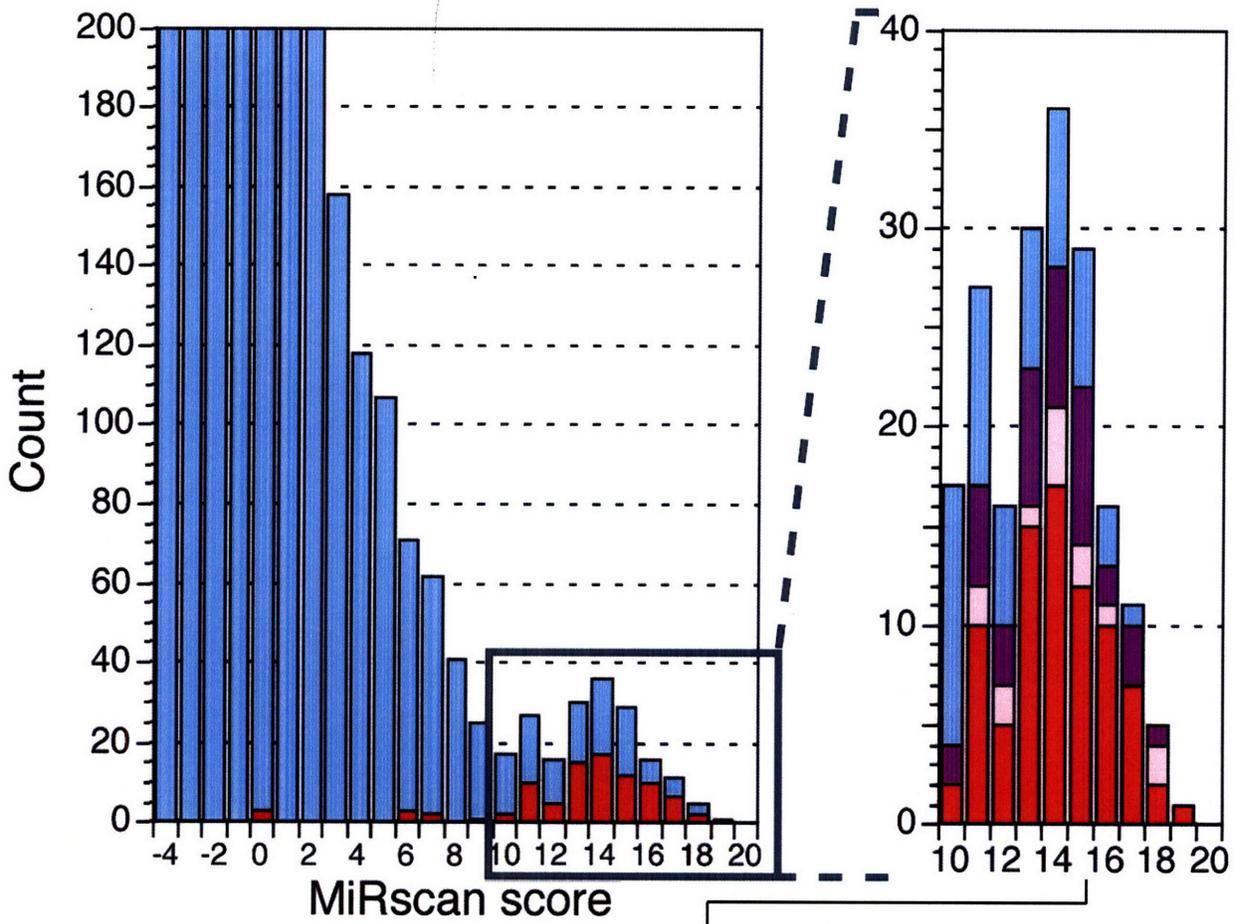
known loci. This minimum specificity value can be used to calculate a lower bound on the number of miRNA genes in mammals as $(188 \times 0.71)/0.74 = 180$. When accounting for the sensitivity of our zebrafish experiments and the incomplete coverage of the genome assemblies used, the lower bound increases to about 200 genes (Supplemental Material).

The 200 to 255 miRNA genes represent nearly 1% of the predicted genes in humans, a fraction similar to that seen for other very large gene families with regulatory roles, such as those encoding transcription-factor proteins. There is no indication that miRNAs are present in single-celled eukaryotes such as yeast. It is tempting to speculate that the substantial expansion of miRNA genes in plants and animals (and the apparent loss of miRNA genes in yeast) is related to their importance in specifying cell differentiation and developmental patterning.

Fig. 1. Computational identification of vertebrate miRNA genes.

The histogram represents the distribution of MiRscan scores for 15,133 human/Fugu consensus structures. Of the 109 reference-set loci, 91 were retained among these aligned segments (red), indicating that at least 80% of the human miRNAs are conserved in fish. The distribution peaks at the score of -4, with a count of 1198, but is truncated at a score of -4 and count of 200 to increase resolution at the high-scoring tail of the distribution. The 188 candidates with scores greater than 10.0 were examined further (expanded portion of the histogram): 81 were in the reference set of known loci (red), 14 were close paralogues of loci in the reference set (≤ 2 point substitutions within the miRNA) or represented cloned human miRNAs for which loci had not been previously reported (pink), and 38 were found in miRNA cDNA libraries made from zebrafish (purple, Supplemental Material).

Figure 1



SUPPLEMENTAL MATERIAL

Materials and Methods

Identification of vertebrate miRNA gene candidates

The computational procedure used to identify vertebrate miRNA genes is summarised in the flowchart of Supplemental Figure 1, with the fraction of known human loci retained at each step of the analysis shown in red. The reference set of known human loci was compiled using 71 previously reported human miRNA loci corresponding to miRNAs cloned from HeLa cells together with 38 human loci corresponding to the closest match to cloned mouse miRNAs with reported loci (Lagos-Quintana et al., 2001; Lagos-Quintana et al., 2002; Mourelatos et al., 2002).

We started with human/mouse BLAT alignments (Kent, 2002) obtained from the UCSC annotation of the December 2001 NCBI human genome assembly with the November 2001 Sanger mouse genome assembly (<http://www.genome.ucsc.edu/downloads.html>). After merging overlapping or closely adjacent blocks of conservation, regions that had over 50% overlap with Ensembl annotated genes (Hubbard et al., 2002) were removed, and the remaining regions were extended by 50 nt on each end. Stem-loops were located by passing a 110nt window through the conserved regions of the human genome (incrementing the position of the window by 3 nt) and folding the window with the program RNAfold (Hofacker et al., 1994) to identify predicted stem-loop structures with a minimum of 25 base pairs and a folding free energy of at least 25 kcal/mol ($\Delta G^{\circ}_{\text{folding}} \leq -25$ kcal/mol). Stem-loops that had fewer base pairs than overlapping stem-loops were culled, leaving ~800,000 human stem-loops. For each human stem-loop, the corresponding mouse sequence was retrieved using WUBLAST (Gish, W. <http://blast.wustl.edu>) and folded with RNAfold. Following sequence alignment with ClustalW (Thompson et al., 1994), a consensus human/mouse structure was generated using Alidot (Hofacker et al., 1994). The Alidot consensus structures were scored with MiRscan, using scoring matrices derived from 50 miRNA loci conserved between *C. elegans* and *C. briggsae* (Lim et al., 2003). MiRscan scores the following seven features of the miRNA and its predicted precursor: i) base pairing of the miRNA within the foldback, ii) base pairing of the rest of the foldback, iii) stringent sequence conservation in the 5' half of the microRNA, iv) slightly less stringent sequence conservation in the 3' half of the miRNA, v) sequence biases in the first five bases of the miRNA (especially a U at the first position), vi) a tendency towards having symmetric rather than asymmetric internal loops and bulges in the miRNA region, and vii) the presence of 2–9 consensus base pairs between the miRNA and the terminal loop region, with a preference for 4–6 bp (Lim et al., 2003). 15,651 human sequences scoring in the top 10% of the human/mouse analysis had at least weak homology to the assembled *Fugu rubripes* genome (Aparicio et al., 2002), as determined by WU-BLAST. 518 sequences corresponding to repetitive Fugu loci (i.e. with more than 20 WU-BLAST hits to the Fugu genome) were culled. (These included seven sequences with scores above 10.0; none had scores above 14.0.) For the remaining human/Fugu pairs, consensus structures were created and scored with MiRscan, again using the scoring matrix trained on nematode miRNA loci. The 188 candidates with scores greater than 10.0 are listed, linked to their Human/Fugu consensus secondary structures (Table S1). Note that as has been customary in miRNA gene identification, we did not determine whether any of the duplicated loci are pseudogenes. The 188 loci with MiRscan scores greater than 10.0 included 81 of the 109 reference set loci. We anticipate that using more complete genome databases and

more sensitive alignments of syntenic regions would enable an even higher proportion of known genes to be captured. Note that at each locus there are two possible stem-loop precursors; in addition to the correct precursor, there is the reverse complement of the correct precursor, which generally will also be predicted to form a stem-loop. For 17 of the 81 predicted loci that matched the reference set, the candidate 21-nt miRNA was on the reverse-complement precursor (see secondary structures linked to Table S1). Of the 64 remaining candidates that resided on the correct precursor, 14 were on the opposite arm of the stem-loop. Thus, 50 of the 81 candidate ~21-nt miRNAs that matched the reference set loci, were on the correct arm of the correct precursor (see secondary structures linked to Table S1). About half (26) of these were correctly positioned at their 5' terminus, and nearly all were within 3 nucleotides of the correct position (see secondary structures linked to Table S1).

Validation of miRNA gene candidates

Of the 188 high-scoring candidates, 93 remained after excluding the 81 candidates that matched the reference set loci and another 14 candidates that were close paralogues of the reference set loci or represented cloned human miRNAs for which loci had not been previously reported. We examined whether any of these 93 newly identified candidates could be found in zebrafish miRNA cDNA libraries. The libraries were generated from total RNA isolated from zebrafish embryos using the protocol developed for *C. elegans* (Lau et al., 2001), which is posted on the web (<http://web.wi.mit.edu/bartel/pub/>). They were made from the following developmental stages: zygote, 256 cell, germ ring, 75% epiboly, tailbud, 11 somite, 24 hours, and 48 hours. 260-1000 clones were sequenced from each library. Of the 93 newly identified miRNA gene candidates, 21 matched sequenced clones (Table S1). A PCR assay was used to test for candidate sequences present within the zebrafish libraries but absent from the set of sequenced clones (Lim et al., 2003). Using unpublished releases of zebrafish genomic sequence, (available from www.sanger.ac.uk/projects/D_reno), primers were designed for candidates with apparent zebrafish homologues (Table S2). Of the 58 candidates tested by this method, 17 were validated (Table S1). For each predicted miRNA locus, the mature miRNA might reside in four alternative locations (the location of the top scoring candidate, the other arm of the stem-loop, or on either side of the reverse-complement stem-loop). MiRscan scores for all four alternatives were generated, and the primers were designed to test the two highest scoring alternatives (Table S2). For 80% of the genes in the reference set, the correct alternative had the highest or second-highest score.

Additional considerations for estimates of gene numbers

Several considerations are each expected to cause small decreases in the sensitivity of detection of miRNAs at various steps in our analysis, the net effect of which leads us to increase our lower bound on the number of miRNA genes in humans from 180 to ~200. These considerations are as follows:

- 1) use of two rather than four specific primers in the PCR assay;
- 2) incomplete coverage of the zebrafish genome by the genome assembly used;
- 3) incomplete coverage of the mouse and human genomes by the genome

assemblies used for the BLAT alignments;

A more detailed description of these considerations is given below.

PCR assay. Some of the loci that were not validated in the PCR assay were probably false negatives. For instance, the PCR assay was used to test for expression from only two of the four possible miRNA locations within each locus. We anticipate that extending the PCR assay to examine all four alternatives rather than the two highest scoring alternatives would validate several additional predictions. Furthermore, positive controls performed in parallel showed that the sensitivity of this assay is about 0.85 even when the correct alternative is used.

Zebrafish genomic sequence. A second consideration is that not all the predictions could be tested using the PCR assay because zebrafish homologues were not found within the available zebrafish genomic sequence (Table S1). As additional zebrafish sequence becomes available, more candidates can be tested and verified. In addition, closer homologues for some previously tested candidates can be identified and tested.

Mouse and human genomic sequence. The third consideration also relates to incomplete genomic databases. Our reference set of known loci was compiled from previously reported cloned miRNAs that matched human and mouse genomic data available in 2001 (Lagos-Quintana et al., 2001; Lagos-Quintana et al., 2002; Mourelatos et al., 2002). Our computational search was based on BLAT alignment of assemblies also from 2001. Because loci present only in more recent assemblies will be absent from both our reference set and our predictions, the estimate of the number of miRNA genes will be low, in proportion to the degree of coverage of the 2001 genomic assemblies.

Together, these three considerations increase our estimate for the lower bound on the number of human miRNA genes from 180 to about 200. The third consideration, stemming from incomplete coverage of the 2001 assemblies, would also raise the upper estimate for the number of genes in humans. On the other hand, among the 188 candidates, there was a clear correlation between the MiRscan score of the candidate and whether or not it was validated (Table S1), which indicates that some of the low-scoring candidates were false positives. We estimate that these false positives offset the false negatives arising from incomplete database coverage, and thus do not increase the upper bound for the number of human miRNA genes beyond 255.

ACKNOWLEDGEMENTS

We thank Elizabeth Wiellette and Hazel Sive for guidance in culturing and staging zebrafish embryos, and Julian Lange and David Page for assistance and use of equipment and facilities. This analysis would not have been possible without the availability of genome sequences from human, mouse, Fugu and zebrafish, and we thank the contributing sequencing centres. We also thank Sam Griffiths-Jones and the miRNA Registry (www.sanger.ac.uk/Software/Rfam/mirna/) for assisting with the naming of newly identified miRNA genes and Mathew Jones-Rhoades for helpful comments on the data.

Supplemental Table S1. 188 predicted human miRNA loci.

For each locus, the coordinates are those of the December 2001 human genome assembly (genome.ucsc.edu). Newly identified genes were validated either by sequencing clones from a cDNA library of zebrafish 18- to 26-nt RNAs, or by detecting the miRNA sequences in this library using a PCR-based method (Lim et al., 2003). Because our PCR method determines the 5' terminus of the miRNA but not its 3' terminus, some newly identified miRNAs that were verified using this method are shown as 21mers, with footnotes indicating that their 3' termini were not defined. The remaining candidates verified by the PCR method (genes for miR-182, miR-183, miR-187, and miR-192) were concurrently identified by the cloning and sequencing effort of the Tuschl group (Lagos-Quintana et al., 2003); in these cases, the 3' termini are those of the sequences reported in reference (Lagos-Quintana et al., 2003). For miR-7 and miR-183, both the miRNA and the sequence from the other side of the precursor (the miRNA* sequence) were detected using the PCR method; in these cases, the miRNA was distinguished from the miRNA* based on the orthologous miRNAs cloned from *Drosophila* and *C. elegans* (Lagos-Quintana et al., 2001; Lim et al., 2003). For miR-182 the PCR method appears to have only detected the miRNA* (Lagos-Quintana et al., 2003).

(The annotated stem loop of miR-140 is depicted as an example. Table with links to other annotated stem-loops can be downloaded from *Science* Online <http://www.sciencemag.org/cgi/content/full/299/5612/1540/DC1>)

Human locus	Score	Human gene	Human miRNA sequence	Validation method (zebrafish miRNA)
CHR 16:[72705524,72705633]	19.25	<i>mir-140</i>	AGUGGUUUUACCCUAUGGUAG	Reference set
CHR 6:[76441142,76441033]	18.61	<i>mir-30a</i>	CUUUCAGUCGGAUGUUUGCAGC	Reference set
CHR 1:[8224808,8224917]	18.58	<i>mir-34</i>	UGGCAGUGUCUUAGCUGGUUGU	Newly identified, verified by cloning (UGGCAGUGUCUUAGCUGGUUGU)
CHR 1:[218380691,218380800]	18.49	<i>mir-29b-1</i>	UAGCACCAUUUGAAAUCAGUGUU	Reference set
CHR 7:[140153438,140153329]	18.46	<i>mir-29b-2</i>	UAGCACCAUUUGAAAUCAGUGUU	Homolog of published miRNA
CHR 3:[166222322,166222431]	18.04	<i>mir-16-3</i>	UAGCAGCACGUAAAUAUUGGCG	Homolog of published miRNA
CHR 14:[104094902,104095011]	17.85	<i>mir-203</i>	GUGAAAUGUUUAGGACCACUAG	Newly identified, verified by cloning (GUGAAAUGUUUAGGACCACUAG)
CHR 9:[76243263,76243154]	17.76	<i>mir-7-1</i>	UGGAAGACUAGUGAUUUUGUU ²	Newly identified, verified by PCR ¹ (UGGAAGACUAGUGAUUUUGUU)
CHR 2:[175251688,175251797]	17.75	<i>mir-10b</i>	UACCCUGUAGAACCGAAUUUGU	Newly identified, verified by cloning (UACCCUGUAGAACCGAAUUUGU)
CHR 2:[132584666,132584775]	17.65	<i>mir-128a</i>	UCACAGUGAACCGGUCUCUUUU	Reference set
CHR 2:[218671450,218671559]	17.56	<i>mir-153-1</i>	UUGCAUAGUCACAAAAGUGA	Reference set
CHR 7:[161989454,161989345]	17.44	<i>mir-153-2</i>	UUGCAUAGUCACAAAAGUGA	Reference set
CHR 15:[59496766,59496875]	17.43			Not detected by PCR
CHR 9:[89781464,89781573]	17.41	<i>mir-27b</i>	UUCACAGUGGCUAAGUUCUG	Reference set
CHR 7:[133435157,133435048]	17.36	<i>mir-96</i>	UUUGGCACUAGCACAUUUUUGC	Reference set
CHR 13:[92040427,92040536]	17.34	<i>mir-17a/mir-91</i>	CAAAGUGCUUACAGUGCAGGUAGU	Reference set
CHR 9:[131994694,131994585]	17.17	<i>mir-123/mir-126a</i>	CAUUAUUACUUUUGGUACGCG	Reference set
CHR 17:[1722038,1721929]	16.91	<i>mir-132</i>	U AACAGUCUACAGCCAUGGUCGC	Reference set
CHR 17:[30570536,30570427]	16.84	<i>mir-108-1</i>	AUAAGGAUUUUUAGGGGCAUU	Homolog of published miRNA
CHR 9:[89781221,89781330]	16.67	<i>mir-23b</i>	AUCACAUUGCCAGGGAUUACCAC	Reference set
CHR 12:[65375803,65375912]	16.63	<i>let-7i</i>	UGAGGUAGUAGUUUGUGCU	Reference set
CHR 17:[1722404,1722295]	16.62	<i>mir-212</i>	U AACAGUCUCCAGUCACGGCC ²	Newly identified, verified by PCR (U AACAGUCUACAGUCAUGGCU)
CHR 1:[90399302,90399411]	16.62			Not detected by PCR
CHR 5:[89092977,89092868]	16.62	<i>mir-131-2</i>	U AAAGCUAGAU AACCGAAAGU	Reference set
CHR 22:[43175818,43175927]	16.62	<i>let-7b</i>	UGAGGUAGUAGGUUGUGUGGUU	Reference set
CHR 20:[61054858,61054749]	16.58	<i>mir-1d</i>	UGGAAUGUAAAAGAAGUAUGUAU	Reference set
CHR 18:[59559211,59559320]	16.54	<i>mir-122a</i>	UGGAGUGUGACAAUGGUGUUUGU	Reference set
CHR 17:[1417960,1417851]	16.47	<i>mir-22</i>	AAGCUGCCAGUUGAAGAACUGU	Reference set
CHR 13:[92041120,92041229]	16.24	<i>mir-92-1</i>	UAUUGCACUUGUCCCGGCCUGU	Reference set
CHR 7:[94811366,94811257]	16.23			Not tested, zebrafish homology not yet found
CHR 17:[60258675,60258566]	16.12	<i>mir-142</i>	CAUAAAGUAGAAAGCACUAC	Reference set
CHR 7:[133435388,133435279]	16.10	<i>mir-183</i>	UAUGGCACUGGUAGAAUUCACUG	Newly identified, verified by PCR ¹ (UAUGGCACUGGUAGAAUUCACUG)
CHR X:[119194533,119194642]	16.04			Not detected by PCR
CHR 1:[173396907,173396798]	15.92	<i>mir-214</i>	ACAGCAGGCACAGACAGGCAG	Newly identified, verified by cloning (ACAGCAGGCACAGACAGGCAG)
CHR 5:[151612723,151612832]	15.79	<i>mir-143</i>	UGAGAUGAAGCACUGUAGCUCA	Reference set
CHR 11:[58914759,58914650]	15.78	<i>mir-192-1</i>	CUGACCUAUGAAUUGACAGCC	Newly identified, verified by PCR (AUGACCUAUGAAUUGACAGCC)
CHR 11:[59435564,59435455]	15.78	<i>mir-192-2</i>	CUGACCUAUGAAUUGACAGCC	Newly identified, verified by PCR (AUGACCUAUGAAUUGACAGCC)
CHR 11:[59544391,59544282]	15.78	<i>mir-192-3</i>	CUGACCUAUGAAUUGACAGCC	Newly identified, verified by PCR (AUGACCUAUGAAUUGACAGCC)
CHR 20:[33400991,33401100]	15.77			Not detected by PCR

<u>CHR 5:[90467504,90467613]</u>	15.70			Not detected by PCR
<u>CHR 5:[3220067,3219958]</u>	15.62			Not detected by PCR
<u>CHR 22:[43174880,43174989]</u>	15.61	<i>let-7a-3</i>	UGAGGUAGUAGGUUGUAUAGUU	Reference set
<u>CHR 1:[93614799,93614690]</u>	15.59			Not detected by PCR
<u>CHR 13:[96029172,96029281]</u>	15.58			Not detected by PCR
<u>CHR 9:[118855135,118855244]</u>	15.57	<i>mir-181a</i>	AACAUUCAACGCUGUCGGUGAGU	Newly identified, verified by cloning (AACAUUCAACGCUGUCGGUGAGU)
<u>CHR 9:[87053162,87053271]</u>	15.55	<i>let-7a-1</i>	UGAGGUAGUAGGUUGUAUAGUU	Reference set
<u>CHR 1:[219827236,219827345]</u>	15.54	<i>mir-205</i>	UCCUUCAUUCCACCGGAGUCUG	Newly identified, verified by cloning (UCCUUCAUUCCACCGGAGUCUG)
<u>CHR 5:[160729220,160729111]</u>	15.48	<i>mir-103-1</i>	AGCAGCAUUGUACAGGGCUAUGA	Reference set
<u>CHR 3:[40955342,40955451]</u>	15.46	<i>mir-26a</i>	UUCAAGUAAUCCAGGAUAGGCU	Reference set
<u>CHR 22:[39011454,39011563]</u>	15.41	<i>mir-33a</i>	GUGCAUUGUAGUUGCAUUG	Reference set
<u>CHR 12:[56846907,56846798]</u>	15.38	<i>mir-196-2</i>	UAGGUAGUUUCAUGUUGUUGGG	Newly identified, verified by cloning (UAGGUAGUUUCAUGUUGUUGGG)
<u>CHR 10:[95470599,95470490]</u>	15.35	<i>mir-107</i>	AGCAGCAUUGUACAGGGCUAUCA	Homolog of published miRNA
<u>CHR X:[129317473,129317364]</u>	15.33	<i>mir-106</i>	AAAAGUGCUUACAGUGCAGGUAGC	Homolog of published miRNA
<u>CHR 9:[87053556,87053665]</u>	15.30	<i>let-7f-1</i>	UGAGGUAGUAGAUUGUAUAGUU	Reference set
<u>CHR 5:[81367589,81367698]</u>	15.27			Not detected by PCR
<u>CHR 1:[218381280,218381389]</u>	15.27	<i>mir-29c</i>	CUAGCACCAUUUGAAAUCGGUU	Reference set
<u>CHR 11:[66445182,66445291]</u>	15.24	<i>mir-130a</i>	CAGUGCAAUGUUAAAAGGGC	Reference set
<u>CHR 2:[174973719,174973828]</u>	15.20			Not detected by PCR
<u>CHR 4:[21933011,21933120]</u>	15.19	<i>mir-218-1</i>	UUGUGCUUGAUCUAACCAUGU ²	Newly identified, verified by PCR (UUGUGCUUGAUCUAACCAUGU)
<u>CHR 8:[62809258,62809367]</u>	15.16	<i>mir-124a-2</i>	UUAAGGCACGCGGUGAAUGCCA	Reference set
<u>CHR 17:[61936443,61936552]</u>	15.06	<i>mir-21</i>	UAGCUUAUCAGACUGAUGUUGA	Reference set
<u>CHR 13:[50013057,50012948]</u>	15.05	<i>mir-16-1</i>	UAGCAGCACGUAAAUAUUGGCG	Reference set
<u>CHR 17:[28101804,28101695]</u>	15.00	<i>mir-144</i>	UACAGUAUAGAUGAUGUACUAG	Reference set
<u>CHR X:[42918827,42918718]</u>	14.99	<i>mir-221</i>	AGCUACAUUGUCUGCUGGGUUUC	Newly identified, verified by cloning (AGCUACAUUGUCUGCUGGGUUUC)
<u>CHR X:[42919663,42919554]</u>	14.94	<i>mir-222</i>	AGCUACAUCUGGCUACUGGGUCUC	Newly identified, verified by cloning (AGCUACAUCUGGCUACUGGGUCUC)
<u>CHR 8:[136408632,136408523]</u>	14.93	<i>mir-30d</i>	UGUAAACAUCGCCGACUGGAAG	Reference set
<u>CHR X:[129316945,129316836]</u>	14.93	<i>mir-19b-2</i>	UGUGCAAUCCAUGCAAAACUGA	Reference set
<u>CHR 3:[38612498,38612607]</u>	14.88	<i>mir-128b</i>	UCACAGUGAACCGGUCUCUUUC	Homolog of published miRNA
<u>CHR 9:[122569704,122569595]</u>	14.85			Not detected by PCR
<u>CHR 2:[58499669,58499560]</u>	14.82			Not tested, apparent homolog of mir-23
<u>CHR 21:[23525088,23525197]</u>	14.81			Not tested, zebrafish homology not yet found
<u>CHR 7:[134473761,134473652]</u>	14.78	<i>mir-29b-3</i>	UAGCACCAUUUGAAAUCAGUGUU	Homolog of published miRNA
<u>CHR 11:[44269678,44269787]</u>	14.73	<i>mir-129-2</i>	CUUUUUGCGGUCUGGGCUUGC	Homolog of published miRNA
<u>CHR 6:[55307706,55307815]</u>	14.71	<i>mir-133b</i>	UUGGUCCCCUUAACCAGCUA	Homolog of published miRNA
<u>CHR 3:[21794989,21794880]</u>	14.69			Not detected by PCR
<u>CHR 9:[87056039,87056148]</u>	14.66	<i>let-7d</i>	AGAGGUAGUAGGUUGCAUAGU	Reference set
<u>CHR 3:[166222169,166222278]</u>	14.66	<i>mir-15b</i>	UAGCAGCACAUCAUGGUUUACA	Reference set
<u>CHR 7:[134473051,134472942]</u>	14.62	<i>mir-29a-1</i>	CUAGCACCAUCUGAAAUCGGUU	Reference set
<u>CHR 1:[154832185,154832076]</u>	14.60			Not detected by PCR
<u>CHR 9:[122421823,122421714]</u>	14.58	<i>mir-199b</i>	CCCAGUGUUUAGACUAUCUGUUC	Newly identified, verified by cloning (CCCAGUGUUCAGACUACCUGUUC)
<u>CHR 7:[131134356,131134465]</u>	14.56	<i>mir-129-1</i>	CUUUUUGCGGUCUGGGCUUGC	Reference set
<u>CHR 19:[69303433,69303542]</u>	14.53	<i>let-7e</i>	UGAGGUAGGAGGUUGUAUAGU	Reference set
<u>CHR 13:[72133104,72132995]</u>	14.52			Not detected by PCR
<u>CHR 21:[14576290,14576399]</u>	14.50	<i>let-7c</i>	UGAGGUAGUAGGUUGUAUGGUU	Reference set
<u>CHR 9:[62636736,62636627]</u>	14.49	<i>mir-204</i>	UUCCCUUGUCAUCCUAUGCCU	Newly identified, verified by cloning

				(UCCCCUUGUCAUCCUAUGCCU)
<u>CHR 5:[151614445,151614554]</u>	14.40	<i>mir-145</i>	GUCCAGUUUCCAGGAAUCCUU	Reference set
<u>CHR 8:[9693506,9693615]</u>	14.28	<i>mir-124a-1</i>	UUAAGGCACGCGGUGAAUGCCA	Reference set
<u>CHR 22:[17113556,17113665]</u>	14.24			Not detected by PCR
<u>CHR 1:[202905852,202905743]</u>	14.22	<i>mir-213</i>	ACCAUCGACCGUUGAUUGUACC	Newly identified, verified by cloning (ACCAUCGACCGUUGAUUGUACC)
<u>CHR 4:[158780085,158779976]</u>	14.22			Not detected by PCR
<u>CHR 13:[92040868,92040977]</u>	14.21	<i>mir-20</i>	UAAAGUGCUUAUAGUGCAGGUAG	Reference set
<u>CHR 18:[20292169,20292060]</u>	14.18	<i>mir-133a-1</i>	UUGGUCCCCUUAACCAGCUGU	Reference set
<u>CHR 16:[59052655,59052764]</u>	14.18	<i>mir-138-2</i>	AGCUGGUGUUGUGAAUC	Reference set
<u>CHR X:[49457711,49457602]</u>	14.13	<i>mir-98</i>	UGAGGUAGUAAGUUGUAUUGUU	Reference set
<u>CHR 17:[49334701,49334592]</u>	14.11	<i>mir-196-1</i>	UAGGUAGUUUCAUGUUGUUGGG	Newly identified, verified by cloning (UAGGUAGUUUCAUGUUGUUGGG)
<u>CHR 11:[127322200,127322091]</u>	14.07	<i>mir-125b-1</i>	UCCUGAGACCCUAAUUGUGA	Reference set
<u>CHR 1:[173402644,173402535]</u>	14.07	<i>mir-199a-2</i>	CCCAGUGUUCAGACUACCUGUUC	Newly identified, verified by cloning (CCCAGUGUUCAGACUACCUGUUC)
<u>CHR 7:[140152718,140152609]</u>	14.05	<i>mir-29a-2</i>	CUAGCACCAUCUGAAAUCGGUU	Reference set
<u>CHR 4:[38938235,38938344]</u>	14.00			Not tested, zebrafish homology not yet found
<u>CHR 1:[202905681,202905572]</u>	13.99	<i>mir-181b</i>	AACAUUCAUUGCUGUCGGUGGGUU	Newly identified, verified by cloning (AACAUUCAUUGCUGUCGGUGGGUU)
<u>CHR 15:[30996132,30996023]</u>	13.96			Not detected by PCR
<u>CHR 12:[6964499,6964608]</u>	13.93	<i>mir-141</i>	AACACUGUCUGGUAAGAUGG	Reference set
<u>CHR 1:[157065254,157065363]</u>	13.92	<i>mir-131-1</i>	UAAAGCUAGUAACCGAAAGU	Reference set
<u>CHR 20:[61044240,61044131]</u>	13.87	<i>mir-133a-2</i>	UUGGUCCCCUUAACCAGCUGU	Reference set
<u>CHR 10:[141581502,141581611]</u>	13.87			Not tested, zebrafish homology not yet found
<u>CHR 10:[141772361,141772470]</u>	13.87			Not tested, zebrafish homology not yet found
<u>CHR 10:[80541846,80541737]</u>	13.86			Not detected by PCR
<u>CHR 18:[20295481,20295372]</u>	13.82	<i>mir-1b</i>	UGGAAUGUAAAGAAGUAUGUAU	Reference set
<u>CHR 13:[92040559,92040668]</u>	13.70	<i>mir-18</i>	UAAGGUGCAUCUAGUGCAGUA	Reference set
<u>CHR X:[119291657,119291548]</u>	13.62	<i>mir-220</i>	CCACACCGUAUCUGACACUUU ²	Newly identified, verified by PCR (CCACAACCGUAUCGGACACUU)
<u>CHR 1:[214725765,214725656]</u>	13.60			Not detected by PCR
<u>CHR 19:[6814521,6814630]</u>	13.58	<i>mir-7-3</i>	UGGAAGACUAGUGAUUUUGUU ²	Newly identified, verified by PCR ¹ (UGGAAGACUAGUGAUUUUGUU)
<u>CHR 5:[161016881,161016772]</u>	13.58	<i>mir-218-2</i>	UUGUGCUUGAUCUAACCAUGU ²	Newly identified, verified by PCR (UUGUGCUUGAUCUAACCAUGU)
<u>CHR 19:[16613842,16613733]</u>	13.49	<i>mir-24-2</i>	UGGCUCAGUUCAGCAGGAACAG	Reference set
<u>CHR 9:[89782024,89782133]</u>	13.49	<i>mir-24-1</i>	UGGCUCAGUUCAGCAGGAACAG	Reference set
<u>CHR 20:[3853130,3853239]</u>	13.48	<i>mir-103-2</i>	AGCAGCAUUGUACAGGGCUAUGA	Reference set
<u>CHR 15:[27942568,27942459]</u>	13.45	<i>mir-211</i>	UCCCCUUGUCAUCCUUCGCCU	Newly identified, verified by cloning (UCCCCUUGUCAUCCUAUGCCU)
<u>CHR 9:[1526810,1526919]</u>	13.43	<i>mir-101-3</i>	UACAGUACUGUGUAACUGA	Reference set
<u>CHR 8:[136404276,136404167]</u>	13.40	<i>mir-30b</i>	UGUAAACAUCCUACACUCAGC	Reference set
<u>CHR 13:[90861308,90861199]</u>	13.37			Not detected by PCR
<u>CHR 11:[127204641,127204532]</u>	13.37	<i>let-7a-4</i>	UGAGGUAGUAGGUUGUAUAGUU	Reference set
<u>CHR 17:[49282057,49281948]</u>	13.36	<i>mir-10a</i>	UACCCUGUAGAUCCGAAUUUGUG	Newly identified, verified by cloning (UACCCUGUAGAUCCGAAUUUGUG)
<u>CHR 13:[92040712,92040821]</u>	13.29	<i>mir-19a</i>	UGUGCAAUUCUAUGCAAACUGA	Reference set
<u>CHR X:[49458662,49458553]</u>	13.17	<i>let-7f-2</i>	UGAGGUAGUAGAUUGUAUAGUU	Reference set
<u>CHR 13:[50013207,50013098]</u>	13.15	<i>mir-15a-1</i>	UAGCAGCACAAUAAUGGUUGUG	Reference set

<u>CHR_16:[12853371,12853262]</u>	13.08	<i>mir-108-2</i>	AUAAGGAUUUUUAGGGGCAUU	Homolog of published miRNA
<u>CHR_1:[103432111,103432220]</u>	13.01	<i>mir-137</i>	UAUUGCUUAAGAAUACGCGUAG	Reference set
<u>CHR_6:[43048796,43048905]</u>	13.01	<i>mir-219</i>	UGAUUGUCCAAACGCAAUUCU ²	Newly identified, verified by PCR (UGAUUGUCCAAACGCAAUUCU)
<u>CHR_12:[57461725,57461834]</u>	12.87	<i>mir-148b</i>	UCAGUGCAUCACAGAACUUUGU	Homolog of published miRNA
<u>CHR_22:[18806271,18806380]</u>	12.87	<i>mir-130b</i>	CAGUGCAAUGAUGAAAGGGC	Homolog of published miRNA
<u>CHR_13:[92041003,92041112]</u>	12.85	<i>mir-19b-1</i>	UGUGCAAUCCAUGCAAACUGA	Reference set
<u>CHR_11:[127368959,127368850]</u>	12.84	<i>let-7a-2</i>	UGAGGUAGUAGGUUGUAUAGUU	Reference set
<u>CHR_2:[58511309,58511418]</u>	12.59	<i>mir-216</i>	UAAUCUCAGCUGGCAACUGUG ²	Newly identified, verified by PCR (UAAUCUCAGCUGGCAACUGUG)
<u>CHR_11:[127374676,127374567]</u>	12.51	<i>mir-100-1</i>	AACCCGUAGAUCGGAACUUUGU	Reference set
<u>CHR_11:[127210360,127210251]</u>	12.51	<i>mir-100-2</i>	AACCCGUAGAUCGGAACUUUGU	Reference set
<u>CHR_18:[34497140,34497031]</u>	12.50	<i>mir-187</i>	UCGUGUCUUGUGUUGCAGCCGG	Newly identified, verified by PCR (UCGUGUCUUGUGUUGCAGCCA)
<u>CHR_20:[22339705,22339814]</u>	12.47			Not detected by PCR
<u>CHR_10:[124638355,124638464]</u>	12.44			Not detected by PCR
<u>CHR_10:[127681756,127681865]</u>	12.44			Not detected by PCR
<u>CHR_20:[61811381,61811490]</u>	12.34	<i>mir-124a-3</i>	UUAAGGCACGCGGUGAAUGCCA	Reference set
<u>CHR_15:[85832771,85832880]</u>	12.30	<i>mir-7-2</i>	UGGAAGACUAGUGAUUUUGUU ²	Newly identified, verified by PCR (UGGAAGACUAGUGAUUUUGUU)
<u>CHR_19:[24459027,24459136]</u>	12.20			Not tested, zebrafish homology not yet found
<u>CHR_7:[16645672,16645781]</u>	12.09			Not detected by PCR
<u>CHR_17:[51122450,51122559]</u>	12.05			Not detected by PCR
<u>CHR_11:[895403,895294]</u>	11.98	<i>mir-210</i>	CUGUGCGUGUGACAGCGGUG	Newly identified, verified by cloning (CUGUGCGUGUGACAGCGGCUA)
<u>CHR_1:[231366714,231366605]</u>	11.97	<i>mir-215</i>	AUGACCUAUGAAUUGACAGAC ²	Newly identified, verified by PCR (AUGACCUAUGAAUUGACAGCC)
<u>CHR_X:[60718647,60718756]</u>	11.92	<i>mir-223</i>	UGUCAGUUUGUCAAUACCCC ²	Newly identified, verified by PCR (UGUCAGUUUGUCAAUACCCC)
<u>CHR_15:[86699903,86700012]</u>	11.91	<i>mir-131-3</i>	UAAAGCUAGAUAAACCGAAAGU	Reference set
<u>CHR_19:[14233893,14233784]</u>	11.85	<i>mir-199a-1</i>	CCCAGUGUUCAGACUACCUGUUC	Newly identified, verified by cloning (CCCAGUGUUCAGACUACCUGUUC)
<u>CHR_6:[76414552,76414443]</u>	11.83	<i>mir-30c</i>	UGUAAACAUCCUACACUCUCAGC	Reference set
<u>CHR_1:[47442083,47442192]</u>	11.79	<i>mir-101-1</i>	UACAGUACUGUGAUAAACUGA	Reference set
<u>CHR_1:[47837341,47837450]</u>	11.79	<i>mir-101-2</i>	UACAGUACUGUGAUAAACUGA	Reference set
<u>CHR_2:[170125323,170125214]</u>	11.79			Not detected by PCR
<u>CHR_15:[38883681,38883790]</u>	11.77			Not tested, zebrafish homology not yet found
<u>CHR_2:[218109200,218109309]</u>	11.65	<i>mir-26b</i>	UUCAAGUAAUUCAGGAUAGGUU	Reference set
<u>CHR_5:[140967392,140967283]</u>	11.59			Not tested, zebrafish homology not yet found
<u>CHR_17:[48643265,48643156]</u>	11.57	<i>mir-152</i>	UCAGUGCAUGACAGAACUUGG	Reference set
<u>CHR_3:[55016281,55016172]</u>	11.53	<i>mir-135-1</i>	UAUGGCUUUUUUUAUCCUAUGUGAU	Reference set
<u>CHR_12:[100165134,100165243]</u>	11.48	<i>mir-135-2</i>	UAUGGCUUUUUUUAUCCUAUGUGAU	Homolog of published miRNA
<u>CHR_2:[58554598,58554707]</u>	11.43	<i>mir-217</i>	UACUGCAUCAGGAACUGAUUGGAU	Newly identified, verified by cloning (UACUGCAUCAGGAACUGAUUGGAU)
<u>CHR_1:[90399570,90399461]</u>	11.39			Not detected by PCR
<u>CHR_13:[50448400,50448291]</u>	11.37	<i>mir-15a-2</i>	UAGCAGCACAAUUGGUUUUGU	Reference set
<u>CHR_3:[54990340,54990231]</u>	11.37	<i>let-7g</i>	UGAGGUAGUAGUUUGUACAGU	Reference set
<u>CHR_17:[51087113,51087222]</u>	11.33			Not tested, zebrafish homology not yet found
<u>CHR_17:[23384662,23384771]</u>	11.32	<i>mir-33b</i>	GUGCAUUGCUGUUGCAUUG	Homolog of published miRNA
<u>CHR_X:[120434504,120434613]</u>	11.24			Not tested, zebrafish homology not yet found

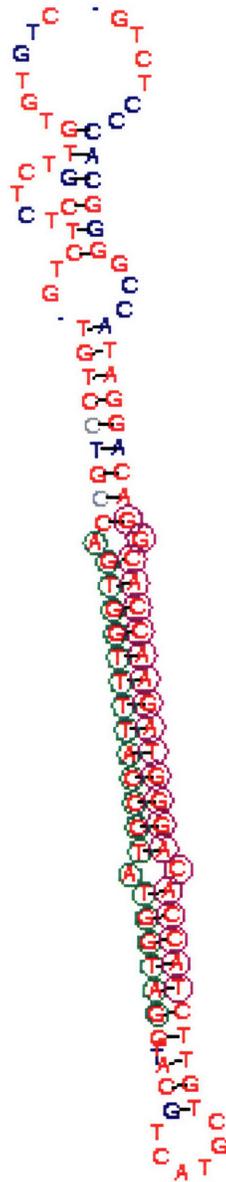
<u>CHR_13:[50448235,50448126]</u>	11.21	<i>mir-16-2</i>	UAGCAGCACGUAAAUAUUGGCG	Reference set
<u>CHR_22:[39875393,39875284]</u>	11.06			Not detected by PCR
<u>CHR_7:[127373003,127373112]</u>	11.04			Not tested, zebrafish homology not yet found
<u>CHR_2:[155447946,155447837]</u>	11.03			Not detected by PCR
<u>CHR_4:[158780484,158780375]</u>	11.01			Not detected by PCR
<u>CHR_10:[82216993,82216884]</u>	10.99			Not detected by PCR
<u>CHR_9:[5165902,5166011]</u>	10.96			Not detected by PCR
<u>CHR_7:[133430866,133430757]</u>	10.95	<i>mir-182</i>	UUUGCAAUGGUAGAACUCACA	Newly identified, verified by PCR (miR-182*,UGGUUCUAGACUUGCCAACUA)
<u>CHR_4:[154056517,154056408]</u>	10.90			Not tested, zebrafish homology not yet found
<u>CHR_18:[75755597,75755706]</u>	10.82			Not detected by PCR
<u>CHR_5:[40367092,40367201]</u>	10.68			Not detected by PCR
<u>CHR_2:[173797566,173797457]</u>	10.57			Not tested, zebrafish homology not yet found
<u>CHR_15:[58026411,58026520]</u>	10.45			Not detected by PCR
<u>CHR_15:[26169137,26169028]</u>	10.42			Not detected by PCR
<u>CHR_11:[136385422,136385531]</u>	10.42			Not detected by PCR
<u>CHR_7:[26822177,26822068]</u>	10.41	<i>mir-148a</i>	UCAGUGCACUACAGAACUUUGU	Reference set
<u>CHR_15:[25719132,25719023]</u>	10.41			Not detected by PCR
<u>CHR_19:[16614142,16614033]</u>	10.38	<i>mir-23a</i>	AUCACAUUGCCAGGGAUUUCC	Reference set
<u>CHR_6:[35221040,35220931]</u>	10.33			Not detected by PCR
<u>CHR_10:[20344745,20344636]</u>	10.32			Not detected by PCR
<u>CHR_19:[16652166,16652275]</u>	10.15	<i>mir-181c</i>	AACAUUCAACCGUCGGUGAGU	Newly identified, verified by cloning (AACAUUCAACGCUGUCGGUGAGU)
<u>CHR_5:[117678742,117678851]</u>	10.00			Not detected by PCR

¹RNA was detected from both the miRNA and the miRNA* sides of the precursors.

²The exact 3' terminus of the miRNA is unknown.

mir-140

CHROMOSOME_16:[72705524,72705633]



The highest-scoring MicroScan prediction is highlighted by purple circles, and the validated miRNA is highlighted by green circles. Human sequence is shown, with nucleotides conserved between human and Fugu shown in red.

Supplemental Table S2. Primers used to test for the presence of candidate miRNA sequences in cDNA libraries of zebrafish miRNAs.

The primers used to validate the zebrafish miRNAs shown in Table S1 are indicated (asterisks).

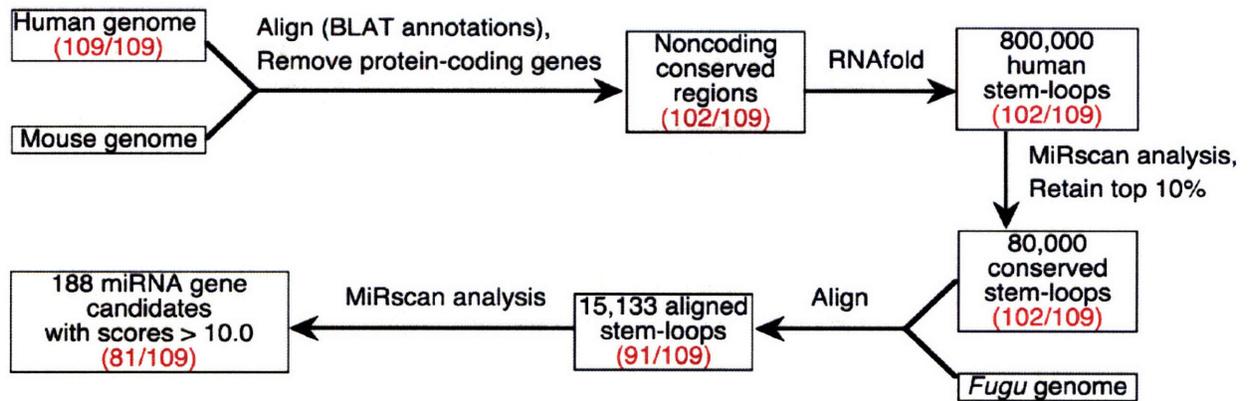
Human locus	MiRscan Score	Primers	
CHR 9: [76243263, 76243154]	17.76	CAACAAAATCACTAGTC*	TGGCAGACTGTGATTTG*
CHR 15: [59496766, 59496875]	17.43	CTGTAGGAATATGTTTG	ACCTAATATATCAAACA
CHR 17: [1722404, 1722295]	16.62	TAGCCATGACTGTAGAC*	GCAGTAAGCAGTCTAGA
CHR 1: [90399302, 90399411]	16.62	ATTAAAAAGTCCCTCTTG	ACCCTTAGGCATCAACA
CHR 7: [133435388, 133435279]	16.10	ACAGTGAATTCACCAG*	TTACCAAAGGGCCATAA*
CHR X: [119194533, 119194642]	16.04	GGGTAGGTGGAATACTA	TATATTATCCACCCAC
CHR 11: [58914650, 58914759]	15.78	GGCTGTCAATTCATAGG*	TGACCTATGAATTGACA
CHR 11: [59435455, 59435564]	15.78	GGCTGTCAATTCATAGG*	TGACCTATGAATTGACA
CHR 11: [59544282, 59544391]	15.78	GGCTGTCAATTCATAGG*	TGACCTATGAATTGACA
CHR 20: [33400991, 33401100]	15.77	TTAAACATCACTGCAAG	AGTTAAGACTTGCAGTG
CHR 5: [90467504, 90467613]	15.70	TTGTCAACAAAACCTGCT	TTGTAAGCAGTTTTGTT
CHR 5: [3220067, 3219958]	15.62	TTCACAGAGAAAACAAC	ATTAACCTATAGCCTTA
CHR 1: [93614799, 93614690]	15.59	CTTCTCAAAGATTTTCC	GACAGTTGAAATCTCTG
CHR 13: [96029172, 96029281]	15.58	CCACTCTCTCATTTATCT	TGCAGATAAACAGAGTG
CHR 5: [81367589, 81367698]	15.27	GGAGAATCAATAGGGCA	CATGGCGCTTCTCTGAC
CHR 2: [174973719, 174973828]	15.20	GAAGAGCATTAAACCATC	TCTGTCATTGTTAATTG
CHR 4: [21933011, 21933120]	15.19	CACATGGTTAGATCAAG*	GGTGCCTTGACAGAACCA
CHR 9: [122569704, 122569595]	14.85	GCAGTTGATGTCCCAA	AGAATTGCGTTTGGGAC
CHR 3: [21794989, 21794880]	14.69	GCCCCATTAATATTTTA	CTAATTAAAAATCAACA
CHR 1: [154832185, 154832076]	14.60	CTGTAGGAATATGTTTG	ACCTAATATATCAAACA
CHR 13: [72133104, 72132995]	14.52	TTATGATAGCTTCCTCA	AACCTTCAAACGAAAAA
CHR 22: [17113556, 17113665]	14.24	GCCTCAATTATTGGAAA	ATTGCTCTTTTCATTAA
CHR 4: [158780085, 158779976]	14.22	GTTGATGCGCCATTTGG	CTTAGGCCAAATGGCGC
CHR 15: [30996132, 30996023]	13.96	TTACATAAAATTAACAG	AAACTTGTTAATTGACT
CHR 10: [80541846, 80541737]	13.86	GATAAAGCCAATAAAAC	CCTGACAGTCTGAGCCA
CHR X: [119291657, 119291548]	13.62	ACCGCATCATGAACACC	AAGTGTCAGATACGGTG*
CHR 1: [214725765, 214725656]	13.60	TCCGAGTCGGAGGAGGA	TCCCTCCGCCGCCGAGG
CHR 19: [6814521, 6814630]	13.58	CAACAAAATCACTAGTC*	AGGCAGACTGTGACTTG*
CHR 5: [161016881, 161016772]	13.58	CACATGGTTAGATCAAG*	GGTGCCTTGACAGAACCA
CHR 13: [90861308, 90861199]	13.37	AATCAGTACTGGATTGC	TTACAGCAATCCAGTAC
CHR 6: [43048796, 43048905]	13.01	AGAATTGCGTTTGGACA*	GCTCCTGATTGTCCAAA
CHR 2: [58511309, 58511418]	12.59	TACTCACAGTTGCCAGC*	AGAACCCCGAGCAGCGCC
CHR 18: [34497140, 34497031]	12.50	CACTGGCTGCAACACAA*	GTGTCTTGTGTTGCAGC
CHR 20: [22339705, 22339814]	12.47	CAAAAATTATCAGCCAG	ACAAAAGGCTGACAGCC
CHR 10: [124638355, 124638464]	12.44	CATGCAAGTATGAAAAT	TCTGATTAATCAAGCCT
CHR 10: [127681756, 127681865]	12.44	TCTGATTAATCAAGCCT	CATGCAAGTATGAAAAT
CHR 15: [85832771, 85832880]	12.30	CAACAAAATCACTAGTC*	GAAGACTAGTGATTTTG
CHR 7: [16645672, 16645781]	12.09	GGTAAATTGCTTGCAAA	CAAAATTGCAAGCAAAT
CHR 17: [51122450, 51122559]	12.05	ATCCACAAAGCTGAACA	GACATGTAGACTCTTTG
CHR 1: [231366714, 231366605]	11.97	GGCTGTCAATTCATAGG*	TGACCTATGAATTGACA

CHR X: [60718647,60718756]	11.92	TGGGGTATTTGACAAAC*	CTTAGAGTATTTGACAG
CHR 2: [170125323,170125214]	11.79	AGCAAGATGCCTGCCTC	TATTATCAGCATCTGCA
CHR 1: [90399570,90399461]	11.39	ATGACTACAAGTTTATG	ATGGCCATGTCCTGTAG
CHR 22: [39875393,39875284]	11.06	CGCCTGTACCTCAACC	TAGCAAGAAAAAGCCCC
CHR 2: [155447946,155447837]	11.03	TC'TTGCTCTAACACTTG	CCTGACAGTCTGAGCCA
CHR 4: [158780484,158780375]	11.01	CGATAAAACATAACTTG	CAAGTTATGTTTTATCG
CHR 10: [82216993,82216884]	10.99	CACCAAAC TGACAGGAA	CAGTAACTTCTGTGCAG
CHR 9: [5165902,5166011]	10.96	GCCAGACTCTCTTCTCG	ACGCCAATAAGAGGTAA
CHR 7: [133430866,133430757]	10.95	TAGTTGGCAAGTCTAGA*	TTTGGCAATGGTAGAAC
CHR 18: [75755597,75755706]	10.82	AATATGTCATATCAAAG	ACAGACCTTCCCATGCA
CHR 5: [40367092,40367201]	10.68	TGAGAGTAGCATGTTTG	GGAGACAAACATGCTAC
CHR 15: [58026411,58026520]	10.45	TAATTGGACAAAGTGCC	
CHR 15: [26169137,26169028]	10.42	ACCGAACAAAGTCTGAC	TCTCGTGCCAGACCCGG
CHR 11: [136385422,136385531]	10.42	AACTACTTCCAGACCAG	
CHR 15: [25719132,25719023]	10.41	GCAAGCGCAGCTCCACA	GGCCGTGAGGACCACAC
CHR 6: [35221040,35220931]	10.33	GAGCTGCTCAGCTGGCC	
CHR 10: [20344745,20344636]	10.32	TGAGAGTAGCATGTTTG	
CHR 5: [117678742,117678851]	10.00	TACTTTTCATTTCCCTC	AATGATGACAGAGAAAG

Supplemental Fig. S1. The computational identification of vertebrate miRNA genes.

The computational procedure used to identify vertebrate miRNA genes is summarised in the flowchart, with the fraction of known human loci retained at each step of the analysis shown in red.

Supplemental Figure 1



REFERENCES

- Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J. M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., *et al.* (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science* 297, 1301-1310.
- Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, S., Tacker, M., and Schuster, P. (1994). Fast folding and comparison of RNA secondary structures. *Monatshefte f Chemie* 125, 167-188.
- Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., Down, T., *et al.* (2002). The Ensembl genome database project. *Nucleic Acids Res* 30, 38-41.
- Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res* 12, 656-664.
- 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. *Current Biology* 12, 735-739.
- 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.
- Lim, L. P., Lau, N. C., Weinstein, E. G., Abdelhakim, A., Yekta, S., Rhoades, M. W., Burge, C. B., and Bartel, D. P. (2003). The microRNAs of *Caenorhabditis elegans*. *Genes Dev* 17, 991-1008.
- Moss, E. G., and Poethig, R. S. (2002). MicroRNAs: something new under the sun. *Cur Biol* 12, 688-690.
- Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. (2002). miRNPs: a novel class of ribonucleoproteins containing numerous microRNAs. *Genes Dev* 16, 720-728.
- Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 22, 4673-4680.

Chapter Two

MicroRNA-directed cleavage of *Hoxb8* mRNA

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

ABSTRACT

MicroRNAs (miRNAs) are endogenous ~22-nucleotide RNAs, some of which are known to play important regulatory roles in animals by targeting the messages of protein-coding genes for translational repression. We find that miR-196, a miRNA encoded at three paralogous locations in the A, B, and C mammalian Hox clusters, has extensive, evolutionarily conserved complementarity to messages of *Hoxb8*, *Hoxc8*, and *Hoxd8*. RNA fragments diagnostic of miR-196-directed cleavage of *Hoxb8* were detected in mouse embryos. Cell culture experiments demonstrated down-regulation of *Hoxb8*, *Hoxc8*, *Hoxd8*, and *Hoxa7* and supported the cleavage mechanism for miR-196-directed repression of *Hoxb8*. These results point to a miRNA-mediated mechanism for the posttranscriptional restriction of Hox gene expression during vertebrate development and demonstrate that metazoan miRNAs can repress expression of their natural targets through mRNA cleavage in addition to inhibiting productive translation.

Over 1% of the predicted mammalian genes encode microRNAs (miRNAs) (Bartel and Bartel, 2003; Lagos-Quintana et al., 2003; Lim et al., 2003). As previously reported for miR-10 (Lagos-Quintana et al., 2003), genes for miR-196 (Lagos-Quintana et al., 2003; Lim et al., 2003) map to homeobox (Hox) clusters (Fig. 1 and table S1). Hox clusters are groups of related transcription factor genes crucial for numerous developmental programs in animals (Krumlauf, 1994). Mammals have four Hox clusters (HoxA to D) containing a total of 39 genes organised into 13 paralogous subgroups (Fig. 1) (Krumlauf, 1994).

The miR-196 miRNAs have intriguing complementarity to sites in the 3' untranslated regions (3' UTRs) of Hox genes representing each cluster. With the exception of a single G:U wobble, pairing between miR-196a and the human *Hoxb8* 3' UTR is perfect. The functional importance of this miR-196 complementary site is supported by its conservation in the fish and frog *Hoxb8* 3' UTRs, despite the divergence of surrounding UTR sequences (Fig. 2A). This conserved, near-perfect pairing suggested that, like natural miRNA targets in plants (Kasschau et al., 2003; Llave et al., 2002; Tang et al., 2003) or engineered miRNA targets in animals (Hutvagner and Zamore, 2002; Zeng et al., 2002), *Hoxb8* mRNA is targeted for cleavage. MicroRNA-directed cleavage can be detected by using a modified form of 5'-RACE (rapid amplification of cDNA ends) because the 3' product of this cleavage has two unusual properties: (i) a 5' terminus that is a suitable substrate, without further modification, for ligation to an RNA adaptor using T4 RNA ligase and (ii) a 5' terminus that maps precisely to the nucleotide that pairs with the tenth nucleotide of the miRNA (Kasschau et al., 2003; Llave et al., 2002). Accordingly, we used this method to examine whether *Hoxb8* mRNA was a natural target of miR-196-directed cleavage. Having determined that miR-196 is expressed during mouse embryogenesis starting at or before day 7 (See materials and methods within supplemental material, Fig. S1), total RNA from day 15 to day 17 mouse embryos was chosen for this analysis. Of the eight 5'-RACE clones that ended in the vicinity of the miR-196 complementary site [i.e., within the 150-nucleotide (nt) segment centring on the complementary site], seven terminated precisely at the position diagnostic of miR-196-directed cleavage (Fig. 2A, the eighth terminated 41 nt downstream).

By analyzing RNA isolated from the mouse, the *Hoxb8* 5'-RACE results validated this miRNA-target interaction in the animal, whereas previous experimental support for predicted mammalian miRNA targets has come from reporter assays in cultured cells (Lewis et al., 2003). Moreover, this experiment detected miRNA-mediated cleavage of the targeted message, whereas, in the previously examined metazoan examples, translational inhibition had been the mechanism of endogenous miRNA-mediated repression (Kasschau et al., 2003). Nevertheless, the *Hoxb8* 5'-RACE results do not rule out the possibility that the predominant mode of silencing is translational inhibition, as illustrated for miR172 regulation of *Arabidopsis APATELA2* (Aukerman and Sakai, 2003; Chen, 2003; Kasschau et al., 2003). To explore the mechanism of *Hoxb8* repression, the miR-196 complementary site from *Hoxb8* or control complementary sites were placed into the 3' UTR of the firefly luciferase reporter gene, and the reporter plasmid was cotransfected into HeLa cells together with a transfection control and either cognate or noncognate miRNAs. As expected, miR-196a inhibited luciferase expression from the construct with the complementary site from *Hoxb8* mRNA (Fig. 2B). Inhibition was essentially the same as that observed for a reporter with perfect antisense complementarity to the miRNA (miR-196a-as), indicating that the conserved G:U wobble involving U5 of miR-196 (Fig. 2A) does not substantially decrease miRNA-directed inhibition. Accompanying this inhibition was a substantial decrease in the amount of reporter mRNA, again similar to that seen with perfect

antisense complementarity (Fig. 2C), indicating that a large fraction of the miR-196-directed repression occurred through mRNA degradation.

Genes from each of the other clusters also appear to be miR-196 targets (Fig. 1). The *Hoxa7* 3' UTR has multiple conserved matches to residues 2 to 8 of miR-196, called "seed matches," which previously identified it as a likely miR-196 target (Fig. 3A) (Lewis et al., 2003). *Hoxc8* and *Hoxd8* UTRs have both seed matches and more extensive complementary sites (Fig. 3A), although none of these sites resemble the *Hoxb8* site in having perfect pairing at their centre (Fig. S2). Segments from all three UTRs imparted miR-196-dependent repression to the luciferase reporter without a substantial decrease in reporter mRNA, indicating predominantly translational inhibition (Fig. 3, B and C). To the extent that these experiments in cell culture reflect regulation in animals, miR-196 represses its targets by two posttranscriptional mechanisms. As reported for engineered targets (Doench et al., 2003; Hutvagner and Zamore, 2002; Zeng et al., 2002; Zeng et al., 2003), the choice of mechanism appears to depend on the nature of the complementary sites, with translational inhibition apparently requiring more complementarity sites but less pairing within each site.

For some Hox genes, posttranscriptional processes combine with complex transcriptional regulation to generate the pattern of Hox expression seen in early development (Brend et al., 2003; Nelson et al., 1996). Our results indicate that miR-196 delimits or dampens the expression of *Hoxb8* and probably also *Hoxc8*, *Hoxd8*, and *Hoxa7*, thus pointing to additional posttranscriptional control of genes in the Hox clusters. The temporal and spatial order of Hox gene expression along the anterior-posterior axis is initially colinear with the physical position of the Hox genes on their chromosomes (Kmita and Duboule, 2003; Krumlauf, 1994). If the same is true for *mir-196* genes, their locations suggest that they would initially be transcribed slightly later than their targets and could help define the posterior boundary of target gene expression. Close orthologues of miR-196 have not been found in invertebrates, although miR-196 is a distant homologue of the *let-7* miRNAs (Lim et al., 2003a). Nonetheless, an unrelated miRNA, miR-iab-4, maps to the corresponding region in the Bithorax complex of insect Hox genes (Aravin et al., 2003), and this miRNA is predicted to target *Ubx* (Stark et al., 2003), an insect counterpart of the Hox 6 to 8 paralogous groups, suggesting that analogous regulation might extend to insects (Fig. 1).

The observation of endogenous miRNA-directed mRNA cleavage in animals raises the question of how many other metazoan miRNA targets might be down-regulated by cleavage. Few other human messages have such extensive complementarity to the known miRNAs. Nonetheless, analysis of off-target siRNA-mediated regulation indicates that extensive complementarity is not always required (Jackson et al., 2003), leaving open the possibility that a large fraction of metazoan miRNA regulation might be achieved through mRNA cleavage

ACKNOWLEDGEMENTS

We thank Wendy Johnston for plasmid construction, John Doench and Phil Sharp for sharing ribonuclease-protection protocols and reagents, Harvey Lodish and David Page for sharing facilities, and Andrew Chess, and Cliff Tabin for comments on this manuscript.

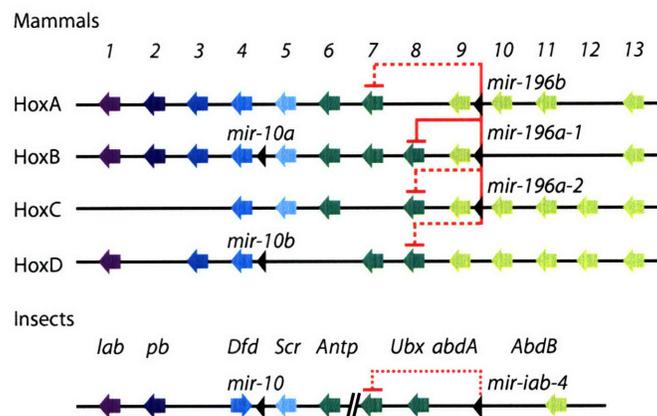


Fig. 1. Genomic organisation of Hox clusters. Colored arrows indicate Hox genes representing 13 paralogous groups; black arrowheads depict miRNA genes. Repression supported by bioinformatic evidence only (dotted red line), cell-culture and bioinformatic evidence (dashed line), or in vivo, cell culture, and bioinformatic evidence (solid line) are indicated. The vertical red line indicates that miRNAs from any of the three loci could repress the targets.

Fig. 2. miR-196–directed cleavage of *Hoxb8* mRNA. (A) Sequence alignment of the miR-196 complementary site in the 3' UTRs of *Hoxb8* genes (Table S2). Absolutely conserved nucleotides are highlighted in black; those of the complementary site are either red (Watson-Crick pairing to miR-196a) or pink (G:U wobble). The vertical arrowhead indicates the 5' end of cleavage products detected by 5'-RACE, with the frequency of clones noted. *Hs*, human; *Mm*, mouse; *Rn*, rat; *Xl*, frog; *Dr*, zebrafish; *Fr*, pufferfish. (B) Response to miR-196a in HeLa cells (Doench et al., 2003). Firefly luciferase reporters containing either the miR-196 complementary site from mammalian *Hoxb8*, or the perfect antisense sequence of miR-196a or miR-1d (*Hoxb8*, miR-196a-as, and miR-1d-as, respectively) were cotransfected with the indicated amount of cognate or noncognate miRNA (See materials and methods within supplemental material). Firefly luciferase activity was normalised to that of the Renilla transfection control, then activity with cognate miRNA (solid) was normalised to the median activity with noncognate miRNA at the same concentration (open). Each box represents the distribution of activity measured for each reporter ($n = 9$; bars define 10th and 90th percentiles; box spans 25th and 75th percentiles; line and number indicate median normalised activity). In each case, the expression observed for the cognate miRNA significantly differed from that for the noncognate miRNA ($P < 0.001$, Mann-Whitney test). (C) Reporter mRNA levels monitored with a ribonuclease protection assay (See materials and methods within supplemental material, (Doench and Sharp, 2004)). Protected fragments of each probe are indicated (arrowheads), and were absent when cells did not receive reporter plasmid (\emptyset). Intensity of firefly relative to Renilla bands (Firefly/Renilla) is shown and normalised to that of the noncognate miRNA. In three independent repetitions of this experiment, normalised mRNA levels were 0.35 ± 0.03 , 0.53 ± 0.02 , and 0.52 ± 0.03 for miR-1d-as, miR-196a-as, and *HOXB8*, respectively (\pm SD).

Figure 2

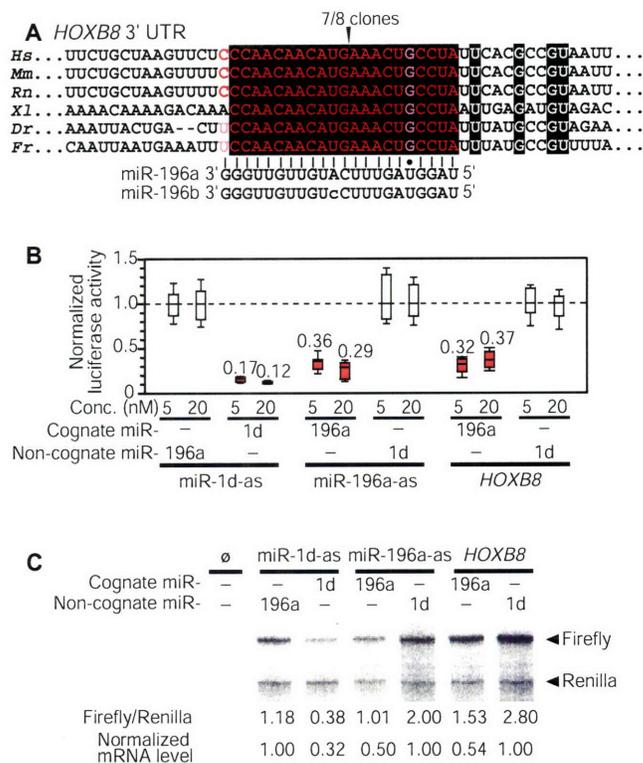
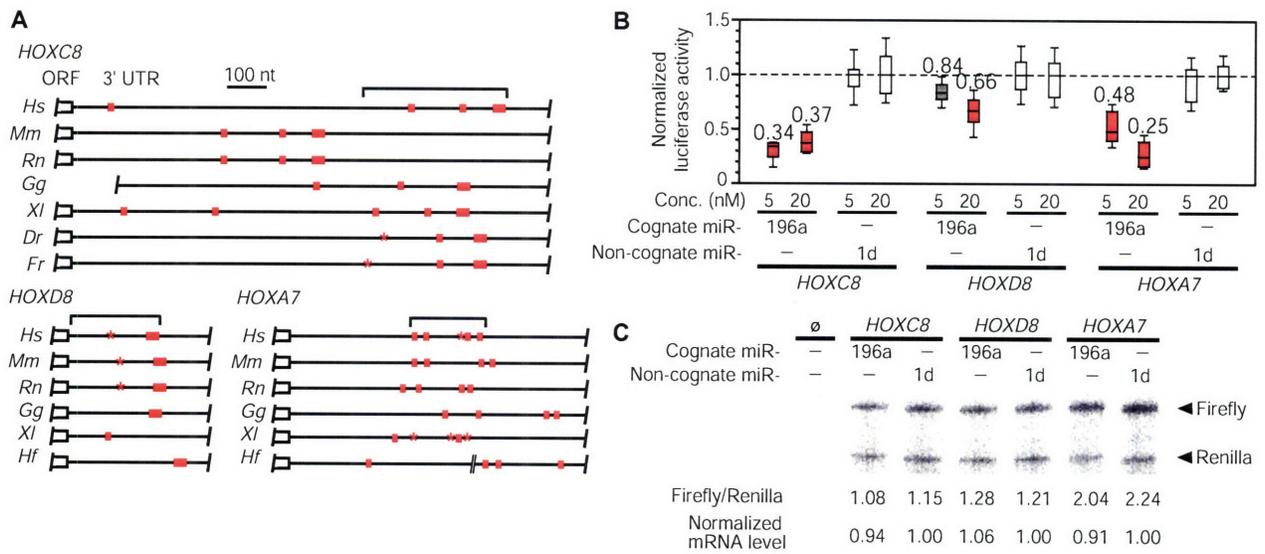


Fig. 3. Additional HOX genes predicted to be regulatory targets of miR-196. (A) miR-196 complementary sites in *Hoxc8*, *Hoxd8*, and *Hoxa7* 3' UTRs (Table S2). Wide red bars indicate extensive complementarity, narrow bars indicate seed matches, and asterisks indicate slightly relaxed seed matches that permit a G:U wobble involving U8 of miR-196 (Fig. S2). Abbreviations as in Fig. 1A, plus *Gg*, chicken, and *Hf*, hornshark. (B) Experimental support for predicted targets. Experiments were analogous to those of Fig. 2B, except the firefly reporters contained fragments of human *Hoxc8*, *Hoxd8*, and *Hoxa7* UTRs bracketed in (A) (Table S3). Red boxes denote statistically significant repression ($P < 0.001$, Mann-Whitney test). (C) Reporter mRNA levels monitored as in Fig. 2C. In three independent repetitions, normalised mRNA levels were 1.01 ± 0.08 , 1.00 ± 0.09 , and 0.95 ± 0.06 for *Hoxc8*, *Hoxd8*, and *Hoxa7*, respectively (\pm SD).

Figure 3



SUPPLEMENTAL MATERIAL

Materials and Methods

5'-RACE of Mouse *Hoxb8*

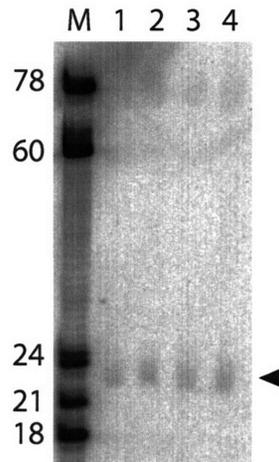
Total RNA from day-15 and day-17 mouse embryo (Clontech) was pooled and subjected to modified 5'-RACE (Llave et al., 2002). The total RNA was ligated to a synthetic RNA (5'-CGACUGGAGCACGAGGACACUGACAUGGACUGAAGGAGUAGAAA-3') and reversely transcribed using a gene-specific primer (5'CCTTTTCAGGCGCAGACAACAGAAC-3'). The cDNA was PCR-amplified with a non-specific forward primer (5'-CGACTGGAGCACGAGGACACTGA-3') and the same reverse primer. The PCR products were further amplified by nested PCR using another forward (5'-GGACTGACATGGACTGAAGGAGTA-3') and reverse (5'-CCATAAAGCAATTCACAGATACAGG-3') primer. PCR DNA was cloned (TOPO TA cloning kit, Invitrogen) and sequenced.

Cell Culture and Luciferase Assays

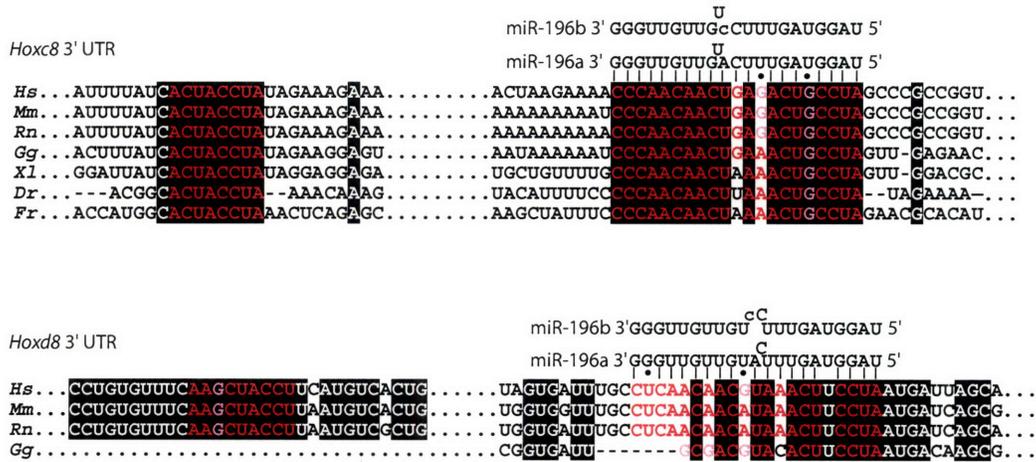
Construction of the firefly luciferase vectors was as described (Lewis et al., 2003). The sequence of each insert is shown (Table S3). Adherent HeLa S3 cells were grown in complete DMEM to 90% confluency in 24-well plates. Cells were transfected with 0.4 µg of the firefly luciferase reporter vector, 0.08 µg of the control vector containing Renilla luciferase, pRL-TK (Promega), and 2.5 or 10 pmol of synthetic miRNA duplex (Table S3; Dharmacon) in a final volume of 0.5 ml using Lipofectamine 2000 (Invitrogen) (Doench et al., 2003). Thirty hours after transfection, firefly and Renilla luciferase activities were measured consecutively using the Dual-luciferase assays (Promega).

RNase Protection Assay

RNase protection assays were modeled after those of Doench and Sharp (Doench and Sharp, 2004). HeLa S3 cells were seeded in 60-mm dishes and transfected with 1.0 µg of firefly luciferase reporter vector, 7.0 µg of pRL-TK, and 100 pmol of synthetic miRNA duplex in a final volume of 5 ml using Lipofectamine 2000. Luciferase assays confirmed that repression was indistinguishable from that seen in the experiments of Fig. 2B. Thirty hours after transfection, total RNA was extracted with the RNAeasy kit (Qiagen) including DNase I treatment. The firefly luciferase probe corresponded to nucleotides 1142–1429 of the mRNA plus non-complementary vector sequence, and the Renilla luciferase probe corresponded to nucleotides 1068–1297 of the mRNA plus non-complementary vector sequence. The radioactive-labeled probes were made by in vitro transcription (MAXIscript, Ambion), and RNase protection assays were performed according to the manual (RPA III, Ambion) using 15 µg of total RNA from each sample. Control experiments confirmed that probes were in excess over the reporter mRNA.



Supplemental Fig. S1. Expression of miR-196 during mouse embryonic development. The Northern blot (Lau et al., 2001) probes total RNA from 7-day, 11-day, 15-day, and 17-day mouse embryos (lanes 1-4, respectively). The oligonucleotide used as a probe was 5'-CCCAACAACATGAACTACCTA-3'. M indicates 33 P-labeled size markers.



Supplemental Fig. S2. Sequence alignment of the miR-196 complementary site, and a representative complementary seed and flanking regions in the 3' UTRs of vertebrate *Hoxc8* and *Hoxd8* genes. Nucleotides absolutely conserved are highlighted in black. Sequences of both miR-196a and miR-196b are shown above the alignment, with the single-nucleotide difference between these miRNAs indicated by a lowercase letter in miR-196b. Nucleotides of the miR-196 complementary site and the seed are either red (Watson-Crick pairing to miR-196a) or pink (G:U wobbles). Species abbreviations as in Fig. 1A and 2A.

Supplemental Table S1. The sequence and genomic context of the vertebrate miR-196 family. The miRNA stem-loops were identified on the basis of homology to known human, mouse, and zebrafish sequences, and grouped according to the Hox cluster in which they reside. Only miRNA genes that were localized to Hox clusters are reported. Residues of the mature miRNA are in bold, while miRNAs validated by cloning are in red. Asterisks represent the positions of perfectly conserved residues (Lagos-Quintana et al., 2003; Lim et al., 2003)

Species	Predicted sequence of miRNA stem-loop	Genomic Context
HoxB cluster (miR-196a-1 in mammals)		
<i>Homo sapiens</i> (Human)	AAUUGGAACUCGUGAGUGAAU UAGGUAGUUUCAUGUUGUUGG CCUUGGUUUC-UGAACACAACAACAUUAAACCACCCGAUUCACGGCAGUUACUGCUC	Between <i>Hoxb9</i> and <i>Hoxb13</i>
<i>Mus musculus</i> (Mouse)	AGCCGGGACUGUUGAGUGAAG UAGGUAGUUUCAUGUUGUUGG CCUUGGUUUC-UGAACACAACGACAUCAAACCACCUUGAUUCACGGCAGUUACUGCUCU	Between <i>Hoxb9</i> and <i>Hoxb13</i>
<i>Rattus norvegicus</i> (Rat)	AACUGGGACUCGUGAGUGAAG UAGGUAGUUUCGUGUUGUUGG CCUUGGUUUC-UGAACACAACAACACCAAACCACCUUGAUUCACGGCAGUUACUGCUCU	Between <i>Hoxb9</i> and <i>Hoxb13</i>
<i>Gallus gallus</i> (Chicken)	-----AACUGCUCUGUGAAU UAGGUAGUUUCAUGUUGUUGG CUUAAAUUU-UAAACACAAGAACAUCAAACUACCUUGAUUUACUCCAGUU-----	In HoxB cluster
<i>Fugu rubripes</i> (Pufferfish)	CACCGGAGCGGU--GUGAU UAGGUAGUUCAAGUUGUUGG CUGAACUCUUGUGAUCACAGGAACCUUGAAACUGCCUGAGUCACGGCAGUACCGCUG	Between <i>Hoxb9</i> and <i>Hoxb13</i>
	* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *	
HoxC cluster (miR-196a-2 in mammals)		
<i>Homo sapiens</i> (Human)	CUCAGCUGAUCUGUGGU UAGGUAGUUUCAUGUUGUUGG GAUUGAGUUUUG--AACUCGGCAACAAGAAACUGCCUGAGUUACAUCAGUCGGUU	Between <i>Hoxc9</i> and <i>Hoxc10</i>
<i>Mus musculus</i> (Mouse)	---AGCUGAUCUGUGGU UAGGUAGUUUCAUGUUGUUGG GAUUGAGUUUUG--AACUCGGCAACAAGAAACUGCCUGAGUUACAUCAGUC---	Between <i>Hoxc9</i> and <i>Hoxc10</i>
<i>Rattus norvegicus</i> (Rat)	---AGCUGAUCUGUGGU UAGGUAGUUUCAUGUUGUUGG AUUGAGUUUUG--AACUCGGCAACAAGAAACUGCCUGAGUUACAUCAGUCGGUU	Between <i>Hoxc9</i> and <i>Hoxc10</i>
<i>Gallus gallus</i> (Chicken)	UGCAGCUGAUCUGUGGU UAGGUAGUUUCAUGUUGUUGG AUUGGCUUUUA--GCUCGGCAACAAGAAACUGCCUUAUUUACGUCAGUUAGUC	Nearby <i>Hoxc10</i>
<i>Xenopus tropicalis</i> (Frog)	--CAGCUGAUCUGUGGU UAGGUAGUUUCAUGUUGUUGG AUUGCUUUUUUUAACGCGGCAACAAGAAACUGCCUUAUUUACGUCAGUU---	In HoxC cluster
<i>Danio rerio</i> (Zebrafish)	---GGCUGUGCGUGGU UAGGUAGUUUCAUGUUGUUGG AUUGGCUUCCU--GGCUCGACAACAAGAAACUGCCUUGAUUACGUCAGUUCGUC	Between <i>Hoxc9a</i> and <i>Hoxc10a</i>
<i>Danio rerio</i> (Zebrafish)	---AGCUGAUCUGUGGU UAGGUAGUUUCAUGUUGUUGG GUUGACUUCU--GGCUCGACAACAAGAAACUGCCUUGAUUACGUCAGUU---	Between <i>Hoxc6b</i> and <i>Hoxc11b</i>
<i>Fugu rubripes</i> (Pufferfish)	CGAAGCUGGAGCGUGGU UAGGUAGUUUCAUGUUGUUGG GAUGGCUUCCU--GGCUCGGCAACAAGAAACUGCCUUGAUUACGUCAGUUCGUC	Between <i>Hoxc9</i> and <i>Hoxc11</i>
	* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *	
HoxA cluster (miR-196b in mammals)		
<i>Homo sapiens</i> (Human)	-ACUGGUCGGUGAU UAGGUAGUUCCUGUUGUUGG A-UCCACCUUUCUCUCGACAGCAGCAGACUGCCUUCAUUACUUC--AGUUG	Between <i>Hoxa9</i> and <i>Hoxa10</i>
<i>Mus musculus</i> (Mouse)	AACUGGUCGGUGAU UAGGUAGUUCCUGUUGUUGG A-UCCACCUUUCUCUCGACAGCAGCAGACUGCCUUCAUUACUUC--AGUUG	Between <i>Hoxa9</i> and <i>Hoxa10</i>
<i>Rattus norvegicus</i> (Rat)	AACUGGUCGGUGAU UAGGUAGUUCCUGUUGUUGG A-UCCACCUUUCUCUCGACAGCAGCAGACUGCCUUCAUUACUUC--AGUUG	Between <i>Hoxa9</i> and <i>Hoxa10</i>
<i>Gallus gallus</i> (Chicken)	---UGCUCUGUGGU UAGGUAGUUUCAUGUUGUUGG GCUCACCUUUCUCUCGACAGCAGCAGAAACUGCCUUAUUUACUUC--AGUUG	In HoxA cluster
<i>Danio rerio</i> (Zebrafish)	GACUGUCGAGUGGU UAGGUAGUUUCAUGUUGUUGG AUUAACAUCAAA-CUCUGCAACGUGAAACUGUCUUAUUUGCCCC--AGUU-	Nearby <i>Hoxa9a</i>
<i>Danio rerio</i> (Zebrafish)	AACUGCUAAGUGAU UAGGUAGUUUCAUGUUGUUGG CUCAUUAUUUAUCCCCGCAACACGAAACUGUCUUAUUUGCCUCGAGUGA	Between <i>Hoxa9b</i> and <i>Hoxa10b</i>
<i>Fugu rubripes</i> (Pufferfish)	GACUGUCGAGUGGU UAGGUAGUUUCAUGUUGUUGG GUCCAUUUCAA-CUCUGCAACUGAAACUGUCUUAUUUGCCCC--AGUUA	Between <i>Hoxa9</i> and <i>Hoxa10</i>
<i>Heterodontus francisci</i> (Hornshark)	AACUGCGUGUGAU UAGGUAGUUUCAUGUUGUUGG G-CUCAAGUCUAUCUCUACAACACGAAACUGCCUGAAUUACUGC--AGUU-	Between <i>Hoxa9</i> and <i>Hoxa10</i>
	* * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * * *	

Supplemental Table S2. Sequence retrieval information. The accession numbers or genomic loci of Hoxa7, Hoxb8, Hoxc8 and Hoxd8 are shown below. In most cases, the 3' UTRs are undefined. When they extend beyond the annotated region, the accession numbers can be used as a starting point to obtain downstream sequences.

Species	Hoxa7	Hoxb8	Hoxc8	Hoxd8
<i>Homo sapiens</i> (Human)	ENSG00000122592 (Ensembl)	ENSG00000120068 (Ensembl)	ENSG00000037965 (Ensembl)	NM_019558.2 (Genbank)
<i>Mus musculus</i> (Mouse)	ENSMUSG00000038236 (Ensembl)	ENSMUSG00000056648 (Ensembl)	ENSMUSG00000001657 (Ensembl)	ENSMUSG00000027102 (Ensembl)
<i>Rattus norvegicus</i> (Rat)	ENSRNOG00000006544 (Ensembl)	ENSRNOG00000007585 (Ensembl)	ENSRNOG00000016382 (Ensembl)	3.57217400-57218670 (RGSC3.1, Ensembl)
<i>Gallus gallus</i> (Chicken)	chr2:31809235-31810235 (UCSC Feb. 2004)	GGU81801* (Genbank)	chr1:87932700-87937357 (UCSC Feb. 2004)	X57158.1 (Genbank)
<i>Xenopus laevis</i> (Frog)	M24752.1 (Genbank)	TC144376 (TIGR XGI ver.5.0)	XLAF001596 (Genbank)	BC060408.1 (Genbank)
<i>Danio rerio</i> (Zebrafish)	Gene not in fish	ENSDARG00000014115 (Ensembl)	Y14544.1 (Genbank)	Gene not in fish
<i>Fugu rubripes</i> (Pufferfish)	Gene not in fish	SINFRUG00000136620 (Ensembl)	SINFRUG00000146328 (Ensembl)	Gene not in fish
<i>Heterodontus francisci</i> (Hornshark)	AF224262.1 (Genbank)	Cluster not in shark	Cluster not in shark	AF224263.1** (Genbank)

* Complete 3' UTR sequence currently unavailable.

** The miR-196 complementary site of horn shark Hoxd8 (CCCAACAACATGAACTGCCTA) resembles that of the Hoxc8 group, and thus was not included in the Hoxd8 alignment (Sup. Fig. S2).

Sequences can be retrieved at the following web locations:

Genbank IDs <http://www.ncbi.nlm.nih.gov>
 Ensembl genes <http://www.ensembl.org>
 UCSC Chicken Genome Assembly <http://genome.ucsc.edu>
 TIGR Frog Gene Index <http://www.tigr.org>

Supplemental Table S3. Reporter inserts and synthetic miRNA sequences. Natural sites of extended complementarity (red) and seed matches (blue) are indicated. Restriction sites introduced to enable directional cloning downstream of the luciferase open reading frame are in lower case. Synthetic miRNA duplexes are shown, with each miRNA in bold and the U's introduced to direct the proper strand into the RISC in red (Schwarz et al., 2003).

Reporter inserts

Hoxb8 3' UTR

5' -gagctc**CCCAACAACATGAAACTGCCTA**tctaga

Hoxc8 3' UTR

5' -gagctcTGCAGTCGCCTCTAAAATCCTACCTAACCATCCCATGGTCACTCGGGCCCATGCCTTCC
TCTCCTTCGCTGTTTGGATTTCTATTCTGTTGGGCCCGCCTTCCTCTGAGCTGCATTAGTGTAGTGCAGAAAT
CACCATAATCACGAAAAATAATAATAATAAAATCTTTAACATA**ACTACCT**AAAGGGAACCTGCAATAATCTTGAAAA
GAAAAAGAGAAAAATTTAAAAATCCTGCTATAGGAGAAAAAAGAGAAAAAATAAAATCAAAAAAAAAAAAA
GAAAGAAAGAAACCTCCAGCGTATTTTATC**ACTACCT**ATAGAAAGAAATCCTGCTTTGAGAGTATTTGTAATGCG
GTTTTGTTGTCGTTTGGTGGCTGCTTATTTCACTAAGAAAA**CCCAACAAC**T**GAGACTGCCTAG**CCCTctaga

Hoxd8 3' UTR

5' -gagctcCCGAAGGCCTGACAAATTAACCTTCTACCTTTAAAATTTACCACAGACTATTAAAACCTAA
TAATCACCATATGCTGTGGACACCACCTATTTTCTTTGTTGGAAAGGACCTTACCTGTGTTTCA**AGCTACCT**TCA
TGCTACTGCTCTTGAGGTTTTCTGTGCTTTGAGAGGGATTTGGGTGTTTAAAAAGTTTCTAGTATCACATAGAA
GCTGTCCCTTGAGCTGTCCATGGAAGGGTAATTTGATACTGACCTTGTAGCTATATTTTATAATGGTTTTTAAAT
GTCTGAGCTAGTGATTTGC**CTCAACAACGTAAACTTCCTA**ATGAtctaga

Hoxa7 3' UTR

5' -gagctcCCAGGCCAGCCGGCCCTGCTCTGGCGCGTCCAAAAT**ACTACCT**AGCACAGGCCTCTGCT
CGAGGCACCCCCAA**ACTACCT**ATGTATCCAGCCCCAGAGGGCCTCCATTTCCAGGAAGTCCCTATGTATCCCAAC
ACTGGCAGACACCCAGCACCACCTCCAGACCCGCAAGAAAGTGAATCTCACT**ACTACCT**ACTCCCCTAA**ACT**
ACCTATTTTTGTGCTGGCTctaga

miR-196a-as

5' -gagctcCCCAACAACATGAAACTACCTAtctaga

miR-1d-as

5' -gagctcATACATACTTCTTTACATTCCAtctaga

Synthetic miRNA duplexes

miR-196a

5' -p**UAGGUAGUUUCAUGUUGUUGG**
GAAU**U**CAUCAAAAGUACAACAACp-5'

miR-1d

5' -p**UGGAAUGUAAAAGAUGUAU**
AUA**UUU**ACAUUUUCUUAUACp-5'

REFERENCES

- Aravin, A. A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. (2003). The small RNA profile during *Drosophila melanogaster* development. *Dev Cell* 5, 337-350.
- Aukerman, M. J., and Sakai, H. (2003). Regulation of Flowering Time and Floral Organ Identity by a MicroRNA and Its APETALA2-Like Target Genes. *Plant Cell* 10, 10.
- Bartel, B., and Bartel, D. P. (2003). MicroRNAs: At the Root of Plant Development? *Plant Physiol* 132, 709-717.
- Brend, T., Gilthorpe, J., Summerbell, D., and Rigby, P. W. (2003). Multiple levels of transcriptional and post-transcriptional regulation are required to define the domain of Hoxb4 expression. *Development* 130, 2717-2728.
- Chen, X. (2003). A MicroRNA as a Translational Repressor of APETALA2 in Arabidopsis Flower Development. *Science*, Published online September 11 2003; 2010.1126/science.1088060.
- Doench, J. G., Peterson, 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.
- 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.
- Kasschau, K. D., Xie, Z., Allen, E., Llave, C., Chapman, E. J., Krizan, K. A., and Carrington, J. C. (2003). P1/HC-Pro, a viral suppressor of RNA silencing, interferes with *Arabidopsis* development and miRNA function. *Dev Cell* 4, 205-217.
- Kmita, M., and Duboule, D. (2003). Organizing axes in time and space; 25 years of colinear tinkering. *Science* 301, 331-333.
- Krumlauf, R. (1994). Hox genes in vertebrate development. *Cell* 78, 191-201.
- Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. (2003). New microRNAs from mouse and human. *RNA* 9, 175-179.
- 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.

Lewis, B. P., Shih, I., 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. (2003). Vertebrate microRNA genes. *Science* *299*, 1540.

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

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.

Nelson, C. E., Morgan, B. A., Burke, A. C., Laufer, E., DiMambro, E., Murtaugh, L. C., Gonzales, E., Tessarollo, L., Parada, L. F., and Tabin, C. (1996). Analysis of Hox gene expression in the chick limb bud. *Development* *122*, 1449-1466.

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

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.

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

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

Chapter Three

Regulation by microRNAs contributes to the functional hierarchy among vertebrate Hox genes

ABSTRACT

The Hox family of homeotic genes encode transcription factors that specify regional identities in vertebrate embryos (Krumlauf, 1994; Maconochie et al., 1996). The miR-196 and miR-10 families are transcribed from genomic loci within clusters of Hox genes, and in turn mediate the posttranscriptional repression of neighbouring Hox transcripts (Yekta et al., 2004). In this report we provide evidence for extensive targeting by miR-196 and miR-10 (Hox miRNAs) of Hox genes located in paralagous groups that are 3' but not 5' of each miRNA locus. Blocking the two Hox miRNAs in avian embryos, results in axial skeletal patterning defects in domains that overlap considerably with Hox target expression. We propose that miR-196 and miR-10 act to refine posterior boundaries at relative levels of expression for multiple Hox genes, thus setting appropriate Hox ratios throughout the anterior-posterior (AP) axis. They further act in concert with more posteriorly expressed Hox genes to impose a functional hierarchy over more anterior ones, thus contributing to the phenomenon of 'posterior prevalence'. The implications of this study place miRNAs as critical developmental regulators in cases where they are spatially coexpressed with their targets and act to restrict broader transcriptional domains in ways that recapitulate and supplement existing interactions at other levels of gene expression.

Introduction

Segmentation of the body plan is a shared characteristic among several distant phyla of bilateral animals, including chordates, arthropods, and annelids (Davis and Patel, 1999). The genomically clustered Hox transcription factors set coordinates of the embryonic AP axis in many animals, and are determinants of individual segmental identities. Multiple Hox clusters have arisen in vertebrates via duplication of a single chordate ancestral cluster. The four clusters present in Mammals, HoxA through D, are located on different chromosomes, range in size between 100-200 KB, and contain 9 to 11 protein coding genes dispersed among 13 paralogous groups, all transcribed from the same strand of DNA. Hox genes are expressed in staggered and overlapping domains in all embryonic germ layers along the AP axis, with sharp anterior and diffuse posterior boundaries. The anterior limit of expression of a Hox mRNA, corresponds to higher transcript levels and the site of manifestation of phenotypic effects—typically anterior homeotic transformations—upon the gene's disruption, and is thus defined as its functional domain. The order of genes within a cluster determines their domain boundaries along the AP axis. In vertebrates, the onset of a gene's expression, initially during gastrulation, also coincides with its genomic position in the cluster. These conserved properties are defined as spatial and temporal colinearity, whereby genes at the 3' end of the Hox cluster as defined by the direction of transcription, are expressed earlier and more anteriorly, whereas more 5' genes appear later and further towards the tail (Krumlauf, 1994; Maconochie et al., 1996). The Hox network also produces a functional hierarchy of posterior genes over anterior ones with respect to segment specification, a phenomenon termed phenotypic suppression in insects, and posterior prevalence in vertebrates (Duboule and Morata, 1994; Krumlauf, 1993; Morata, 1993).

Hox genes are subject to regulation by microRNAs, a class of noncoding RNAs, which mature to form single-stranded 21mers associated with the RNA-Induced Silencing Complex (RISC), and mediate the repression of a third of mRNAs by recognition through direct base-pairing (Bartel, 2004). Embedded within vertebrate Hox clusters are several miRNAs falling into two families, which are transcribed in the same orientation as Hox genes and approximate Hox rules of expression with some differences (Mansfield et al., 2004; Yekta et al., 2004). Both miRNA families have high expression throughout the neural tube, and lower levels in the paraxial mesoderm with undefined anterior limits and broad posterior expression through the tail. The

miR-10 family is located between Hox4 and Hox5 paralogues, or within the intron of Hoxb4 in some organisms, and is conserved in genomic location and in sequence to arthropods. The vertebrate-specific miR-196, situated between Hox9 and Hox10, has been shown to repress neighbouring Hox genes. Absence from invertebrate chordates, and presence of multiple paralogues in vertebrate clusters, reveal an origin in the common ancestor of vertebrates predating the first cluster duplication. This was probably followed by subsequent loss of one copy of the gene in the HoxD cluster, yielding a final copy number of three miR-196 genes in modern tetrapods. There is also evidence for functional convergence of a Hox repressor at a similar location in flies, where a functional analogue of miR-196, miR-iab-4, arising from an orthologous genomic locus, and sharing no sequence homology, targets the downstream homeobox gene *Ubx* (Ronshaugen et al., 2005; Yekta et al., 2004). Expression of miR-196 is lower in the forelimb than in the hindlimb, where the miRNA acts as an inhibitor of *Hoxb8* and prevents its induction by ectopic retinoic acid (Hornstein et al., 2005).

By examining the genomic distribution of a more complete set of predicted targets, and suppressing the effects of each miRNA family in the avian embryo, we provide evidence for the contribution of miRNA-mediated interactions to achieving posterior prevalence. We propose that constraints produced by the requirement for functional hierarchy within the Hox family, have led to the recapitulation of similar modes of interaction at multiple levels of gene expression.

Methods

Prediction and genomic arrangement of the Hox targets of miR-196 and miR-10

Using the minimal miRNA targeting assumptions of (Lewis et al., 2005), we predicted targets of miR-10 and miR-196 in the human, mouse, chick, zebrafish and fugu Hox clusters without making any conservation requirements. A target contains at least one canonical 7-8mer seed match to miR-196 (8-mer seed match sequence: ACTACCTA) or miR-10 (8-mer seed match sequence: ACAGGGA) within its 3' untranslated region (UTR). The latter was defined according to preexisting database annotation in mammals, or as 2 KB of sequence 3' to the stop codon in teleosts where UTR annotation is insufficient. In the chick, where 3' UTR annotations are also poor, we relied on a combination of conservation to mammals, as well as EST availability to defined 3' ends of transcripts. In mammals, the 3' UTR of Hoxb1 was extended beyond existing

annotation. The presence of overlapping EST's and conservation above surrounding intergenic sequence propelled the inclusion of the downstream sequence. 3' UTR sequences used in the analysis are included in the supplementary material section. Exceptions to the requirement of a seed match for targeting, were genes in paralogous group 8 that contain non-classical complementary sites to miR-196, and of which the mammalian sequences have been experimentally validated (Yekta et al., 2004). These were therefore included in the analysis as targets of miR-196. Hox genes were divided into upstream and downstream groups referring to their genomic position relative to the miRNA locus in the direction of transcription. Downstream genes are transcriptional units 3' to the miRNA locus, whereas upstream genes are 5' transcriptional units. For example, the mammalian HoxA, HoxB and HoxC clusters each bear a single miR-196 locus upstream of the paralogous group 9 and in the intergenic space downstream of group 10 where the paralogue has not been lost. HoxD genes, which have no neighbouring miR-196 locus, were categorized as their corresponding paralogues on other clusters. As such, Hox1-9 were taken as downstream and Hox10-13 as upstream genes in the analysis.

Chick embryos

Fertilised eggs were obtained from SPAFAS and incubated at 37 °C. Embryos were staged according to (Hamburger and Hamilton, 1951).

Antagomir injections

Antagomirs are 3'-cholesteryl-conjugated ribonucleic acids with the following stabilizing backbone chemical modifications: 2'-methoxy groups throughout, and phosphorothioates substituting four 3'-terminal and two 5'-terminal phosphodiester linkages (Kruzfeldt et al., 2005). Oligonucleotides with sequences complementary to mature miRNAs, or corresponding mismatched controls were synthesized and RP-HPLC-purified (Dharmacon), dissolved in water, and injected at 0.56 µM in sterile phosphate buffer saline into the extra-embryonic omphalomesenteric (vitelline) veins of stage-14-18 chick embryos.

antagomiR-196a, 5'-CCCAACAACAUGAAACUACCUA; antagomiR-10b, 5'-
ACAAAUUCGGUUCUACAGGGUA; 5mm-antagomiR-196a, 5'-
CCCcACAcCAUGcAACUcCgUA; 5mm-antagomiR-10b, 5'-

ACAcAUUgGGUUGUACAuGuUA; antagomiR-223, 5'-GGGGUAAUUUGACAAACUGACA;
antagomiR-375, 5'-UAACGCGAGCCGAACGAACAAA.

Whole-mount *in situ* hybridisation

Embryos were collected and fixed in 4% paraformaldehyde overnight. Hybridisation with the probe for *Hoxb8* was performed as described (Nelson et al., 1996).

Skeletal preparation and staining

Bone and cartilage were differentially stained using alizarin red S (ossified material) and alcian blue (cartilage) according to a method modified from (McLeod, 1980). Ten-day old chick embryos were harvested in phosphate buffer saline, and extra-embryonic membrane and internal organs were removed. Embryos were dehydrated for one day in 10 volumes of 95% ethanol, one day in 10 volumes of acetone, and stained overnight in (15 mg/ml; 0.015% w/v) alcian blue, alizarin red S (5 mg/ml; 0.005% w/v), 5% glacial acetic acid and 60% ethanol. Excess stain was washed off three times for 30 minutes each in 95% ethanol, and tissues were cleared overnight or longer at 4 °C in 1% KOH. The skeletons were brought to 100% glycerol by passage through increasing glycerol:KOH series (25%, 50%, and 75%).

Analysis of skeletal phenotypes

Ten-day old skeletons were assessed for deviations from a wildtype axial body pattern of 14 cervical (C1-C14), 7 thoracic (T1-T7), 4 lumbar (L1-L4) and up to 19 sacrocaudal (S1-Cn) vertebrae. Phenotypic variation of individual vertebra was scored as either a transformation or a malformation. Unilateral, bilateral, partial or complete homeotic transformations were grouped together and counted equally in the scoring. Abnormalities that could not clearly be recognised as a transformation were counted as malformations. Deletions of caudal (coccygeal) vertebrae were also treated as malformations. P-values for the significance of the frequency of defects at each vertebral segment between experimental and control treatments were obtained by Fisher's Exact test.

Quantitative real-time PCR for chick Hox mRNAs

Embryos were harvested 30 hrs following antagomir injections at stage 15. Tails were dissected at the posterior level of the hind limb bud, and total RNA was isolated by trizol. Oligo dT-primed cDNAs were synthesized from 350 ng of DNase treated total RNA, and diluted 50 folds for quantitative real-time PCR. Six antagomiR-196a and eight antagomiR-223-treated tails were subjected to quantification for levels of various Hox mRNAs. Three replicate runs were done in parallel on each sample using SYBER Green DNA Master Mix on the Applied Biosystems ABI Prism 7700 Sequence Detection System, with the following primers:

5'-GAPDH, 5'-GACGTGCAGCAGGAACACTA; 3'-GAPDH, 5'-CTTGGACTTTGCCAGAGAGG; 5'-Hoxa3, 5'-GGCACGCGTAGGAAATACAT; 3'-Hoxa3, 5'-GCCTTCTTTCCCCCTATCTG; 5'-Hoxa5, 5'-ACACCCGGTATCAGACCTTG; 3'-Hoxa5, 5'-CTGAGAGGCAAAGAGCGTGT; 5'-Hoxa7, 5'-AGGAAAGCAACCTGCACAAC; 3'-Hoxa7, 5'-TCTGGTAGCGGGTGTAGGTC; 5'-Hoxa9, 5'-CTTACACCAAGCACCAGACG; 3'-Hoxa9, 5'-CTCTCGGTGAGGTTGAGGAG; 5'-Hoxa10, 5'-CAGACAGACAAGTTAAAATCTGGTT; 3'-Hoxa10, 5'-GAAATTAAGTTGGCTGTGAGC; 5'-Hoxa11, 5'-TCAGATTAGAGAGCTAGAAAGGGAAT; 3'-Hoxa11, 5'-TACTTGGCGGTCGGTCAG; 5'-Hoxb7, 5'-CGGCAAACCTACACCAGGTA; 3'-Hoxb7, 5'-TTCATGCGCCTGTTCTGG; 5'-Hoxb8, 5'-AACCTACAGCCGCTACCAGA; 3'-Hoxb8, 5'-GAGACCTCGATCCTCCGTTT; 5'-Hoxb9, 5'-AGTCTGGCCACTTCGTGTCT; 3'-Hoxb9, 5'-GAAAAAGCGATGCCCTTACA; 5'-Hoxc8, 5'-GAACCTCCAGCATCTCCAAC; 3'-Hoxc8, 5'-CTCGGCAGAGCTTCATATCC.

Relative RNA amounts were calculated using the $\Delta\Delta C_t$ method. The statistical significance of the difference between ΔC_t values (normalised to GAPDH, p-value < 0.05) of antagomiR-196a and antagomiR-223 treated tails was calculated using the Mann-Whitney U test. Mean values of the fold change in expression of six antagomiR-196a-treated tails over the average of eight antagomiR-223-treated tails for each gene primer set was measured and reported.

Results

A number of Hox genes have been previously identified and validated as targets of the hox cluster-embedded miR-196. The mammalian *Hoxb8* is an atypical target as it has a single perfectly complementary site with the exception of a G:U wobble pair at position 5 from the 5' end of miR-196. It has been shown to be a target of miRNA-directed cleavage in the mouse. Fragment of 3' UTRs of *Hoxb8*, *Hoxc8*, *Hoxd8* and *Hoxa7* containing target sites also mediate the repression of reporters in HeLa cells (Yekta et al., 2004). Much has been learned concerning the prediction of microRNA targets in the last few years. We revisited the question of Hox targets and expanded previously known targets of miR-196 and miR-10 in the human, mouse, chick, zebrafish and fugu Hox clusters. Hox 3' UTRs had at least one and up to six canonical 7-8mer seed matches to miR-196 or miR-10 (Supplementary Materials). Although conservation was not required for predicting individual targets, the majority of the seed matches were in fact conserved in multiple genomes. In three cases where there was a discrepancy between the human and mouse targets, the history of the seed match was assessed by an examination of multiple mammalian genome alignments. All three cases involved a single site that was not conserved in both human and mouse, but was present in at least one other mammal. *Hoxa4* seems to have gained a 7mer-1A seed match to miR-196 in the rodent lineage. The site is absent in primates and in other basal mammals. *Hoxb13* has gained a 7mer-1A seed match to miR-196 in the primate lineage but is absent in other mammals. *Hoxd1* has retained an 8mer match to miR-10 in the mouse, but has lost it in rat and in the primate lineage. The basal opossum does not have the site but six other mammals basal to both primates and rodents do have a 7mer-m8 site to miR-10. In a number of cases including the murine *Hoxa4* and *Hoxa7*, where there is evidence of alternative polyadenylation, miRNA sites appear in the longer isoform, and may contribute to isoform-specific regulation.

On average miR-196 was predicted to target 27% of Hox genes in the five vertebrates genomes examined, implying that potentially there is a considerable amount of posttranscriptional regulation internal to Hox clusters governed by this miRNA. These predicted target genes however, were not evenly distributed throughout the clusters. Regardless of conservation, a significant majority of target 3' UTRs belonged to genes that lie in paralogous groups transcriptionally downstream of, or with expression boundaries anterior to miR-196 loci based

on presumed homeotic patterns. For example, in humans, there were ten downstream targets of miR-196, and only a single upstream one. Within the downstream region, more than half of the predicted targets are in the immediate 3' vicinity of the miRNA locus, that is, within the central Hox genes of paralogous groups 5-9. Likewise, the higher fraction of downstream Hox genes predicted to be targets of miR-196 is higher than the fraction of upstream Hox genes, with a vertebrate average of 38% vs. 4% (Table 1a).

The trend in non-random distribution of predicted target genes was also significant for the more ancient Hox miRNA family, miR-10. The genomic position of miR-10 in the clusters (between Hox4 and Hox5) dictates that there were fewer 3' Hox genes available for targeting. Although this miRNA appears to target a smaller number of Hox genes than does miR-196 (16% in vertebrates), the genomic arrangement of targets follows the same one-sided skew observed for miR-196. Here again, a higher fraction of the downstream genes compared with upstream ones were predicted as targets of miR-10 (a vertebrate average of 37% vs. 8%; Table 1b).

In *Drosophila*, the analogue of miR-196, miR-iab-4 gives rise to miRNAs from both arms of the precursor hairpin, miR-iab-4-5p and miR-iab-4-3p, both represented among sequences cloned from cDNA libraries. miR-iab4-5p has target sites in the 3' UTRs of downstream genes *Antp* and *Ubx*, and miR-iab-4-3p has sites in the 3' UTRs of downstream genes *Scr*, *Antp* and *Ubx* (Data not shown; (Ronshaugen et al., 2005)) It appears that vertebrates and fly pathways involving the targeting of downstream Hox genes by a miR-iab-4/miR-196 have converged independently. This trend is not observed for the *Drosophila* miR-10 and not yet examined in other non-chordate animals. As such, it is not clear whether regulation of Hox genes by this miRNA arose independently in chordates, or whether it was shared in the last common ancestor of arthropods and chordates and subsequently lost in *Drosophila*.

Based on prediction made from sequences of fish, birds and mammals, separated by over 400 million years of evolutionary distance, both miR-196 and miR-10 appear to have specialised in regulating Hox messages, specifically those more downstream in the Hox cluster. Table 2 lists predicted Hox targets of miR-196 and miR-10 in chick. To validate that these are indeed true targets *in vivo*, we employed the previously established technology of antagomirs (Krutzfeldt et

al., 2005) to block miRNA function in chick embryos, and examine the effects of miRNA inhibition on predicted Hox target expression, focusing specifically on miR-196.

Antagomirs are chemically modified oligonucleotides resistant to enzymatic degradation and designed to base-pair to and sequester RISC-bound miRNAs *in vivo*. We showed they can be delivered systemically to whole embryos by injection into the circulatory system where a 3'-conjugated cholesteryl moiety facilitates cellular uptake from the bloodstream. Antagomirs complementary in sequence to miR-196a (antagomiR-196a), and the control miR-223 (antagomiR-223), were injected into stage-15 embryos. Tails were dissected 30 hours following injection, and normalised levels of several Hox mRNAs were quantified by real-time PCR (Fig. 1a). Predicted target Hox genes were consistently upregulated upon antagomiR-196a treatment, with statistically significant increase in levels of *Hoxb7*, *Hoxb9* and *Hoxc8*. No significant change was observed in levels of non-targets *Hoxa3*, *Hoxa10* and *Hoxa11*. Increases in target mRNA expression levels in adult mice subjected to antagomir treatment have previously been reported (Kruzfeldt et al., 2005). This increase is presumably a result of the derepression of RISC-bound mRNAs due to competitive binding of antagomirs, and/or the escape of nascent mRNAs from miRNA-mediated destabilization due to lowered availability of functional miRNAs. AntagomiR-196a treatment also led to an increase in a Hox target mRNA, as determined by whole-mount *in situ* hybridisation: a reproducible ectopic expansion by 1 to 2 somites in the posterior domain of *Hoxb8* was observed in the paraxial mesoderm in 3-day old embryos 30 hours after injection (n = 3 of 7, Fig. 1b).

Alterations in Hox gene expression affect the differentiation of somites, or the relative growth and differentiation of somite derivatives, apparent in the form of patterning defects of the axial skeleton, often homeotic transformations of a vertebral segment (Krumlauf, 1994). Vertebrae develop from migrating sclerotome cells, one of three major somitic components, with high dependence on signals from the notochord and surrounding tissues. Antagomirs with sequences complementary to miR-196a and miR-10b (antagomiR-10b) were injected into stage 15-18 embryos. As controls for non-specific effects of the reagent, two unrelated miRNAs not been implicated in early developmental roles, miR-223, which is specifically expressed in granulocytes, and miR-375, which is specifically expressed in the pituitary gland and in pancreatic islet cells, and mismatched versions of miR-196a and miR-10b were also targeted by

complementary antagomirs. Phenotypic consequences on the developing embryonic skeleton were examined seven days after injection in ten-day old embryos, when most cartilage has formed, ossification of ribs has begun to take place, and individual vertebrae have acquired much of their distinct features.

The chick embryonic axial skeleton exhibits naturally occurring variation—most commonly in the lumbosacral region—to an extent that differs substantially from one flock to another. Wildtype variation from the most common pattern of C14/T7/L4/SC ~19 is depicted (Fig. 3). Complete or partial transformations of the first lumbar vertebra to a thoracic one bearing an ectopic eighth rib were recurrent (21%), as were anterior shifts in the identities of L2-S1 vertebrae. Also observed, were differences in the total number of lumbar vertebrae. Overall, a background rate of 33% of wildtype skeletons differed from the standard pattern, with an average of two altered vertebral segments per individual embryo (Table 3).

The array of defects observed in antagomir-treated embryos was scored on a per-segment basis, with each affected vertebra counting as defective—either with a malformation or a homeotic transformation—regardless of the nature or severity of the abnormality. Typical phenotypes are depicted in (Fig. 2). Cervical vertebra 1 (C1) deformities consisted of unfused neural arches at the dorsal midline. Posteriorising homeotic transformations were observed at C2, with the appearance of C3- and C4-specific foramina, while the disappearance of the same structure at C4 was interpreted as a C4 to C5 posterior transformation. Posteriorising transformations of C14 to a rib-bearing T1 were observed. Typical malformations included unfused vertebrae at the dorsal midline and absent spinous processes; unfused elements in the dorsal lamina (common in the upper ribcage); compressed vertebral bodies and compacted vertebrae; adjacent vertebral discs that were entirely fused together; asymmetric or staggered vertebral discs, at times unilaterally deleted; and, deletions of the most caudal vertebrae. Rib defects consisted of fusions, absence of sternal fusion at T3; abnormal rib morphology such as shortened, thickened, bifurcated or twisted ribs; and abnormal or asymmetric rib articulation at the transverse processes.

Summaries of frequency of occurrence and individual variations of skeletal phenotypes are listed (Table 3). 62% of antagomiR-196a-treated and 76% of antagomiR-10b-treated bird embryos displayed one or more defects, and on average, six and nine vertebral segments respectively,

were abnormal in affected individuals. Treatment of embryos with antagomirs targeting miR-375 and miR-223 did not lead to significant skeletal defects. These embryos resembled wildtype cases, with a non-significant trend towards additional defects in the anterior cervical and upper thoracic regions. Similar results were observed for the 5mm-antagomiR-196a control. For these three controls, 21-27% of antagomir-treated animals had one or more defects, with an expressivity of one to two segments per affected individual, suggesting that while there might be nonspecific toxic effects of antagomir-treatment on skeletal patterning, the phenotypes observed for antagomiR-10b and antagomiR-196a can be attributed to sequence-specific events presumably involving the miRNAs. The 5mm-antagomiR-10b produced a higher-than-background fraction of defects, 48%, and an average number of four vertebral segments were affected. This expressivity, still less than half that of treatment with antagomiR-10b, suggested that hybridisation to miR-10b might not have been entirely abolished. Alternatively, off-target toxicity might account for some of the observed defects.

A wide range of skeletal defects variable in aspect and distribution along the A-P axis was observed upon blocking miR-196 and miR-10, implying severe and pleiotropic phenotypes, despite the late-stage introduction of the reagent. The distribution of phenotypes along the skeletal axis showed that several vertebral segments were affected with a statistically significant frequency of defects in antagomir-196a and antagomir-10b—treated skeletons compared to the wildtype, antagomiR-223 and antagomiR-375—treated controls (Fig. 3). Defects in avian embryos treated with antagomiR-196a occurred throughout the skeletal axis with significant effects observed posterior to C14, mostly concentrated in the upper thoracic region. Caudal deletions occurred with moderate statistical significance. Tail defects were often visible in embryos harvested early after injection linking the phenotype to early misregulation events. C14 transformation and the corresponding increase in rib number were also observed at a low rate in miR-196a knockdowns. Outgrowth of sclerotome cells has begun by stage 15, so we asked if more transformations would result from injection at an earlier developmental stage when fewer somites have formed. We injected stage-10-13 embryos in the left presomitic mesoderm (psm) cavity, expecting to expose a more restricted area to higher levels of antagomirs. Correspondingly, defects were localised near the site of injections and were more severe. C14 to T1 transformation was observed in 51% (n = 35) of embryos injected with antagomiR-196a

while 24% (n = 49) of embryos injected with 5mm-antagomiR-196a displayed the same phenotype (McGlenn and Tabin, data not shown).

Suppression of miR-10b led to a comparable but more extensive phenotype with a surprising number of posterior defects. Significant numbers of C2 to C3 posterior transformations were detected, as were a high number of caudal deletions.

Discussion

The broad phenotypes of the suppression of miR-196 and miR-10 point to possible action throughout the paraxial mesoderm and inducing tissues to set appropriate levels of many Hox and non-Hox target mRNAs. The transcriptional domains of both miRNAs approximate homeotic expression patterns (Mansfield et al., 2004); data not shown). The anterior limit of miR-196 expression is expected to be slightly more posterior to *Hoxb9*, which has an anterior limit in the paraxial mesoderm up to prevertebra-3 (pv3) in E9.5 mice that shifts caudally in the upper thoracic region by E12.5 (Chen and Capecchi, 1997). The anterior limit of miR-10 is expected to be caudal or equivalent to that of *Hoxb4*, which has a boundary in the paraxial mesoderm at pv2 in E10.5 mice (Brend et al., 2003). In fact, there is considerable overlap between the distribution of defects, the presumed domains of the miRNAs, and expression of their Hox targets. The skeletal phenotypes produced by treatment of embryos with antagomirs, are a composite of general toxicity, off-target events, and the intended sequence-specific effects on miRNAs. It is impossible given the data here to attribute defects to specific Hox misregulation events, but rather, it seems likely that in the absence of normal levels of either miR-10 or miR-196, multiple Hox genes are misexpressed, and contribute to the observed range of defects.

Blocking Hox miRNAs produced anterior phenotypes that lie within expected boundaries of target Hox genes. The most anteriorly expressed targets of miR-196, are *Hoxb1* and *Hoxa5*, both with expression boundaries anterior to the cervical to thoracic transition. *Hoxb1* has an anterior boundary in the hindbrain, while *Hoxa5* is expressed up to the level of somite 8 in the mouse (Aubin et al., 1998; Burke et al., 1995). The defects of miR-196 knockdown were within the posterior regions of the expression domains of these anterior targets, suggesting that the miRNA may be responsible for regulating posterior boundaries. The thoracic defects however do overlap

with functional domains of Hox paralogues in groups 7 through 9 defined through loss-of-function studies (Chen and Capecchi, 1997; Chen et al., 1998; van den Akker et al., 2001). It is therefore possible that the realm of miRNA control extends to broader, or a subset of cells within expression domains of some target genes, rather than being limited to regulation within posterior ends of target expression. Lastly, the lumbar through caudal defects overlay posterior limits of most targeted Hox genes, again coinciding with miRNA regulation of posterior expression boundaries of Hox genes.

The anterior targets of miR-10, *Hoxb1*, *Hoxa3*, and *Hoxb3*, have anterior expression boundaries in the hindbrain and in the occipital somites (Burke et al., 1995; Manzanares et al., 2001; Rossel and Capecchi, 1999). Loss-of-function of group 3 genes leads to a number of atlas and axis defects (Chisaka and Capecchi, 1991; Manley and Capecchi, 1997), overlaying spatially with the observed C2 (Axis) to C3 posterior transformations in miR-10 knockdowns. Other defects of miR-10 suppression seem to map to posterior domains of its targets, suggesting that misexpression of the most 3' Hox genes in posterior regions can lead to considerable patterning defects.

Transformations at C14 were observed for both miRNAs, with significance of the phenotype upon early introduction of antagomiR-196. Shift to a posterior identity at this boundary position and formation of an ectopic rib is common in mutants of the central Hox region, examples include loss-of-function of paralogues in group 4 (Horan et al., 1995a; Horan et al., 1995b; Horan et al., 1994) and 5 (Jeannotte et al., 1993), and gain-of-function of *Hoxb7* and *Hoxb8* (Charite et al., 1995; McLain et al., 1992). The transformation is also observed upon disruption of global regulators of Hox expression; including treatment with retinoic acid (Kessel and Gruss, 1991), loss-of-function of the retinoic acid metabolizing enzyme, CYP26A1 (Abu-Abed et al., 2001), and mutations in polycomb repressors of Hox genes, *mel-18*, *rae28*, and *bmi-1* (Akasaka et al., 1996; McLain et al., 1992; Takihara et al., 1997). Changes in Hox gene expression are observed in several cancers. Human children with embryonal cancers have a 125-fold increased incidence of cervical ribs, and children born with a cervical rib have a 120-fold increased likelihood of early childhood cancer (Galis, 1999). The recurrence of this phenotype implies that boundary segments are most susceptible to transformations (Chen et al., 1998), and perturbations in Hox ratios cause shifts in identity at regional boundaries, and the corresponding expansion of

the thorax. Statistically significant detection of this phenotype upon early miR-196 suppression provides further evidence for the role of the miRNA in the maintenance of appropriate Hox ratios. The presence of miRNAs in the cluster with other Hox genes results in a shared local DNA environment, which may lead to similar responses to non-specific perturbations that affect global transcription, and allow for overall maintenance of constant relative Hox ratios. This shared response may be a driving force in maintaining gene linkage within clusters.

The conserved trend in the genomic distribution of the miRNA-target Hox genes within the clusters can be described as a statistically significant enrichment of miRNA seed matches in genes transcriptionally downstream and/or anterior in expression to the miRNA loci, relative to a depletion of sites in posterior and upstream genes (Table 1). The implied net effect is that where expression overlaps, downstream genes are more likely to be repressed by the miRNA than upstream ones. The genomic distribution of Hox targets and Hox rules of expression lead to the prediction that the Hox miRNAs contribute to restricting the posterior limits of Hox genes within domains of transcriptional competence. In fact upon suppression of miR-196, the expression domain of *Hoxb8* increases towards the tail in the paraxial mesoderm. The high number of defects in posterior regions, that is, posterior to the rostral limit of target Hox genes, further implicates the miRNAs in restricting posterior domains of the Hox targets. The genomic distribution of miR-196 targets also suggests that within the expression domain of posterior Hox genes (Hox10-13), the miRNA acts as a repressor of transcriptionally active downstream genes, but not of the upstream Hox10-13, which have evolved to escape repression (Table 1a), thus contributing to the expected functional dominance of the posterior group over the anterior one. Further support for this idea comes from the statistically significant increase in mRNA levels of three downstream Hox targets of miR-196 in the tail upon knockdown, with no observed upregulation of either upstream Hox gene examined.

As new segments or units of serial repetition arise in evolution, or form during ontogeny, it is expected that molecular events leading to final morphologies base themselves on preexisting or default pathways, and evolve novel routes that override and alter and coopt the default pathway. The posterior prevalence model describes a functional hierarchy that is a remnant of an ancestral property of the Hox cluster, observed even in the absence of segmentation, or absolute spatial and temporal colinearity, as is the case in the highly divergent Nematode Hox cluster (Burglin

and Ruvkun, 1993). The phenomenon was originally postulated based on phenotypic observations made in *Drosophila* larvae with mutations at the *extra sex combs (esc)* locus that inactivate polycomb group repressors and cause general derepression of Hox expression. The resulting segmental pattern in *esc* mutants was that governed by the most posterior acting Hox gene, *Abd-B*, such that the head, thoracic and abdominal segments, morphed into the most posterior abdominal segment, A8. Mutant *esc* larvae that further lacked *Abd-B*, developed with a reiteration of A4 segments, typically specified by *abd-A*, the next more posterior gene. Mutant *esc* larvae with deletion of all abdominal Hox genes, *Ubx*, *abd-A*, and *Abd-B*, developed with reiterations of thoracic segments normally specified by *Scr* and *Antp*. When these two were eliminated in addition to the three abdominal genes, *esc* larvae had cephalic segments throughout. This study defined a hierarchy of homeotic gene function, where posterior and 5' genes were epistatic to anterior and 3' genes (Struhl, 1983). Further experiments showed that phenotypic suppression is not primarily due to transcriptional cross-regulation. Ubiquitous expression of Hox genes under promoters known to be transcriptionally irrepressible, led to transformations only in regions anterior to the functional domain of the gene. For example, the thoracic *Antp*, when expressed ubiquitously can suppress Hox genes of the head and cause posterior transformations of head segments into the thorax, but does not affect the abdomen, where *Antp* is suppressed by abdominal genes such as *Ubx* (Gibson and Gehring, 1988; Morata, 1993). Similar observations were made in vertebrates, where for instance the introduction of a *Hoxd4* transgene under the transcriptional control of the promoter of *Hoxa1* led to an expected rostral shift in the anterior boundary of *Hoxd4* expression, a gene that is not a target of miR-10 or miR-196. The transgenic embryos exhibited posterior transformations of the occipital bones at the base of the skull towards structures that resemble characteristics of the segmented vertebral column, in particular of the first two cervical vertebrae. However, while levels of the transgene were also higher in the endogenous *Hoxd4* and more posterior expression domains, the phenotypes were limited to the ectopic anterior boundary (Lufkin et al., 1992). The posterior prevalence model explains the general trends of homeotic phenotypes with loss-of-function often leading to anterior transformation at rostral boundaries of expression; in the absence of a Hox gene, more anterior acting genes that are typically suppressed are now permitted to function. The model also explains why gain-of-function or ectopic expression of a Hox gene generally causes posterior transformations in regions anterior to the endogenous domain, where the ectopic

expression can suppress resident homeotic genes. These tendencies generally hold true for the fly, but not always for vertebrates see for example (Pollock et al., 1995), where deviations from this general rule, and defects other than homeotic transformations appear. In general vertebrate systems appear to be more sensitive to quantitative differences in Hox gene expression.

The molecular mechanisms for the functional hierarchy, largely unsolved, are attributed directly to various properties of the Hox genes (Morata, 1993) and their downstream targets. Within a functional domain, a gene's dominance over downstream ones is asserted foremost quantitatively due to higher levels of transcription. Posterior acting Hox proteins are also more efficient than anterior ones at exerting their downstream function, likely in competing for overlapping DNA target binding sites, or for binding to interaction partners, see for example (Williams et al., 2006). The hierarchy also appears although without observable phenotypic consequence at the level of transcriptional cross-regulation, whereby posterior genes direct the repression of anterior ones (Morata, 1993). Derepression of target genes in their posterior domains is expected to be phenotypically suppressed, and thus irrelevant due to dominance of more posterior Hox genes. We find this misregulation to be of functional consequence however, given the abundance of defects in posterior axial regions in our study. These are attributed to derepressed target Hox genes that override endogeneous suppression by posterior Hox genes. Hence, to a certain extent, the dominance of genes that have evolved to escape Hox miRNA control, results from a direct contribution of miRNA-mediated repression of target Hox genes, and in the absence of suppression by Hox miRNAs, altered levels of multiple target Hox proteins disrupt normal skeletal development. The contribution of miRNA-mediated repression to the posterior prevalence model, adds an additional layer of regulatory interactions at the posttranscriptional level, demonstrating that the ancient functional requirement for hierarchy among Hox genes, dictates that a general mode of regulation be reiterated by multiple layers of gene interactions, at all levels of gene expression. The additive effects of layering modes of regulation are likely to provide a stabilizing effect that buffers against perturbations in expression.

Genes involved in essential and ancient developmental processes, are limited in variation, and maintained by stabilizing selection as many subsequent integrated processes depend on them, thus constituting highly constrained systems. The Hox:miRNA interactions unraveled thus far in insect and vertebrate systems, demonstrate a layer of gene interactions—within which

hierarchies and individual contributions remain to be deciphered—that confers buffering capacity or robustness to a genetic network, and has evolved independently toward meeting the demands of an overall constraint.

Table 1. Genomic Distribution of Hox genes targeted by of miR-196 and miR-10

Predicted targets contain within their 3' UTR one or more canonical 7-8mer seed match. Hox genes targeted by miR-196 (a), and miR-10 (b), were categorised according to their genomic location in the Hox clusters relative to the miRNA locus. In the case of miR-196, Hox1-9 were grouped together as downstream, and Hox10-13 were considered upstream, regardless of which cluster they belonged to. Similarly Hox1-4 were downstream of miR-10, whereas Hox5-13 were upstream. P-values for the likelihood of the observed genomic distributions in each species were obtained by the Fisher's Exact Test.

a) miR-196 targets in the Hox cluster

Species (% Hox genes predicted as targets)	Downstream targets/All downstream genes (Hox1-9)	Upstream targets/All upstream genes (Hox10-13)	P-value ($P_{total} = 2.6 \times 10^{-8}$)
Fugu (27)	12/32	1/16	0.020
Zebrafish (31)	14/32	1/17	0.005
Chick (25)	8/22	0/10	0.030
Human (28)	10/27	1/12	0.068
Mouse (26)	10/27	0/12	0.013

b) miR-10 targets in the Hox cluster

Species (% Hox genes predicted as targets)	Downstream targets/All downstream genes (Hox1-4)	Upstream targets/All upstream genes (Hox5-13)	P-value ($P_{total} = 2.4 \times 10^{-9}$)
Fugu (29)	9/15	5/33	0.003
Zebrafish (16)	5/14	3/35	0.033
Chick (9)	3/10	0/22	0.024
Human (13)	3/12	2/27	0.159
Mouse (15)	4/12	2/27	0.060

Table 2. Avian Hox genes targeted by miR-196 and miR-10

Hox genes with 7-8mer seed matches to miR-196 (a) or miR-10 (b) in their 3' UTRs are listed.

a) Targets of miR-196 in avian Hox clusters

Hox target	8-mer	7-mer 1A	7-mer m8
Hoxa5	1		
Hoxa7	3	1	1
Hoxa9	1		1
Hoxb1			1
Hoxb7*	1	2	
Hoxb8	1		
Hoxb9			1
Hoxc8			1

b) Targets of miR-10 in avian Hox clusters

Hox target	8-mer	7-mer 1A	7-mer m8
Hoxa3	1		
Hoxb1			1
Hoxb3	1		

* Hoxb4 and Hoxb7 homologues in the chick genome do not appear as linked to the remainder of HoxB cluster genes on chromosome 27. Instead, they map to unknown chromosomal fragments. Their sequence homology and Hox-like expression pattern may suggest that their genomic placement is a result of incomplete and inaccurate assembly of the chick genome, and thus they were included in the analysis in a manner similar to other genomes.

Table 3. Penetrance and Expressivity of skeletal defects

Stage-15-18 embryos were treated with various antagomirs and vertebral defects were scored in day-ten skeletons. * P-value < 0.001 determined by the Chi-squared test of significance of difference from the expected wildtype frequency.

Condition	Total	Defective	Penetrance	Average segments per affected individual
Wild type	73	24	33%	2
antagomiR-196a	61	38	62% *	6
antagomiR-10b	46	35	76% *	9
5mm-antagomiR-196a	59	16	27%	2
5mm-antagomiR-10b	27	13	48%	4
antagomiR-223	47	10	21%	1
antagomiR-375	52	13	25%	2

Fig 1. Injection of an antagomir complementary in sequence to miR-196a caused upregulation of several target Hox mRNAs. (a) Quantitative real-time PCR assay on several Hox mRNAs containing one or more miR-196 target sites (name in red) or no sites (name in black) in their 3' UTR. Each box represents the distribution of Δ Ct values normalised to GAPDH for antagomiR-223-treated tails (grey box, n = 8) and antagomiR-196a-treated tails (red box, n = 6). Bars define 10th and 90th percentiles; box spans 25th and 75th percentiles; and the line indicates median Δ Ct. Means for fold increase in mRNA levels of antagomiR-196a-treated tails over antagomiR-223-treated tails and statistical significance (* p-value < 0.05, ** p-value < 0.001, Mann-Whitney-U test) are indicated above box plots for each gene primer set. (b) Whole-mount *in situ* hybridisation of *Hoxb8*. Left, wildtype, right, stage-24 embryo two days after antagomiR-196a injection. Arrow points to ectopic expression of *Hoxb8*.

Fig 2. Typical skeletal defects induced by antagomir treatment. (a) Wildtype day ten skeleton, b, c,e,f,h) antagomiR-10b-; and d,g) antagomiR-196a—treated skeletons. (b) C2 to C3 posterior transformation, (c) C14 to T1 posterior transformation, d) T1-T2 fusion defects, (e) abnormal rib morphology, (f) lumbosacral fusion of multiple vertebrae, (g) unilaterally deleted and staggered sacral vertebrae, (h) caudal deletion. Arrows point to defective areas.

Fig 3. Summary of the distribution of defects along the chick skeletal axis. Defective vertebral segments are shown as filled horizontal bars either malformations (red) or transformations (black) along the horizontal axis, which corresponds to chick vertebral segments. Each horizontal line is representative of an individual embryo, and stacked lines represent the entire experimental group. The asterisk represents statistical significance (p-value < 0.05) of the frequency of defects at a given position paired with miR-223, miR-375 and wildtype controls, calculated by the Fisher's Exact Test.

Fig. 4. Summary of Genomic Data

Figure 1

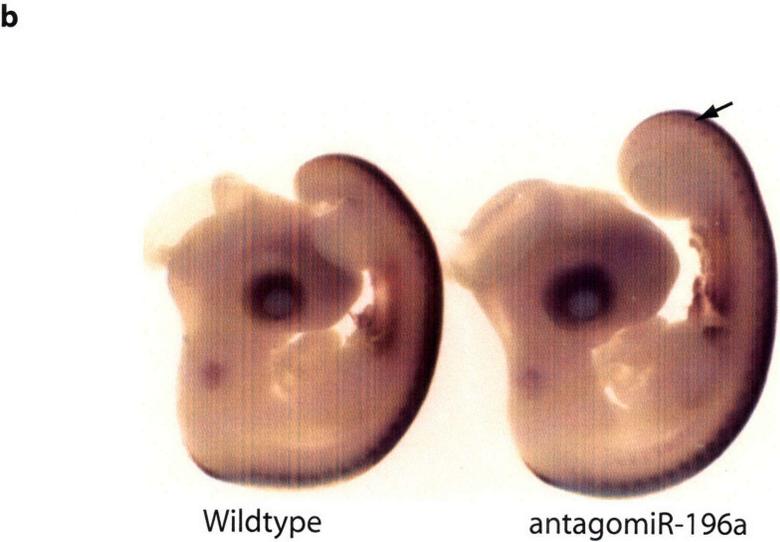
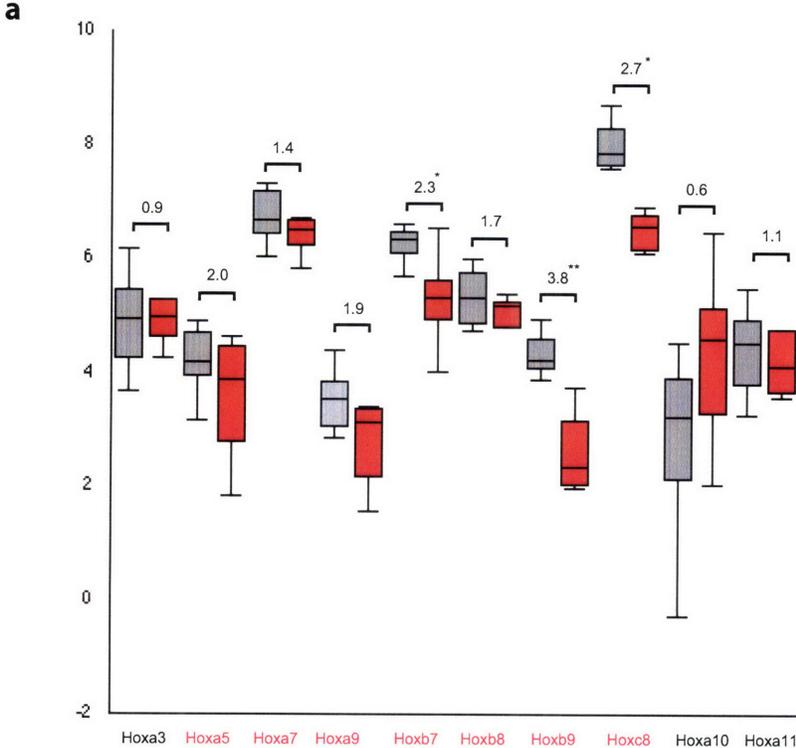


Figure 2

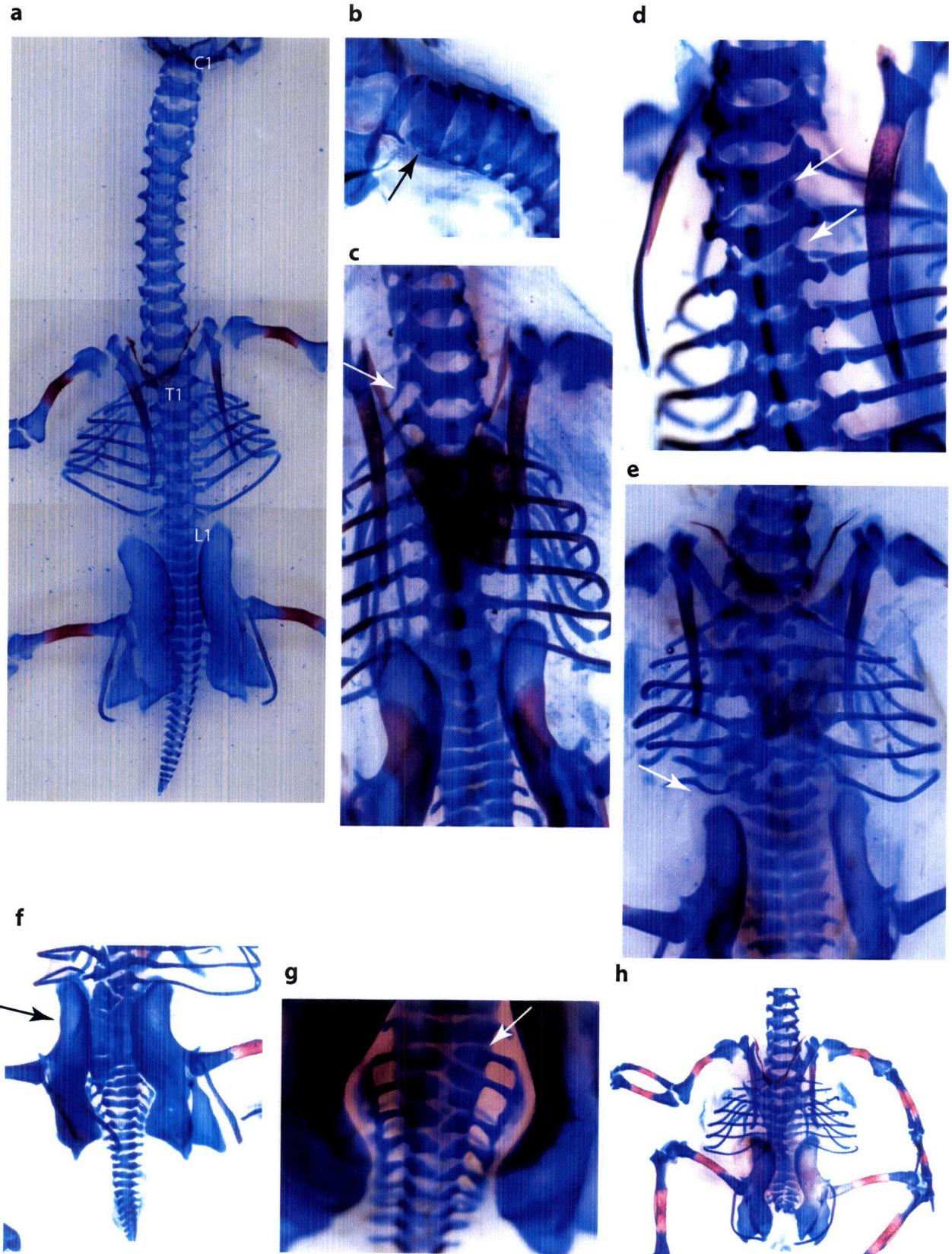
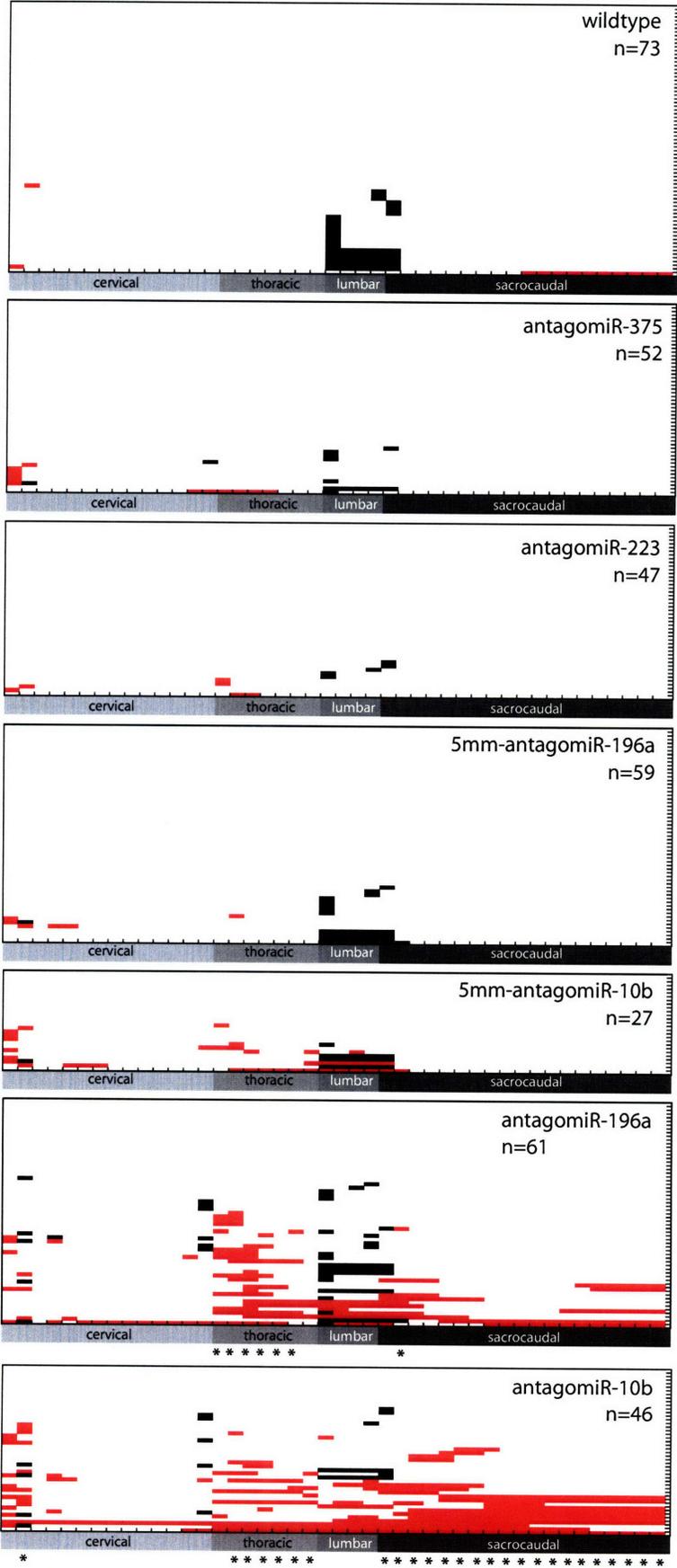
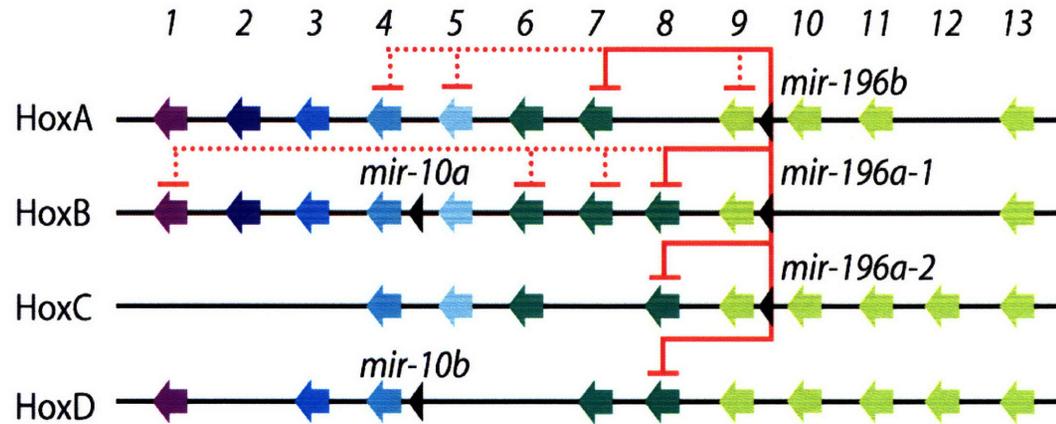


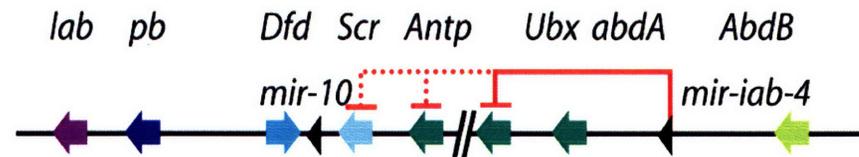
Figure 3



Mouse

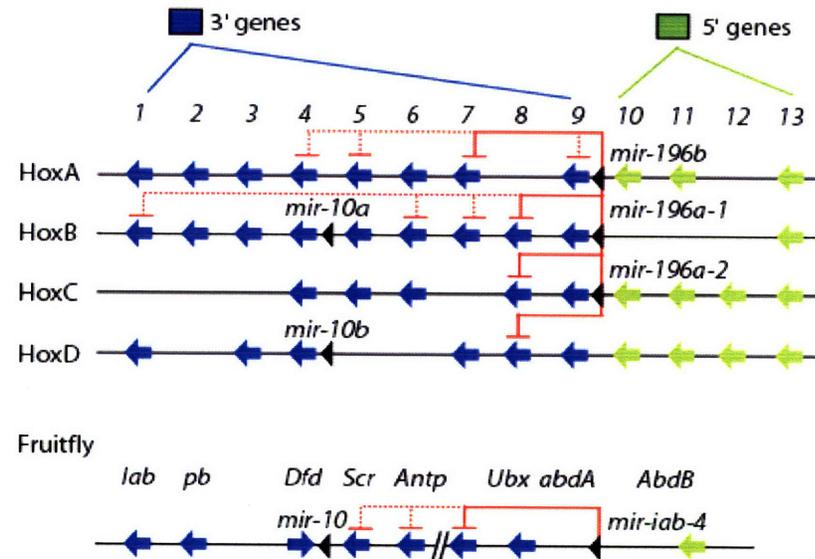
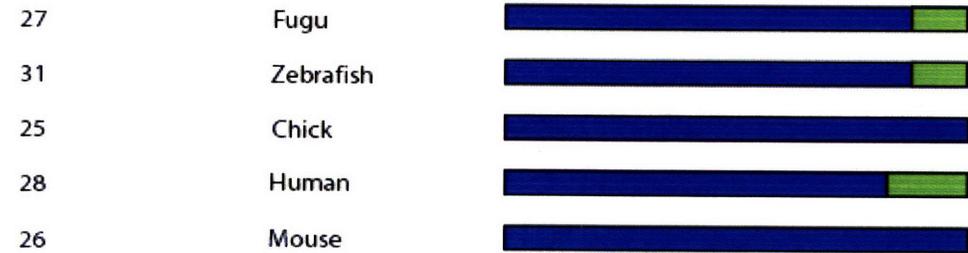


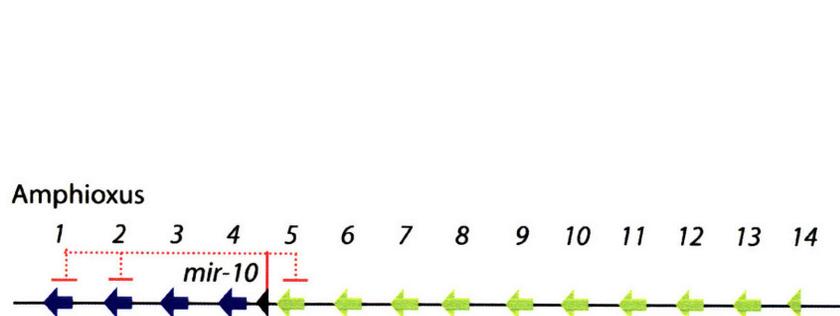
Fruitfly



	3' genes	5' genes	p-value
Fugu	12/32	1/16	0.020
Zebrafish	14/32	1/17	0.005
Chick	8/22	0/10	0.030
Human	10/27	1/12	0.068
Mouse	10/27	0/12	0.013

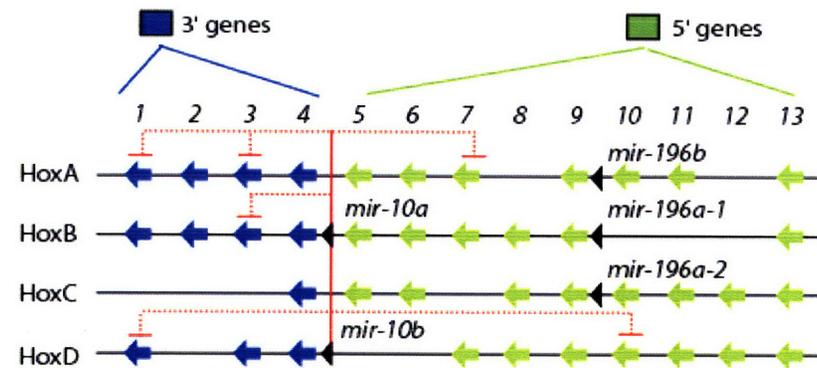
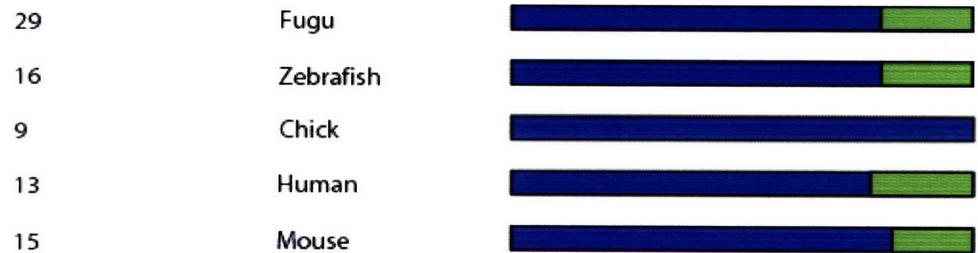
% Hox genes predicted as targets





	3' genes	5' genes	<i>p</i> -value
Fugu	9/15	5/33	0.003
Zebrafish	5/14	3/35	0.033
Chick	3/10	0/22	0.024
Human	3/12	2/27	0.159
Mouse	4/12	2/27	0.060

% Hox genes predicted as targets



ACKNOWLEDGEMENTS

The experimental work with Chick embryos presented in this chapter was largely performed in the Tabin laboratory. I thank all the members of the Tabin and Cepko labs for their valuable assistance and technical expertise. This work started with Jennifer Mansfield, and is an on-going collaboration with Eddy McGlinn. I would like to thank Cliff Tabin and Eddy McGlinn for their comments on this Chapter.

REFERENCES

- Abu-Abed, S., Dolle, P., Metzger, D., Beckett, B., Chambon, P., and Petkovich, M. (2001). The retinoic acid-metabolizing enzyme, CYP26A1, is essential for normal hindbrain patterning, vertebral identity, and development of posterior structures. *Genes Dev* *15*, 226-240.
- Akasaka, T., Kanno, M., Balling, R., Mieza, M. A., Taniguchi, M., and Koseki, H. (1996). A role for mel-18, a Polycomb group-related vertebrate gene, during theanteroposterior specification of the axial skeleton. *Development* *122*, 1513-1522.
- Aubin, J., Lemieux, M., Tremblay, M., Behringer, R. R., and Jeannotte, L. (1998). Transcriptional interferences at the Hoxa4/Hoxa5 locus: importance of correct Hoxa5 expression for the proper specification of the axial skeleton. *Dev Dyn* *212*, 141-156.
- Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* *116*, 281-297.
- Brend, T., Gilthorpe, J., Summerbell, D., and Rigby, P. W. (2003). Multiple levels of transcriptional and post-transcriptional regulation are required to define the domain of Hoxb4 expression. *Development* *130*, 2717-2728.
- Burglin, T. R., and Ruvkun, G. (1993). The *Caenorhabditis elegans* homeobox gene cluster. *Curr Opin Genet Dev* *3*, 615-620.
- Burke, A. C., Nelson, C. E., Morgan, B. A., and Tabin, C. (1995). Hox genes and the evolution of vertebrate axial morphology. *Development* *121*, 333-346.
- Charite, J., de Graaff, W., and Deschamps, J. (1995). Specification of multiple vertebral identities by ectopically expressed Hoxb-8. *Dev Dyn* *204*, 13-21.
- Chen, F., and Capecchi, M. R. (1997). Targeted mutations in hoxa-9 and hoxb-9 reveal synergistic interactions. *Dev Biol* *181*, 186-196.
- Chen, F., Greer, J., and Capecchi, M. R. (1998). Analysis of Hoxa7/Hoxb7 mutants suggests periodicity in the generation of the different sets of vertebrae. *Mech Dev* *77*, 49-57.
- Chisaka, O., and Capecchi, M. R. (1991). Regionally restricted developmental defects resulting from targeted disruption of the mouse homeobox gene hox-1.5. *Nature* *350*, 473-479.
- Davis, G. K., and Patel, N. H. (1999). The origin and evolution of segmentation. *Trends Cell Biol* *9*, M68-72.
- Duboule, D., and Morata, G. (1994). Colinearity and functional hierarchy among genes of the homeotic complexes. *Trends Genet* *10*, 358-364.
- Galis, F. (1999). Why do almost all mammals have seven cervical vertebrae? Developmental constraints, Hox genes, and cancer. *J Exp Zool* *285*, 19-26.

- Gibson, G., and Gehring, W. (1988). Head and thoracic transformations caused by ectopic expression of Antennapedia during *Drosophila* development. *Development* *102*, 657-675.
- Hamburger, V., and Hamilton, H. L. (1951). A series of normal stages in the development of the chick embryo. 1951. *Dev Dyn* *195*, 231-272.
- Horan, G. S., Kovacs, E. N., Behringer, R. R., and Featherstone, M. S. (1995a). Mutations in paralogous Hox genes result in overlapping homeotic transformations of the axial skeleton: evidence for unique and redundant function. *Dev Biol* *169*, 359-372.
- Horan, G. S., Ramirez-Solis, R., Featherstone, M. S., Wolgemuth, D. J., Bradley, A., and Behringer, R. R. (1995b). Compound mutants for the paralogous *hoxa-4*, *hoxb-4*, and *hoxd-4* genes show more complete homeotic transformations and a dose-dependent increase in the number of vertebrae transformed. *Genes Dev* *9*, 1667-1677.
- Horan, G. S., Wu, K., Wolgemuth, D. J., and Behringer, R. R. (1994). Homeotic transformation of cervical vertebrae in *Hoxa-4* mutant mice. *Proc Natl Acad Sci U S A* *91*, 12644-12648.
- Hornstein, E., Mansfield, J. H., Yekta, S., Hu, J. K., Harfe, B. D., McManus, M. T., Baskerville, S., Bartel, D. P., and Tabin, C. J. (2005). The microRNA miR-196 acts upstream of *Hoxb8* and *Shh* in limb development. *Nature* *438*, 671-674.
- Jeannotte, L., Lemieux, M., Charron, J., Poirier, F., and Robertson, E. J. (1993). Specification of axial identity in the mouse: role of the *Hoxa-5* (*Hox1.3*) gene. *Genes Dev* *7*, 2085-2096.
- Kessel, M., and Gruss, P. (1991). Homeotic transformations of murine vertebrae and concomitant alteration of Hox codes induced by retinoic acid. *Cell* *67*, 89-104.
- Krumlauf, R. (1993). Mouse Hox genetic functions. *Curr Opin Genet Dev* *3*, 621-625.
- Krumlauf, R. (1994). Hox genes in vertebrate development. *Cell* *78*, 191-201.
- 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.
- 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.
- Lufkin, T., Mark, M., Hart, C. P., Dolle, P., LeMeur, M., and Chambon, P. (1992). Homeotic transformation of the occipital bones of the skull by ectopic expression of a homeobox gene. *Nature* *359*, 835-841.
- Maconochie, M., Nonchev, S., Morrison, A., and Krumlauf, R. (1996). Paralogous Hox genes: function and regulation. *Annu Rev Genet* *30*, 529-556.
- Manley, N. R., and Capecchi, M. R. (1997). Hox group 3 paralogous genes act synergistically in the formation of somitic and neural crest-derived structures. *Dev Biol* *192*, 274-288.

- Mansfield, J. H., Harfe, B. D., Nissen, R., Obenaus, J., Srineel, J., Chaudhuri, A., Farzan-Kashani, R., Zuker, M., Pasquinelli, A. E., Ruvkun, G., *et al.* (2004). MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nat Genet* 36, 1079-1083.
- Manzanares, M., Bel-Vialar, S., Ariza-McNaughton, L., Ferretti, E., Marshall, H., Maconochie, M. M., Blasi, F., and Krumlauf, R. (2001). Independent regulation of initiation and maintenance phases of *Hoxa3* expression in the vertebrate hindbrain involve auto- and cross-regulatory mechanisms. *Development* 128, 3595-3607.
- McLain, K., Schreiner, C., Yager, K. L., Stock, J. L., and Potter, S. S. (1992). Ectopic expression of *Hox-2.3* induces craniofacial and skeletal malformations in transgenic mice. *Mech Dev* 39, 3-16.
- McLeod, M. J. (1980). Differential staining of cartilage and bone in whole mouse fetuses by alcian blue and alizarin red S. *Teratology* 22, 299-301.
- Morata, G. (1993). Homeotic genes of *Drosophila*. *Curr Opin Genet Dev* 3, 606-614.
- Nelson, C. E., Morgan, B. A., Burke, A. C., Laufer, E., DiMambro, E., Murtaugh, L. C., Gonzales, E., Tessarollo, L., Parada, L. F., and Tabin, C. (1996). Analysis of Hox gene expression in the chick limb bud. *Development* 122, 1449-1466.
- Pollock, R. A., Sreenath, T., Ngo, L., and Bieberich, C. J. (1995). Gain of function mutations for paralogous Hox genes: implications for the evolution of Hox gene function. *Proc Natl Acad Sci U S A* 92, 4492-4496.
- Ronshaugen, M., Biemar, F., Piel, J., Levine, M., and Lai, E. C. (2005). The *Drosophila* microRNA *iab-4* causes a dominant homeotic transformation of halteres to wings. *Genes Dev* 19, 2947-2952.
- Rossel, M., and Capecchi, M. R. (1999). Mice mutant for both *Hoxa1* and *Hoxb1* show extensive remodeling of the hindbrain and defects in craniofacial development. *Development* 126, 5027-5040.
- Struhl, G. (1983). Role of the *esc+* gene product in ensuring the selective expression of segment-specific homeotic genes in *Drosophila*. *J Embryol Exp Morphol* 76, 297-331.
- Takahara, Y., Tomotsune, D., Shirai, M., Katoh-Fukui, Y., Nishii, K., Motaleb, M. A., Nomura, M., Tsuchiya, R., Fujita, Y., Shibata, Y., *et al.* (1997). Targeted disruption of the mouse homologue of the *Drosophila* polyhomeotic gene leads to altered anteroposterior patterning and neural crest defects. *Development* 124, 3673-3682.
- van den Akker, E., Fromental-Ramain, C., de Graaff, W., Le Mouellic, H., Brulet, P., Chambon, P., and Deschamps, J. (2001). Axial skeletal patterning in mice lacking all paralogous group 8 Hox genes. *Development* 128, 1911-1921.

Williams, M. E., Lehoczky, J. A., and Innis, J. W. (2006). A group 13 homeodomain is neither necessary nor sufficient for posterior prevalence in the mouse limb. *Dev Biol* 297, 493-507.

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

Future Directions

Projects aimed at extending phenotypic consequences of miRNA-mediated repression of Hox genes to mouse models are ongoing. Complete knockouts of all three copies of miR-196 will resolve partial phenotypes obtained through knockdowns in the chick, and address ambiguities as to timing requirements for miR-196 with respect to patterning. The knockouts are designed so that reporters will substitute miRNA hairpins, and reporter expression will mimic that of the pri-miRNA, and provide detailed information about the expression of the miRNA, currently unavailable by other experimental means such as the miRNA sensor or LNA-probe hybridisation technologies. A triple knockout of miR-196 in the mouse will also be highly useful system for assessing miR-196 effects on temporal and spatial aspect of target Hox expression. Definite answers regarding domains of miRNA-mediated repression can be obtained by performing various analyses on target Hox mRNA distribution and levels. In situ hybridisation and qRT-PCR at different positions along the AP axis while technically challenging, may provide information about ratios of specific genes (paralogues in particular) in the absence or presence of the miRNA.

A complete knockout of miR-196 provides excellent material for microarray analysis, perhaps at different developmental stages enabling the identification of the miRNA targets, and the targets of Hox transcription factors. Tail tissue may be a potentially good source of for initial assessment of miR-196-dependent gene fluctuations. Microarray experiments may also be performed on antagomir-injected mice or chick, for a comparative study allowing for the identification of conserved regulation, and providing greater confidence about individual targets. The mouse knockout model however, remains a better system for microarray analysis, as it offers less variability in biological sampling, and a more reliable 3' UTR annotation.

Studies aimed at directing knock-ins of disrupted seed matches into target Hox genes are also on going for several Hox7 and Hox8 targets of miR-196. Such mouse models will shed light on function of particular miRNA-target relationships within specific domains of action of the miRNA, and mutants crossed to miR-196.

Given its extended effect on skeletal patterning in chicks, miR-10 is also a suitable candidate for targeted gene-disruption in the mouse. With only two copies, it is a more accessible knockout than miR-196. Furthermore, the smaller number of Hox targets, make it easier to parse out Hox contribution to specific skeletal defects.

In response to retinoic acid (RA) treatment cultured stem cells undergo a differentiation program toward neuronal fates. RA also triggers a temporal activation of Hox genes reflecting their gene order. This system used to study RA-dependent expression of Hox cluster miRNAs, and its effects on Hox genes. It could also be a starting point for biochemical and cellular assays probing miRNA-target relationships.

Plants and animal cells both possess the catalytic activity to direct the site-specific miRNA-directed cleavage of an mRNA target. However, while plant miRNAs principally use chemical cleavage to silence their targets, very few animal miRNAs possess sufficient sequence complementarity to their targets to mediate cleavage. Is there a biological relevance to the nature of complementarity between some vertebrate *Hoxb8* 3' UTRs and miR-196? Is the rapid clearing of *Hoxb8* required for some biological events to explain its phylogenetic conservation? While it appears that in cultured cells the extent of silencing conferred by the *Hoxb8* target site is

comparable to the canonical sites in *Hoxa7*, it may be that the messages are degraded at different rates. Examining the kinetics of decay of these mRNAs induced by miR-196 in cultured cells would partially address these questions. Substitution of the extended near-perfect target site of *Hoxb8* with the 4-5 classical seed matches of *Hoxa7* by homologous recombination in the mouse, would further address issues regarding the biological requirement for cleavage of *Hoxb8*.

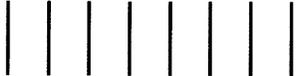
Is Ago2 association different for miR-196 and other miRNAs, thus favouring it for cleavage? Is there actual cellular competition between miR-34 targeting of and the miR-196 extended site in *Hoxb8*, as suggested by the complementarity of miR-34 with the 3' of this site? Are the two miRNAs functionally complementary when coexpressed, with no difference in the mode of silencing? Available cell-based assays to address these questions are sufficiently advanced, and the Hox/miR-196 system provides a biological framework to pursue them.

The Hox miRNAs also have other interesting predicted targets, of which, Ephrins 7 and Ephrin 4 have already been shown to be regulated by Hox genes. If these genes prove to be functional targets based on microarray or other experiments, knock-ins of their target sites may be worth pursuing.

Finally, an extension of the phylogenetic tree of miR-196 is of interest, as it would provide information about the early evolution of the miRNA. Amphioxus is a basal chordate with a single Hox cluster lacking a copy of miR-196, or related sequences. This absence implies that the miRNA emerged after divergence of cephalochordates, or that there was a lineage-specific gene loss in amphioxus. It may be interesting to see if amphioxus has a functional orthologue of this miRNA, similar to *drosophila*, where an unrelated miRNA is expressed from an orthologous genomic location, and capable of repressing neighbouring Hox genes situated 3' to its locus. It is possible to computationally predict amphioxus-specific miRNA hairpins in the intergenic space between amphiox9 and amphiox10; detect any seed matches in downstream Hox 3' UTRs, and assay for miRNA expression by Northern blotting.

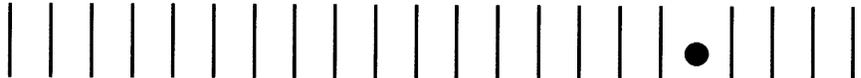
miR-34b

GUUAGUCGAUUAAUGUGACGGAUC 5'



Hoxb8 3' UTR

CCCAACAACAUGAAACUGCCUA 3'



miR-196a

GGGUUGUUGUACUUUGAUGGAU 5'

Appendix I

The microRNAs of *Caenorhabditis elegans*

Lee P. Lim,^{1,2,3,4} Nelson C. Lau,^{1,2,3} Earl G. Weinstein,^{1,2,3} Aliaa Abdelhakim,^{1,2,3} Soraya Yekta,^{1,2} Matthew W. Rhoades,^{1,2} Christopher B. Burge,^{1,5} and David P. Bartel^{1,2,6}

¹Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA, and ²Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

MicroRNAs (miRNAs) are an abundant class of tiny RNAs thought to regulate the expression of protein-coding genes in plants and animals. In the present study, we describe a computational procedure to identify miRNA genes conserved in more than one genome. Applying this program, known as MiRscan, together with molecular identification and validation methods, we have identified most of the miRNA genes in the nematode *Caenorhabditis elegans*. The total number of validated miRNA genes stands at 88, with no more than 35 genes remaining to be detected or validated. These 88 miRNA genes represent 48 gene families; 46 of these families (comprising 86 of the 88 genes) are conserved in *Caenorhabditis briggsae*, and 22 families are conserved in humans. More than a third of the worm miRNAs, including newly identified members of the *lin-4* and *let-7* gene families, are differentially expressed during larval development, suggesting a role for these miRNAs in mediating larval developmental transitions. Most are present at very high steady-state levels—more than 1000 molecules per cell, with some exceeding 50,000 molecules per cell. Our census of the worm miRNAs and their expression patterns helps define this class of noncoding RNAs, lays the groundwork for functional studies, and provides the tools for more comprehensive analyses of miRNA genes in other species.

[**Keywords:** miRNA; noncoding RNA; computational gene identification; Dicer]

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

Received January 13, 2003; accepted in revised form February 25, 2003.

Noncoding RNAs (ncRNAs) of ~22 nucleotides (nt) in length are increasingly recognized as playing important roles in regulating gene expression in animals, plants, and fungi. The first such tiny regulatory RNA to be identified was the *lin-4* RNA, which controls the timing of *Caenorhabditis elegans* larval development (Lee et al. 1993; Wightman et al. 1993). This 21-nt RNA pairs to sites within the 3' untranslated region (UTR) of target mRNAs, specifying the translational repression of these mRNAs and triggering the transition to the next developmental stage (Lee et al. 1993; Wightman et al. 1993; Ha et al. 1996; Moss et al. 1997; Olsen and Ambros 1999). A second tiny riboregulator, *let-7* RNA, is expressed later in development and appears to act in a similar manner to trigger the transition to late-larval and adult stages (Reinhart et al. 2000; Slack et al. 2000). The *lin-4* and *let-7* RNAs are sometimes called small temporal RNAs (stRNAs) because of their important roles in

regulating the timing of larval development (Pasquinelli et al. 2000). The *lin-4* and *let-7* stRNAs are now recognized as the founding members of a large class of ~22-nt ncRNAs termed microRNAs (miRNAs), which resemble stRNAs but do not necessarily control developmental timing (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001).

Understanding the biogenesis and function of miRNAs has been greatly facilitated by analogy and contrast to another class of tiny ncRNAs known as small interfering RNAs (siRNAs), first identified because of their roles in mediating RNA interference (RNAi) in animals and posttranscriptional gene silencing in plants (Hamilton and Baulcombe 1999; Hammond et al. 2000; Parrish et al. 2000; Zamore et al. 2000; Elbashir et al. 2001a; Klahre et al. 2002). During RNAi, long double-stranded RNA (either a bimolecular duplex or an extended hairpin) is processed by Dicer, an RNase III enzyme, into many siRNAs that serve as guide RNAs to specify the destruction of the corresponding mRNA (Hammond et al. 2000; Zamore et al. 2000; Bernstein et al. 2001; Elbashir et al. 2001a). Although these siRNAs are initially short double-stranded species with 5' phosphates and 2-nt 3' overhangs characteristic of RNase III cleavage products, they eventually become incorporated as single-stranded RNAs into a ribonucleoprotein com-

³These authors contributed equally to this work.

⁴Present address: Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034.

Corresponding authors.

⁵E-MAIL cburge@mit.edu; FAX (617) 452-2936.

⁶E-MAIL dbartel@wi.mit.edu; FAX (617) 258-6768.

Article published online ahead of print. Article and publication date are at <http://www.genesdev.org/cgi/doi/10.1101/gad.1074403>.

one that is evolutionarily conserved (Ambros et al. 2003).

Some miRNAs might be difficult to isolate by cloning, due to their low abundance or to biases in cloning procedures. Thus, computational identification of miRNAs from genomic sequences would provide a valuable complement to cloning. Recent advances have been made in the computational identification of ncRNA genes through comparative genomics, and complex algorithms have been developed to identify ncRNAs in general (Argaman et al. 2001; Rivas et al. 2001; Wassarman et al. 2001), as well as specific ncRNA families such as tRNAs and snoRNAs (Lowe and Eddy 1997, 1999).

In the present study, we describe a computational procedure to identify miRNA genes. By using this procedure, together with extensive sequencing of clones (3423 miRNA clones were sequenced), we have detected 30 additional miRNA genes, including previously unrecognized *lin-4* and *let-7* homologs. Extrapolation of the computational analysis indicates that miRNA gene identification in *C. elegans* is now approaching saturation, and that no more than 120 miRNA genes are present in this species. We also identify those genes with intriguing expression patterns during larval development and conditions of nutrient stress, and we show that most miRNAs are expressed at very high levels, with some present in as many copies per cell as the highly abundant U6 snRNA. This extensive census of worm miRNAs and their expression patterns establishes the general properties of this gene class and provides resources and tools for studies of miRNA function in nematodes and other organisms.

Results

Computational prediction of *C. elegans* miRNA genes

We developed a computational tool to specifically identify miRNAs that are conserved in two genomes and have the features characteristic of known miRNAs. To identify miRNAs in nematodes, the *C. elegans* genome was first scanned for hairpin structures with sequences that were conserved in *Caenorhabditis briggsae*. About 36,000 hairpins were found that satisfied minimum requirements for hairpin structure and sequence conservation. This procedure cast a sufficiently wide net to capture 50 of the 53 miRNAs previously reported to be conserved in the two species (Lau et al. 2001; Lee and Ambros 2001). These 50 published miRNA genes served as a training set for the development of a program called MiRscan, which was then used to assign scores to each of the 36,000 hairpins, evaluating them based on their similarity to the training set with respect to the following features: base pairing of the miRNA portion of the fold-back, base pairing of the rest of the fold-back, stringent sequence conservation in the 5' half of the miRNA, slightly less stringent sequence conservation in the 3' half of the miRNA, sequence biases in the first five bases of the miRNA (especially a U at the first position), a tendency toward having symmetric rather than asym-

metric internal loops and bulges in the miRNA region, and the presence of two to nine consensus base pairs between the miRNA and the terminal loop region, with a preference for 4–6 bp (Fig. 1A).

The distribution of MiRscan scores for the ~36,000 hairpins illustrated the ability of MiRscan to discern the 50 miRNA genes of the training set, which fell mostly in the high-scoring tail of the distribution (Fig. 2). Of the features evaluated by MiRscan, base-pairing potential and sequence conservation played primary roles in distinguishing known miRNAs (Fig. 1B). Some of the other conserved hairpins also scored highly; 35 had scores exceeding 13.9, the median score of the 58 known miRNAs (Fig. 2B). These 35 hairpins were carried forward as the top miRNA candidates predicted by MiRscan.

Molecular identification of miRNA genes

Our initial cloning and sequencing of small RNAs from mixed-stage *C. elegans* had identified 300 clones that represented 54 unique miRNA sequences (Lau et al. 2001). For the present study, this approach for identifying miRNAs was scaled-up ~10-fold. In an effort to identify miRNAs not normally expressed in mixed-stage logarithmically growing hermaphrodite worms, RNA was also cloned from populations of *him-8* worms, starved L1, and dauer worms. The *him-8* population was ~40% males, whereas the normal (N2) population was nearly all hermaphrodites (Broverman and Meneely 1994). Starved L1 and dauer worms are arrested in development at larval stages L1 and L3, respectively, with dauer worms having undergone morphological changes that enhance survival after desiccation or other harsh conditions.

As before, some clones matched *Escherichia coli*, the food source of the worms, others corresponded to fragments of annotated *C. elegans* RNAs. Nevertheless, 3423 clones were classified as miRNA clones (Table 1). Most of these represented the 58 miRNA genes previously identified in *C. elegans* (Lau et al. 2001; Lee and Ambros 2001). For example, *lin-4* was represented by 125 clones, *let-7* by 17 clones, and *mir-52* by 404 clones (Table 1). The remaining miRNA clones represented 23 newly identified miRNA loci.

In total, 80 loci were represented by cloned miRNAs (Table 1). Of these, 77 had the classical features of *C. elegans* miRNA genes, in that they had the potential to encode stereotypic hairpin precursor molecules with the 20- to 25-nt cloned RNAs properly positioned within an arm of the hairpin so as to be excised during Dicer processing, and their expression was manifested as a detectable Northern signal in the 20- to 25-nt range. Three other loci, *mir-41*, *mir-249*, and *mir-229*, were also included. The *mir-41* and *mir-249* RNAs were not detected on Northern blots but were still classified as miRNAs because these RNAs and their predicted hairpin precursors appear to be conserved in *C. briggsae*.

The *mir-229* locus was also classified as a miRNA gene, even though it appears to derive from an unusual fold-back precursor. Its precursor appears to be larger

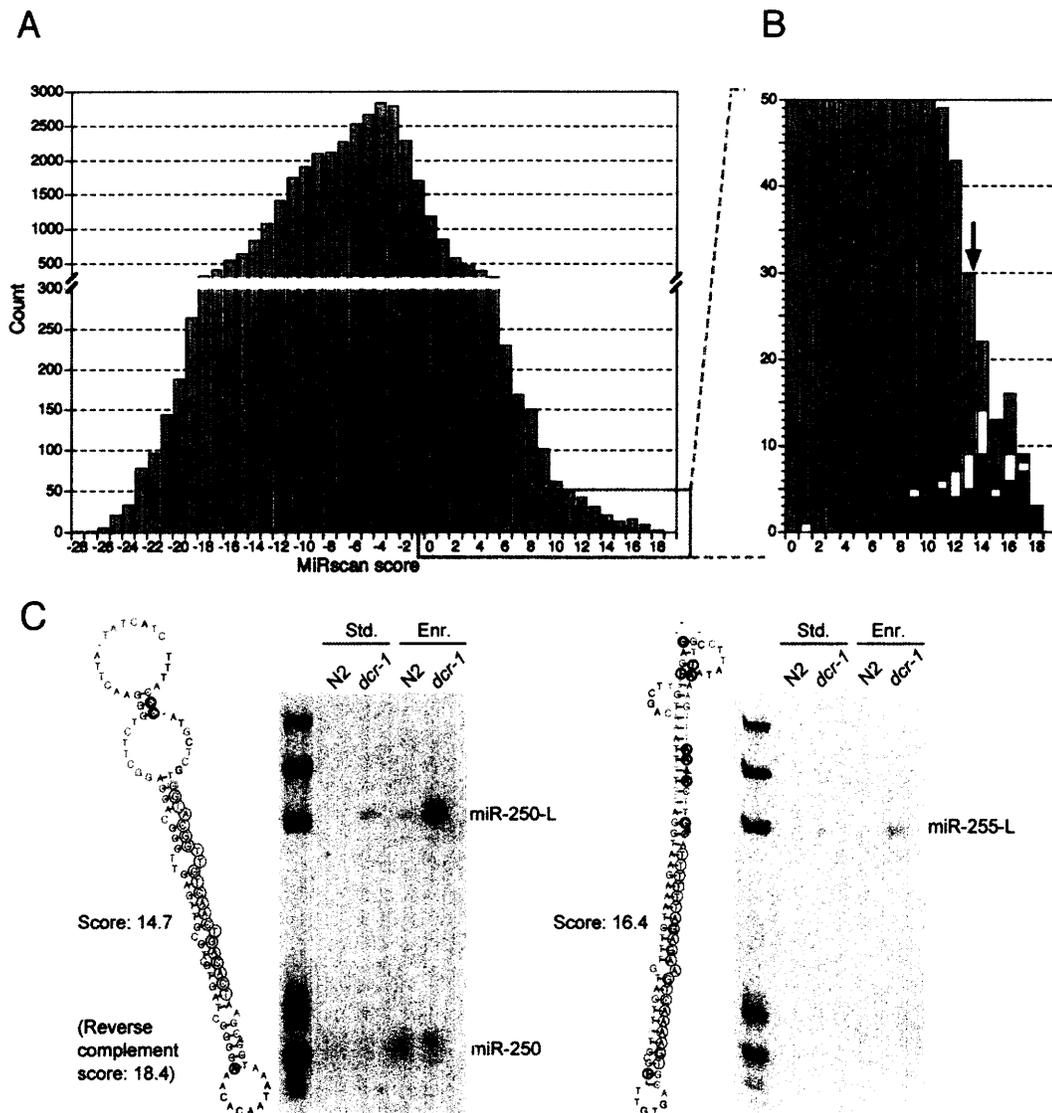


Figure 2. Computational identification of miRNA genes. (A) The distribution of MiRscan scores for 35,697 *C. elegans* sequences that potentially form stem loops and have loose conservation in *C. briggsae*. Note that the Y-axis is discontinuous so that the scores of the 50 previously reported miRNA genes that served as the training set for MiRscan can be more readily seen (red). Scores for these 50 genes were jackknifed to prevent inflation of their values because of their presence in the training set. (B) An expanded view of the high-scoring tail of the distribution. This view captures 49 of the 50 genes of the training set (red). The median score of the 58 previously reported miRNA loci that satisfy the current criteria for designation as miRNA genes (Ambros et al. 2003) is 13.9 (green arrow). Note that this median score was the midpoint between the scores of the 29th and 30th highest-scoring loci of the 50-member training set; namely, it was designated the median score after including the 8 previously reported miRNA genes that were not in the training set because they were lost during the identification of conserved hairpins, usually because they lacked sufficient *C. briggsae* homology. Scores of genes validated by cloning are indicated (yellow), as are scores of six genes that have not yet been cloned but were verified by Northern analysis (purple). (C) Examples of miRNA genes identified by MiRscan with the Northern blots that served to validate them. Stem-loops were annotated as in Figure 1A, except the DNA rather than RNA sequence is depicted. The Northern blots show analysis of RNA from either wild-type (N2) or *dcr-1* worms, isolated using either our standard protocol (Std.) or an additional polyethylene glycol precipitation step to enrich for small RNAs (Enr.). Homozygous worms of the *dcr-1* population have reduced Dicer activity, increasing the level of miRNA precursors (e.g., miR-250-L and miR-255-L), which facilitated the validation of miRNA loci, especially those for which the mature miRNA was not detected (e.g., miR-255). RNA markers (left lane) are 18, 21, 24, 60, 78, and 119 nt. The miR-250 stem loop shown received a MiRscan score of 14.7. The mir-250 reverse complement received an even greater score of 18.4, but was not detected by Northern analysis. Thus, the predicted *mir-250* gene was assigned the score of the higher-scoring, although incorrect, alternative stem loop (Table 1; Fig. 2B).

among the set of 35 high-scoring miRNA gene candidates and served to validate these 10 candidates.

The remaining 25 candidate miRNAs that had not been cloned were tested by Northern blots. RNA from

these miRNAs were missed in the current set of 3423 sequenced miRNA clones.

To investigate whether these miRNAs eventually would have been identified after further cloning and sequencing of our cDNA library of small RNA sequences, a PCR assay was used to detect the presence of these miRNAs in the library. By using a primer specific to the 3' segment of the predicted miRNA, together with a second primer corresponding to the adapter sequence attached to the 5' terminus of all the small RNAs, the 5' segment of the miRNA was amplified, cloned, and sequenced. This procedure validated five of the six predicted miRNAs for which at least a precursor could be detected on Northern blots, including two of the candidates (miR-253 and miR-254) for which a mature ~22-nt RNA was not detected on Northern blots. In addition, it identified the 5' terminus of these five miRNAs, which is difficult to achieve with confidence when using only bioinformatics and hybridization.

Combining the cloning and expression data, 16 of the 35 computationally identified candidates were validated (10 from cloning, five from Northern blots plus the PCR assay, and one from Northern blots only, which validated the precursor but did not identify the mature miRNA). Of the remaining 19 candidates, four could be readily classified as false positives. They appear to be nonannotated larger ncRNA genes, in that probes designed to hybridize to these candidates hybridized instead to high-molecular-weight species that remained constant in the samples from *dcr-1* worms. The remaining 15 new candidates with high MiRscan scores but without any Northern signal might also be false positives, or they might be authentic miRNAs that are expressed at low levels or in only very specific cell types or circumstances. Considering the extreme case in which all the nonvalidated candidates are false positives, the minimum specificity of MiRscan for the *C. elegans/C. briggsae* analysis can be calculated as $(29 + 16)/(29 + 35)$, or 0.70, at a sensitivity level that detects half of the 58 previously known miRNAs. A summary of the miRNA genes newly identified by validating computational candidates (16 genes) or by cloning alone (13 genes) is shown in Table 2, and predicted stem-loop precursors are shown in Supplemental Material. Table 2 also includes one additional gene, *mir-239b*, which was identified based on its homology with *mir-239a* and its MiRscan score of 13.6.

Evolutionary conservation of miRNAs

The 88 *C. elegans* miRNA genes identified to this point were grouped into 48 families, each comprising one to eight genes (data not shown). Within families, sequence identity either spanned the length of the miRNAs or was predominantly at their 5' terminus. All but two of these families extended to the miRNAs of *C. briggsae*. The two families without recognizable *C. briggsae* orthologs each comprised a single miRNA (miR-78 and miR-243). Thus, nearly all (>97%) of the *C. elegans* miRNAs identified had apparent homologs in *C. briggsae*, and all but six of these *C. elegans* miRNAs (miR-72, miR-63, miR-

64, miR-66, miR-229, and miR-247) had retained at least 75% sequence identity to a *C. briggsae* ortholog. Of the 48 *C. elegans* miRNA families, 22 also had representatives among the known human miRNA genes (Fig. 3). In that these 22 families included 33 *C. elegans* genes, it appears that at least a third (33/88) of the *C. elegans* miRNA genes have homologs in humans and other vertebrates.

Developmental expression of miRNAs

The expression of 62 miRNAs during larval development was examined and compiled together with previously reported expression profiles (Lau et al. 2001) to yield a comprehensive data set for the 88 *C. elegans* miRNAs (Fig. 4). RNA from wild-type embryos, the four larval stages (L1 through L4), and young adults was probed, as was RNA from *glp-4* (*bn2*) young adults, which are severely depleted in germ cells (Beanan and Strome 1992). Nearly two thirds of the miRNAs appeared to have constitutive expression during larval development (Fig. 4A). These miRNAs might still have differential expression during embryogenesis, or they might have tissue-specific expression, as has been observed for miRNAs of larger organisms in which tissues and organs can be more readily dissected and examined (Lee and Ambros 2001; Lagos-Quintana et al. 2002; Llave et al. 2002a; Park et al. 2002; Reinhart et al. 2002).

Over one third of the miRNAs had expression patterns that changed during larval development (Fig. 4B,C), and there were examples of miRNA expression initiating at each of the four larval stages (Fig. 4B). Expression profiles for miR-48 and miR-241 (which are within 2 kb of each other in the *C. elegans* genome) were similar to those previously reported for *let-7* RNA and miR-84 (Fig. 4B; Reinhart et al. 2000; Lau et al. 2001). In fact, these four miRNAs appear to be paralogs, with all four miRNAs sharing the same first eight residues (Fig. 3). Another newly identified miRNA, miR-237, is a paralog of the other canonical stRNA, *lin-4* RNA (Fig. 3), although miR-237 exhibited an expression pattern distinct from *lin-4* RNA (Fig. 4E). The existence of these paralogs, as well as other families of miRNAs with expression initiating at the different stages of larval development, supports the idea that *lin-4* and *let-7* miRNAs are not the only stRNAs with important roles in the *C. elegans* heterochronic pathway.

Expression usually remained constant once it initiated, as has been seen for *lin-4* and *let-7* miRNA expression (Fig. 4A,B). Exceptions to this trend included the miRNAs of the *mir-35-mir-41* cluster, which were expressed transiently during embryogenesis (Lau et al. 2001); miR-247, which was expressed transiently in larval stage 3 (and dauer); and miR-248, which was most highly expressed in dauer (Fig. 4C,D). miR-234 was expressed in all stages, but expression was highest in both L1 worms (which had been starved shortly before harvest to synchronize the worm developmental staging) and dauer worms, suggesting that this miRNA might be induced as a consequence of nutrient stress.

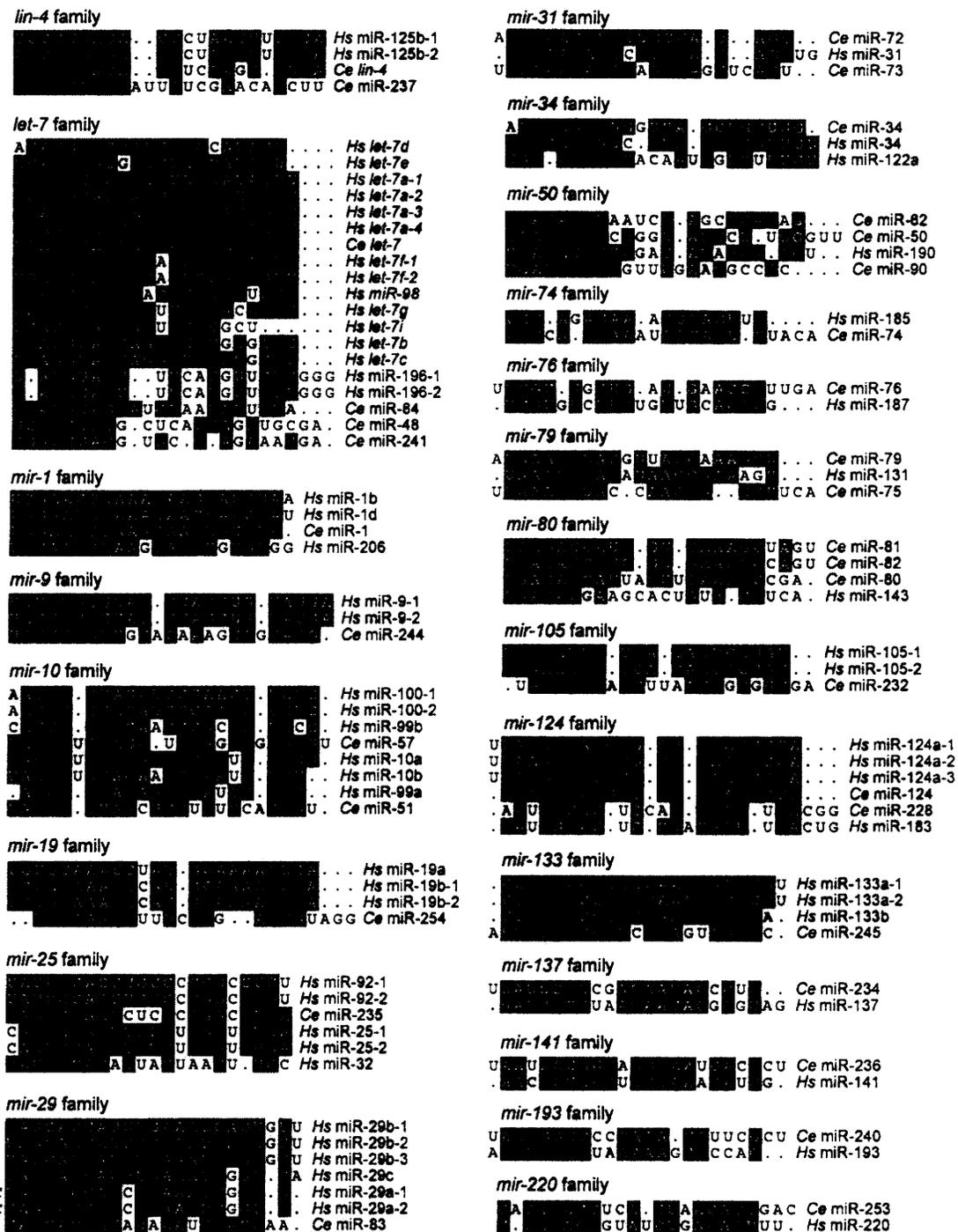


Figure 3. Alignments of *C. elegans* and human miRNA sequences that can be grouped together in families. Human miRNAs (Hs) are those identified in human cells (Lagos-Quintana et al. 2001; Mourelatos et al. 2002) or are orthologs of miRNAs identified in other vertebrates (Lagos-Quintana et al. 2002, 2003; Lim et al. 2003).

mRNAs from mouse tissues [Hastie and Bishop 1976].] Perhaps high concentrations of miRNAs are needed to saturate the relevant complementary sites within the target mRNAs, which might be recognized with low affinity because of the noncanonical pairs or bulges that

appear to be characteristic of the animal miRNA-target interactions.

Because these numbers represent molecular abundance averaged over all the cells of the worm, including cells that might not be expressing the miRNA, there are

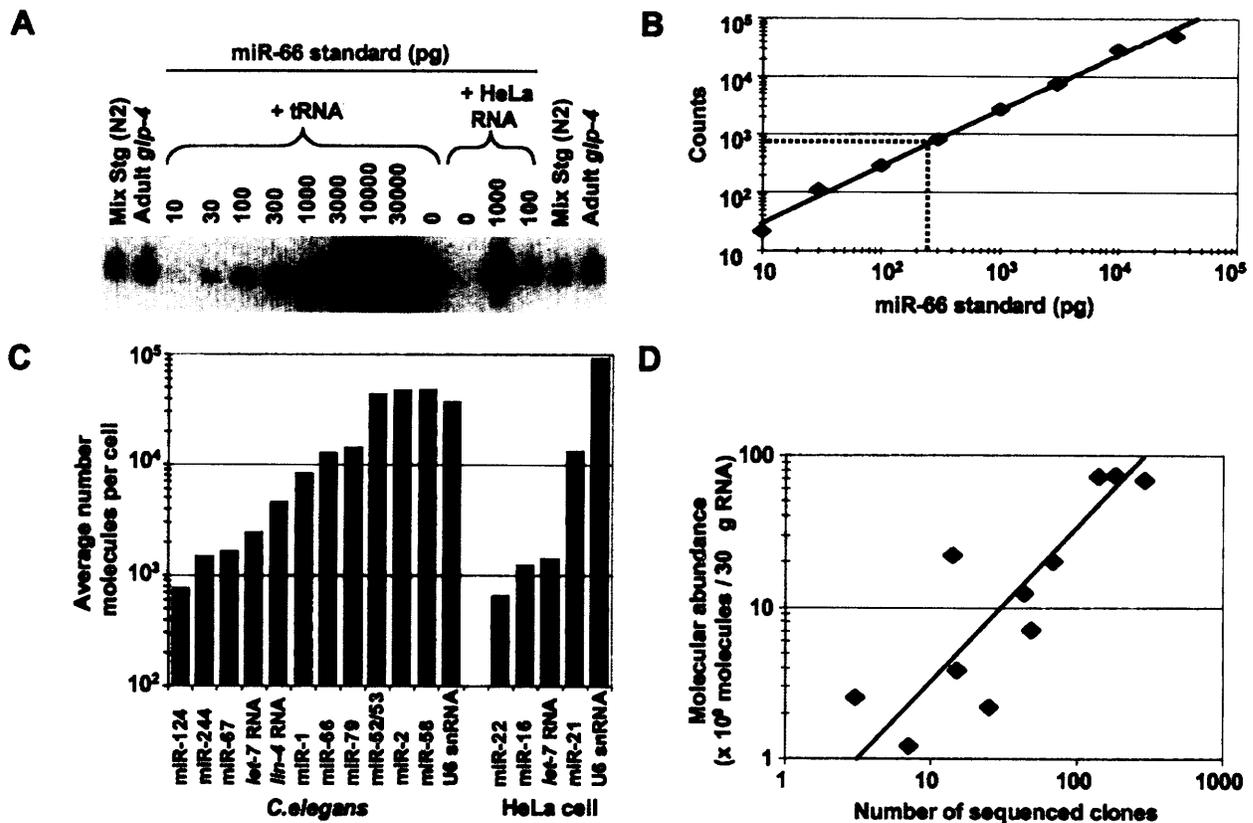


Figure 5. Quantitative analysis of miRNA expression. (A) Northern blot used to quantify the abundance of miR-66. RNA prepared from the wild-type (N2) mixed-stage worms used in cloning and from *glp-4(bn2)* young adult worms were run in duplicate with a concentration course of synthetic miRNA standard. The signal from the standard did not change when total RNA from HeLa cells replaced *E. coli* tRNA as the RNA carrier, showing that the presence of other miRNAs did not influence membrane immobilization of the miRNA or hybridization of the probe. (B) Standard curve from quantitation of miR-66 concentration course. The best fit to the data is a line represented by the equation $y = 3.3x^{0.96}$ ($R^2 = 0.99$). Interpolation of the average signal in the *glp-4* lanes indicates that the *glp-4* samples contain 240 pg of miR-66 (broken lines). (C) Molecular abundance of miRNAs and U6 snRNA. Amounts of the indicated RNA species in the *glp-4* samples were determined as shown in A and B. The average number of molecules per cell was then calculated considering the number of animals used to prepare the sample, and the yield of a radiolabeled miRNA spiked into the preparation at an early stage of RNA preparation. Analogous experiments were performed to determine the amounts of the indicated human miRNAs in HeLa RNA samples. (D) Correlation between miRNA molecular abundance and cloning frequency. The number of molecules in the mixed-stage RNA samples was determined as described for the *glp-4* samples and then plotted as a function of the number of times the miRNAs was cloned from this mixed-stage population (Table 1). The line is best fit to the data and is represented by the equation $y = 0.32x$ ($R^2 = 0.78$).

dance of the miRNA within the mixed-stage RNA preparation was compared with the number of clones generated from that preparation (Fig. 5D). The strong positive correlation observed between the molecular abundance and the number of times the miRNAs were cloned indicated that systematic biases in the cloning procedure were not major. At most, these miRNAs were over- or underrepresented fivefold in the sequenced set relative to their actual abundance as measured by quantitative Northern blots. We cannot rule out the possibility that certain miRNAs not yet cloned might be refractory to our cloning procedure, for example, because of a propensity to form secondary structures that preclude adaptor ligation reactions. Nonetheless, on the whole, the cloning frequencies can be used to approximate the molecular abundance of the miRNAs, and we have no reason to

suspect that the set of miRNAs identified by cloning differs in any substantive way, other than an overall higher steady-state expression level, from the complete set of *C. elegans* miRNAs.

Other endogenous ~22-nt RNAs of *C. elegans*

Of the 4078 *C. elegans* clones, a large majority represented authentic miRNAs (3423 clones, Table 1). The next most abundant class represented degradation fragments of larger ncRNAs, such as tRNA and rRNA (447 clones) and introns (18 clones). The remaining clones represented potential Dicer products that were not classified as miRNAs. Some corresponded to sense (18 clones) or antisense (23 clones) fragments of known or predicted mRNAs and might represent endogenous

MiRscan will improve with a more complete and assembled *C. briggsae* genome. We anticipate that using only those sequences conserved in a syntenic alignment of the two genomes would capture fewer of the background sequences, enabling the authentic miRNAs to be more readily distinguished from the false positives.

Improvement would also come from bringing in a third nematode genome, particularly a genome more divergent than those of *C. elegans* and *C. briggsae*. The advantage of such an additional genome is illustrated by our application of MiRscan to the identification of vertebrate miRNAs using three genomes. The version of MiRscan described here, which had been trained on the set of 50 miRNAs conserved in worms, was applied to the assembled human genome, shotgun reads of the mouse genome, and the assembled pufferfish (*Fugu*) genome (Lim et al. 2003). This analysis had a specificity of ≥ 0.71 at a sensitivity that detected three fourths of the previously known vertebrate miRNAs. The accuracy of the vertebrate analysis was therefore substantially improved over that of the *C. elegans/C. briggsae* analysis, even though the vertebrate genomes are 4–30 times larger than those of *C. elegans* and *C. briggsae*, and are expected to have a correspondingly higher number of background sequences. This improved performance can be attributed to using three genomes, as well as to the evolutionary distance between the mammalian and fish genomes, which are distant enough to reduce the number of fortuitously high scoring sequences, yet close enough to retain most of the known miRNAs.

Other improvements in the computational identification of miRNAs will come with the definition of additional sequence and structural features that specify which sequences are transcribed, processed into miRNAs, and loaded into the miRNP. With the exception of sequence conservation, the features that MiRscan currently uses to identify miRNAs (Fig. 1A) are among those that the cell also uses to specify the biogenesis of miRNAs and miRNPs. The utility of these parameters for MiRscan (Fig. 1B) is a function of both the degree to which these features are correctly modeled (or have already been used to restrict the number of miRNA candidates; see Fig. 1B legend) and their relative importance in vivo. Clearly, much of what defines a miRNA in vivo remains to be determined. Sequence elements currently unavailable for MiRscan include transcriptional promoter and termination signals. Additional sequence and structural features important for processing of the primary transcript and the hairpin precursors also remain to be identified (Lee et al. 2002).

miRNA biogenesis

The presence of miRNA* species, observed now for 14 of the *C. elegans* miRNAs (Fig. 6; Lau et al. 2001), provides evidence for the idea that Dicer processing of miRNA precursors resembles that of siRNA precursors (Hutvagner and Zamore 2002; Reinhart et al. 2002). We suspect that with more extensive sequencing of clones,

miRNA* sequences will be found for a majority of the miRNA precursors, a notion supported by the identification of additional miRNA* sequences using our PCR assay (data not shown). As observed for both *MIR156d* and *MIR169* in plants (Reinhart et al. 2002), the miRNA:miRNA* segments are typically presented within the predicted precursor, paired to each other with 2-nt 3' overhangs (Fig. 6)—a structure analogous to that of a classical siRNA duplex. This is precisely the structure that would be expected if both the miRNA and the miRNA* were excised from the same precursor molecule, and the miRNA* fragments were transient side-products of productive Dicer processing. An alternative model for miRNA biogenesis and miRNA* formation, which we do not favor but cannot rule out, is that the Dicer complex normally excises a ~22-nt RNA from only one side of a miRNA precursor but it sometimes binds the precursors in the wrong orientation and excises the wrong side. In an extreme version of the favored model, the production of the miRNA* would be required for miRNA processing and miRNP assembly; in a less extreme version, miRNA* production would be an optional off-pathway phenomenon. The idea that ~22-nt RNAs might be generally excised from both sides of the same precursor stem loop brings up the question of why the miRNAs and miRNA*s are present at such differing levels. With the exception of miR-34* (sequenced 17 times), none of the miRNA*s is represented by more than three sequenced clones. Perhaps the miRNAs are stabilized relative to their miRNA* fragments because they preferentially enter the miRNP/RISC complex. Alternatively, both the miRNA and the miRNA* might enter the complex, but the miRNA might be stabilized by interactions with its targets.

Five of the newly identified miRNAs are within annotated introns, all five in the same orientation as the predicted mRNAs. When considered together with the previously identified miRNAs found within annotated introns (Lau et al. 2001), 10 of 12 known *C. elegans* miRNAs predicted to be in introns are in the same orientation as the predicted mRNAs. This bias in orientation, also reported recently for mammalian miRNAs (Lagos-Quintana et al. 2003), suggests that some of these miRNAs are not transcribed from their own promoters but instead derive from the excised pre-mRNA introns (as are many snoRNAs), and it is easy to imagine regulatory scenarios in which the coordinate expression of a miRNA with an mRNA would be desirable.

The number of miRNA genes in *C. elegans* and other animals

In addition to providing a set of candidate miRNAs, MiRscan scoring provides a means to estimate the total number of miRNA genes in *C. elegans*. A total of 64 loci have scores greater than the median score of the 58 initially reported *C. elegans* miRNAs (Fig. 2B). Note that this set of 58 miRNAs includes not only the 50 conserved miRNAs of the training set but also the eight previously reported miRNAs that were not in our set of

their importance in specifying cell differentiation and developmental patterning, and that the extra layer of gene regulation afforded by miRNAs was crucial for the emergence of multicellular body plans. The identification of most of the worm miRNAs and the quantitation of the number of genes remaining to be found are important steps toward understanding the evolution of this intriguing class of genes and placing them within the gene regulatory circuitry of these and other animals.

Materials and methods

Computational identification of stem loops

Potential miRNA stem loops were located by sliding a 110-nt window along both strands of the *C. elegans* genome (Worm-Base release 45, <http://www.wormbase.org>) and folding the window with the secondary structure-prediction program RNAfold (Hofacker et al. 1994) to identify predicted stem-loop structures with a minimum of 25 bp and a folding free energy of at least 25 kcal/mole ($\Delta G_{\text{folding}} \leq -25$ kcal/mole). Sequences that matched repetitive elements were discarded, as were those with skewed base compositions not observed in known miRNA stem loops and those that overlapped with annotated coding regions. Stem loops that had fewer base pairs than overlapping stem loops were also culled. *C. briggsae* sequences with at least loose sequence similarity to the remaining *C. elegans* sequences were identified among *C. briggsae* shotgun sequencing reads (November 2001 download from <http://www.ncbi.nlm.nih.gov/Traces>) using WU-BLAST with default parameters and a non-stringent cutoff of $E < 1.8$ (W. Gish, <http://blast.wustl.edu>). These *C. briggsae* sequences were folded with RNAfold to ensure that they met the minimal requirements for a hairpin structure as described above. This procedure yielded ~40,000 pairs of potential miRNA hairpins. For each pair of potential miRNA hairpins, a consensus *C. elegans/C. briggsae* structure was generated using the alidot and pfrali utilities from the Vienna RNA package (Hofacker et al. 1998; Hofacker and Stadler 1999; <http://www.tbi.univie.ac.at/~ivo/RNA>). To create RNA consensus structures, alidot and pfrali combine a Clustal alignment (Thompson et al. 1994) of a pair of sequences with either the minimum free energy structures of these sequences (alidot) derived using the Zuker algorithm (Zuker 1994) or the base pairing probability matrices of these sequences (pfrali) derived using the McCaskill algorithm (McCaskill 1990).

MiRscan

Of the ~40,000 pairs of hairpins, 35,697 had the minimal conservation and base pairing needed to receive a MiRscan score. Among this set were 50 of the 53 previously published miRNAs that were reported to be conserved between *C. elegans* and *C. briggsae* (Lau et al. 2001; Lee and Ambros 2001). [miR-53 is included as a previously reported conserved miRNA because it is nearly identical to miR-52, which has a highly conserved *C. briggsae* ortholog (Lau et al. 2001; Lee and Ambros 2001). The three conserved genes missing from the ~36,000 pairs of hairpins were *mir-56*, *mir-75*, and *mir-88*. The reverse complements of *mir-75* and *mir-88* were later observed among the ~36,000 hairpins and given scores (Table 1).] The MiRscan program was developed to discriminate these 50 known miRNA hairpins from background sequences in the set of ~36,000 hairpins. For a given 21-nt miRNA candidate, MiRscan makes use of the seven features derived from the consensus hairpin structure illus-

trated in Figure 1A: x_1 , "miRNA base pairing," the sum of the base-pairing probabilities for pairs involving the 21-nt candidate miRNA; x_2 , "extension of base pairing," the sum of the base-pairing probabilities of the pairs predicted to lie outside the 21-nt candidate miRNA but within the same helix; x_3 , "5' conservation," the number of bases conserved between *C. elegans* and *C. briggsae* within the first 10 bases of the miRNA candidate; x_4 , "3' conservation," the number of conserved bases within the last 11 bases of the miRNA candidate; x_5 , "bulge symmetry," the number of bulged or mismatched bases in the candidate miRNA minus the number of bulged or mismatched bases in the corresponding segment on the other arm of the stem loop; x_6 , "distance from loop," the number of base pairs between the loop of the stem loop and the closest end of the candidate; and x_7 , "initial pentamer," the specific bases at the first five positions at the candidate 5' terminus.

For a given feature i with a value x_i , MiRscan assigns a log-odds score

$$s_i(x_i) = \log_2 \left(\frac{f_i(x_i)}{g_i(x_i)} \right),$$

where $f_i(x_i)$ is an estimate of the frequency of feature value x_i in miRNAs derived from the training set of 50 known miRNAs, and $g_i(x_i)$ is an estimate of the frequency of feature value x_i among the background set of ~36,000 hairpin pairs. The overall score assigned to a candidate miRNA is simply the sum of the log-odds scores for the seven features:

$$S = \sum_{i=1..7} s_i(x_i).$$

To score a given hairpin, MiRscan slides a 21-nt window representing the candidate miRNA along each arm of the hairpin, assigns a score to each window, and then assigns the hairpin the score of its highest-scoring window. In order to be evaluated, a window was required to be two to nine consensus base pairs away from the terminal loop.

For features x_1 , x_3 , x_4 , x_5 , and x_6 , f_i and g_i were obtained by smoothing the empirical frequency distributions from the training and background sets, respectively, using the R statistical package (<http://lib.stat.cmu.edu/R/CRAN>) with a triangular kernel. Because x_1 and x_2 are not independent of each other, the relative contribution of x_2 was decreased by computing f_2 and g_2 separately subject to the conditions $x_2 \geq 9$ and $x_1 < 9$, in order to account for this dependence. For x_7 , a weight matrix model (WMM) was generated for the five positions at the miRNA 5' terminus. The background WMM, g_7 , was set equal to the base composition of the background sequence set. The miRNA WMM, f_7 , was derived from the position-specific base frequencies of the 50 training set sequences, using standard unit pseudo-counts and normalizing for the contributions of related miRNAs.

Because both strands of the *C. elegans* genome were analyzed, both a hairpin sequence and its reverse complement were sometimes included in the set of ~36,000 stem loops. For representation in Figure 2, in such cases both sequences were considered as a single locus that received the score of the higher scoring hairpin. Also, to prevent overscoring of the 50 known miRNA loci within the training set, each known miRNA locus was assigned a jackknife score calculated by using a training set consisting of the other 49 miRNAs. MiRscan is available for use (<http://genes.mit.edu/mirscan>).

RNA cloning and bioinformatic analyses

Small RNAs were cloned as described previously (Lau et al. 2001), using the protocol available on the Web (<http://web>).

- Broverman, S.A. and Meneely, P.M. 1994. Meiotic mutants that cause a polar decrease in recombination on the X chromosome in *Caenorhabditis elegans*. *Genetics* **136**: 119–127.
- Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**: 346–354.
- Caudy, A.A., Myers, M., Hannon, G.J., and Hammond, S.M. 2002. Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes & Dev.* **16**: 2491–2496.
- C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**: 2012–2018.
- Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282**: 2022–2028.
- Djikeng, A., Shi, H., Tschudi, C., and Ullu, E. 2001. RNA interference in *Trypanosoma brucei*: Cloning of small interfering RNAs provides evidence for retroposon-derived 24–26-nucleotide RNAs. *RNA* **7**: 1522–1530.
- Elbashir, S.M., Lendeckel, W., and Tuschl, T. 2001a. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Dev.* **15**: 188–200.
- Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. 2001b. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20**: 6877–6888.
- Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**: 23–34.
- Ha, I., Wightman, B., and Ruvkun, G. 1996. A bulged lin-4/lin-14 RNA duplex is sufficient for *Caenorhabditis elegans* lin-14 temporal gradient formation. *Genes & Dev.* **10**: 3041–3050.
- Hall, I.M., Shankaranarayana, G.D., Noma, K., Ayoub, N., Cohen, A., and Grewal, S.I. 2002. Establishment and maintenance of a heterochromatin domain. *Science* **297**: 2232–2237.
- Hamilton, A.J. and Baulcombe, D.C. 1999. A novel species of small antisense RNA in posttranscriptional gene silencing. *Science* **286**: 950–952.
- Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. 2002. Two classes of short interfering RNA in RNA silencing. *EMBO J.* **21**: 4671–4679.
- Hammond, S.C., Bernstein, E., Beach, D., and Hannon, G.J. 2000. An RNA-directed nuclease mediates posttranscriptional gene silencing in *Drosophila* cells. *Nature* **404**: 293–296.
- Hastie, N.D. and Bishop, J.O. 1976. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9**: 761–774.
- Hofacker, I.L. and Stadler, P.F. 1999. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput. Chem.* **15**: 401–414.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshfte f. Chemie* **125**: 167–188.
- Hofacker, I.L., Fekete, M., Flamm, C., Huynen, M.A., Rauscher, S., Stolorz, P.E., and Stadler, P.F. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26**: 3825–3836.
- Hutvagner, G. and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297**: 2056–2060.
- 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.
- Ishizuka, A., Siomi, M.C., and Siomi, H. 2002. A *Drosophila* fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes & Dev.* **16**: 2497–2508.
- Ketting, R.F., Fischer, S.E.J., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H.A. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes & Dev.* **15**: 2654–2659.
- Klahre, U., Crete, P., Leuenberger, S.A., Iglesias, V.A., and Meins, F. 2002. High molecular weight RNAs and small interfering RNAs induce systemic posttranscriptional gene silencing in plants. *Proc. Natl. Acad. Sci.* **99**: 11981–11986.
- 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.
- Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. 2003. New microRNAs from mouse and human. *RNA* **9**: 175–179.
- Lai, E.C. 2002. MicroRNAs are complementary to 3'UTR motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30**: 363–364.
- Lander E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitz Hugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- 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., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO J.* **21**: 4663–4670.
- Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003. Vertebrate microRNA genes. *Science* **299**: 1540.
- Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002a. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**: 1605–1619.
- Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. 2002b. Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297**: 2053–2056.
- Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**: 955–964.
- . 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**: 1168–1171.
- Martinez, J., Patkaniowska, A., Urlaub, H., Lührmann, R., and Tuschl, T. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**: 563–574.
- McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**: 1105–1119.
- Mochizuki, K., Fine, N.A., Fujisawa, T., and Gorovsky, M.A. 2002. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in Tetrahymena. *Cell* **110**: 689–699.

Appendix II

The microRNAs of *Caenorhabditis elegans*

Lee P. Lim,^{1,2,3,4} Nelson C. Lau,^{1,2,3} Earl G. Weinstein,^{1,2,3} Aliaa Abdelhakim,^{1,2,3} Soraya Yekta,^{1,2} Matthew W. Rhoades,^{1,2} Christopher B. Burge,^{1,5} and David P. Bartel^{1,2,6}

¹Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA, and ²Whitehead Institute for Biomedical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA

MicroRNAs (miRNAs) are an abundant class of tiny RNAs thought to regulate the expression of protein-coding genes in plants and animals. In the present study, we describe a computational procedure to identify miRNA genes conserved in more than one genome. Applying this program, known as MiRscan, together with molecular identification and validation methods, we have identified most of the miRNA genes in the nematode *Caenorhabditis elegans*. The total number of validated miRNA genes stands at 88, with no more than 35 genes remaining to be detected or validated. These 88 miRNA genes represent 48 gene families; 46 of these families (comprising 86 of the 88 genes) are conserved in *Caenorhabditis briggsae*, and 22 families are conserved in humans. More than a third of the worm miRNAs, including newly identified members of the *lin-4* and *let-7* gene families, are differentially expressed during larval development, suggesting a role for these miRNAs in mediating larval developmental transitions. Most are present at very high steady-state levels—more than 1000 molecules per cell, with some exceeding 50,000 molecules per cell. Our census of the worm miRNAs and their expression patterns helps define this class of noncoding RNAs, lays the groundwork for functional studies, and provides the tools for more comprehensive analyses of miRNA genes in other species.

[**Keywords:** miRNA; noncoding RNA; computational gene identification; Dicer]

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

Received January 13, 2003; accepted in revised form February 25, 2003.

Noncoding RNAs (ncRNAs) of ~22 nucleotides (nt) in length are increasingly recognized as playing important roles in regulating gene expression in animals, plants, and fungi. The first such tiny regulatory RNA to be identified was the *lin-4* RNA, which controls the timing of *Caenorhabditis elegans* larval development (Lee et al. 1993; Wightman et al. 1993). This 21-nt RNA pairs to sites within the 3' untranslated region (UTR) of target mRNAs, specifying the translational repression of these mRNAs and triggering the transition to the next developmental stage [Lee et al. 1993; Wightman et al. 1993; Ha et al. 1996; Moss et al. 1997; Olsen and Ambros 1999]. A second tiny riboregulator, *let-7* RNA, is expressed later in development and appears to act in a similar manner to trigger the transition to late-larval and adult stages (Reinhart et al. 2000; Slack et al. 2000). The *lin-4* and *let-7* RNAs are sometimes called small temporal RNAs (stRNAs) because of their important roles in

regulating the timing of larval development (Pasquinelli et al. 2000). The *lin-4* and *let-7* stRNAs are now recognized as the founding members of a large class of ~22-nt ncRNAs termed microRNAs (miRNAs), which resemble stRNAs but do not necessarily control developmental timing (Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001).

Understanding the biogenesis and function of miRNAs has been greatly facilitated by analogy and contrast to another class of tiny ncRNAs known as small interfering RNAs (siRNAs), first identified because of their roles in mediating RNA interference (RNAi) in animals and posttranscriptional gene silencing in plants (Hamilton and Baulcombe 1999; Hammond et al. 2000; Parrish et al. 2000; Zamore et al. 2000; Elbashir et al. 2001a; Klahre et al. 2002). During RNAi, long double-stranded RNA (either a bimolecular duplex or an extended hairpin) is processed by Dicer, an RNase III enzyme, into many siRNAs that serve as guide RNAs to specify the destruction of the corresponding mRNA (Hammond et al. 2000; Zamore et al. 2000; Bernstein et al. 2001; Elbashir et al. 2001a). Although these siRNAs are initially short double-stranded species with 5' phosphates and 2-nt 3' overhangs characteristic of RNase III cleavage products, they eventually become incorporated as single-stranded RNAs into a ribonucleoprotein com-

³These authors contributed equally to this work.

⁴Present address: Rosetta Inpharmatics, 12040 115th Avenue NE, Kirkland, Washington 98034.

Corresponding authors.

⁵E-MAIL cburge@mit.edu; FAX (617) 452-2936.

⁶E-MAIL dbartel@wi.mit.edu; FAX (617) 258-6768.

Article published online ahead of print. Article and publication date are at <http://www.genesdev.org/cgi/doi/10.1101/gad.1074403>.

Lim et al.

plex, known as the RNA-induced silencing complex (RISC; Hammond et al. 2000; Elbashir et al. 2001a,b; Nykäken et al. 2001; Martinez et al. 2002; Schwarz et al. 2002). The RISC identifies target messages based on perfect (or nearly perfect) antisense complementarity between the siRNA and the mRNA, and then the endonuclease of the RISC cleaves the mRNA at a site near the middle of the siRNA complementarity (Elbashir et al. 2001a,b). Similar pathways have been proposed for gene silencing in plants and fungi, with siRNAs targeting mRNA for cleavage during posttranscriptional gene silencing and heterochromatic siRNAs targeting chromatin for histone methylation, triggering heterochromatin formation and consequent transcriptional gene silencing (Hamilton and Baulcombe 1999; Vance and Vaucheret 2001; Hall et al. 2002; Hamilton et al. 2002; Pickford et al. 2002; Reinhart and Bartel 2002; Volpe et al. 2002; Zilberman et al. 2003).

miRNAs have many chemical and functional similarities to the siRNAs. Like siRNAs they are processed by Dicer, and so they are the same length and possess the same 5'-phosphate and 3'-hydroxyl termini as siRNAs (Grishok et al. 2001; Hutvagner et al. 2001; Ketting et al. 2001; Lau et al. 2001; Park et al. 2002; Reinhart et al. 2002). They are also incorporated within a ribonucleoprotein complex, known as the miRNP, which is similar if not identical to the RISC (Caudy et al. 2002; Hutvagner and Zamore 2002; Ishizuka et al. 2002; Martinez et al. 2002; Mourelatos et al. 2002). In fact, many plant miRNAs match their predicted mRNA targets with near-perfect antisense complementarity, as if they were functioning as siRNAs within a RISC complex (Rhoades et al. 2002), and the plant miR171 and miR165/166 have been shown to specify cleavage of their mRNA targets (Llave et al. 2002b; Tang et al. 2003). The *C. elegans* and *Drosophila* miRNAs do not have as pronounced a tendency to pair with their targets with near-perfect complementarity (Rhoades et al. 2002). Nonetheless, some might still direct cleavage of their targets, as suggested by the observation that miRNAs and siRNAs with 3–4 mismatches with their targets can still direct cleavage in plant and animal lysates (Tang et al. 2003). Furthermore, the *let-7* miRNA is present within a complex that can cleave an artificial RNA target when such a target is perfectly complementary to the miRNA (Hutvagner and Zamore 2002). The known biological targets of *lin-4* and *let-7* RNAs have several mismatches within the central region of the miRNA complementary sites, perhaps explaining why in these particular cases, the miRNAs specify translational repression rather than mRNA cleavage during *C. elegans* larval development (Lee et al. 1993; Wightman et al. 1993; Ha et al. 1996; Moss et al. 1997; Olsen and Ambros 1999; Reinhart et al. 2000; Slack et al. 2000; Hutvagner and Zamore 2002).

Regulatory targets for most animal miRNAs have not yet been identified. Prediction of plant miRNA targets has led to the proposal that many plant miRNAs function to clear from differentiating cells mRNAs encoding key transcription factors, thereby facilitating plant development and organogenesis (Rhoades et al. 2002). Con-

fidant computational prediction of animal miRNA targets has relied on experimental evidence to first narrow the number of candidate mRNAs (Lai 2002). Nonetheless, as seen for the plant miRNAs, the sequences of the animal miRNAs are generally highly conserved in evolution. For example, 91 of the 107 miRNAs cloned from mammals are detected in the pufferfish (*Fugu rubripes*) genome, implying that they have important functions preserved during vertebrate evolution (Lim et al. 2003).

The first step in a systematic approach to identifying the biological roles of miRNAs is to find the miRNA genes themselves. Because gene-prediction programs had not been developed to identify miRNAs in genomic sequence, miRNA gene identification has been primarily achieved by cloning the small RNAs from size-fractionated RNA samples, sometimes specifically enriching in miRNAs by first immunoprecipitating the miRNP complex or by using a cloning protocol specific for the 5' phosphate and 3' hydroxyl found on Dicer products (Lagos-Quintana et al. 2001, 2002, 2003; Lau et al. 2001; Lee and Ambros 2001; Llave et al. 2002a; Mourelatos et al. 2002; Park et al. 2002; Reinhart et al. 2002). Once small RNAs have been cloned, the challenge is to differentiate the authentic miRNAs from other RNAs present in the cell, particularly from endogenous siRNAs. Because both miRNAs and siRNAs are Dicer products and both can act to specify mRNA cleavage, miRNAs cannot be differentiated based on their chemical composition or their functional properties. However, miRNAs can be distinguished from siRNAs based on their biogenesis and evolutionary conservation: (1) They are 20- to 24-nt RNAs that derive from endogenous transcripts that can form local RNA hairpin structures; (2) these hairpins are processed such that a single miRNA molecule ultimately accumulates from one arm of each hairpin precursor molecule; (3) the sequences of the mature miRNAs and their hairpin precursors are usually evolutionarily conserved; and (4) the miRNA genomic loci are distinct from and usually distant from those of other types of recognized genes, although a few are found within predicted introns but not necessarily in the same orientation as the introns. Endogenous siRNAs differ in that (1) they derive from extended dsRNA, (2) each dsRNA precursor gives rise to numerous different siRNAs, (3) they generally display less sequence conservation, and (4) they often perfectly correspond to the sequences of known or predicted mRNAs, transposons, or regions of heterochromatic DNA (Aravin et al. 2001; Djikeng et al. 2001; Elbashir et al. 2001a; Lau et al. 2001; Llave et al. 2002a; Mochizuki et al. 2002; Reinhart and Bartel 2002; Reinhart et al. 2002). Regarding this fourth criterion, miRNAs can also perfectly correspond to sequences of their mRNA targets, but when they do, they still derive from loci distinct from those of their mRNA targets (Llave et al. 2002a,b; Reinhart et al. 2002). Because miRNAs are primarily distinguished based on their biogenesis and evolutionary conservation, the current norms for identification and validation of miRNA genes include experimental evidence for endogenous expression of the miRNA, coupled with evidence of a hairpin precursor, preferably

one that is evolutionarily conserved (Ambros et al. 2003).

Some miRNAs might be difficult to isolate by cloning, due to their low abundance or to biases in cloning procedures. Thus, computational identification of miRNAs from genomic sequences would provide a valuable complement to cloning. Recent advances have been made in the computational identification of ncRNA genes through comparative genomics, and complex algorithms have been developed to identify ncRNAs in general (Argaman et al. 2001; Rivas et al. 2001; Wassarman et al. 2001), as well as specific ncRNA families such as tRNAs and snoRNAs (Lowe and Eddy 1997, 1999).

In the present study, we describe a computational procedure to identify miRNA genes. By using this procedure, together with extensive sequencing of clones (3423 miRNA clones were sequenced), we have detected 30 additional miRNA genes, including previously unrecognized *lin-4* and *let-7* homologs. Extrapolation of the computational analysis indicates that miRNA gene identification in *C. elegans* is now approaching saturation, and that no more than 120 miRNA genes are present in this species. We also identify those genes with intriguing expression patterns during larval development and conditions of nutrient stress, and we show that most miRNAs are expressed at very high levels, with some present in as many copies per cell as the highly abundant U6 snRNA. This extensive census of worm miRNAs and their expression patterns establishes the general properties of this gene class and provides resources and tools for studies of miRNA function in nematodes and other organisms.

Results

Computational prediction of *C. elegans* miRNA genes

We developed a computational tool to specifically identify miRNAs that are conserved in two genomes and have the features characteristic of known miRNAs. To identify miRNAs in nematodes, the *C. elegans* genome was first scanned for hairpin structures with sequences that were conserved in *Caenorhabditis briggsae*. About 36,000 hairpins were found that satisfied minimum requirements for hairpin structure and sequence conservation. This procedure cast a sufficiently wide net to capture 50 of the 53 miRNAs previously reported to be conserved in the two species (Lau et al. 2001; Lee and Ambros 2001). These 50 published miRNA genes served as a training set for the development of a program called MiRscan, which was then used to assign scores to each of the 36,000 hairpins, evaluating them based on their similarity to the training set with respect to the following features: base pairing of the miRNA portion of the fold-back, base pairing of the rest of the fold-back, stringent sequence conservation in the 5' half of the miRNA, slightly less stringent sequence conservation in the 3' half of the miRNA, sequence biases in the first five bases of the miRNA (especially a U at the first position), a tendency toward having symmetric rather than asym-

metric internal loops and bulges in the miRNA region, and the presence of two to nine consensus base pairs between the miRNA and the terminal loop region, with a preference for 4–6 bp (Fig. 1A).

The distribution of MiRscan scores for the ~36,000 hairpins illustrated the ability of MiRscan to discern the 50 miRNA genes of the training set, which fell mostly in the high-scoring tail of the distribution (Fig. 2). Of the features evaluated by MiRscan, base-pairing potential and sequence conservation played primary roles in distinguishing known miRNAs (Fig. 1B). Some of the other conserved hairpins also scored highly; 35 had scores exceeding 13.9, the median score of the 58 known miRNAs (Fig. 2B). These 35 hairpins were carried forward as the top miRNA candidates predicted by MiRscan.

Molecular identification of miRNA genes

Our initial cloning and sequencing of small RNAs from mixed-stage *C. elegans* had identified 300 clones that represented 54 unique miRNA sequences (Lau et al. 2001). For the present study, this approach for identifying miRNAs was scaled-up ~10-fold. In an effort to identify miRNAs not normally expressed in mixed-stage logarithmically growing hermaphrodite worms, RNA was also cloned from populations of *him-8* worms, starved L1, and dauer worms. The *him-8* population was ~40% males, whereas the normal (N2) population was nearly all hermaphrodites (Broverman and Meneely 1994). Starved L1 and dauer worms are arrested in development at larval stages L1 and L3, respectively, with dauer worms having undergone morphological changes that enhance survival after desiccation or other harsh conditions.

As before, some clones matched *Escherichia coli*, the food source of the worms, others corresponded to fragments of annotated *C. elegans* RNAs. Nevertheless, 3423 clones were classified as miRNA clones (Table 1). Most of these represented the 58 miRNA genes previously identified in *C. elegans* (Lau et al. 2001; Lee and Ambros 2001). For example, *lin-4* was represented by 125 clones, *let-7* by 17 clones, and *mir-52* by 404 clones (Table 1). The remaining miRNA clones represented 23 newly identified miRNA loci.

In total, 80 loci were represented by cloned miRNAs (Table 1). Of these, 77 had the classical features of *C. elegans* miRNA genes, in that they had the potential to encode stereotypic hairpin precursor molecules with the 20- to 25-nt cloned RNAs properly positioned within an arm of the hairpin so as to be excised during Dicer processing, and their expression was manifested as a detectable Northern signal in the 20- to 25-nt range. Three other loci, *mir-41*, *mir-249*, and *mir-229*, were also included. The *mir-41* and *mir-249* RNAs were not detected on Northern blots but were still classified as miRNAs because these RNAs and their predicted hairpin precursors appear to be conserved in *C. briggsae*.

The *mir-229* locus was also classified as a miRNA gene, even though it appears to derive from an unusual fold-back precursor. Its precursor appears to be larger

Lim et al.

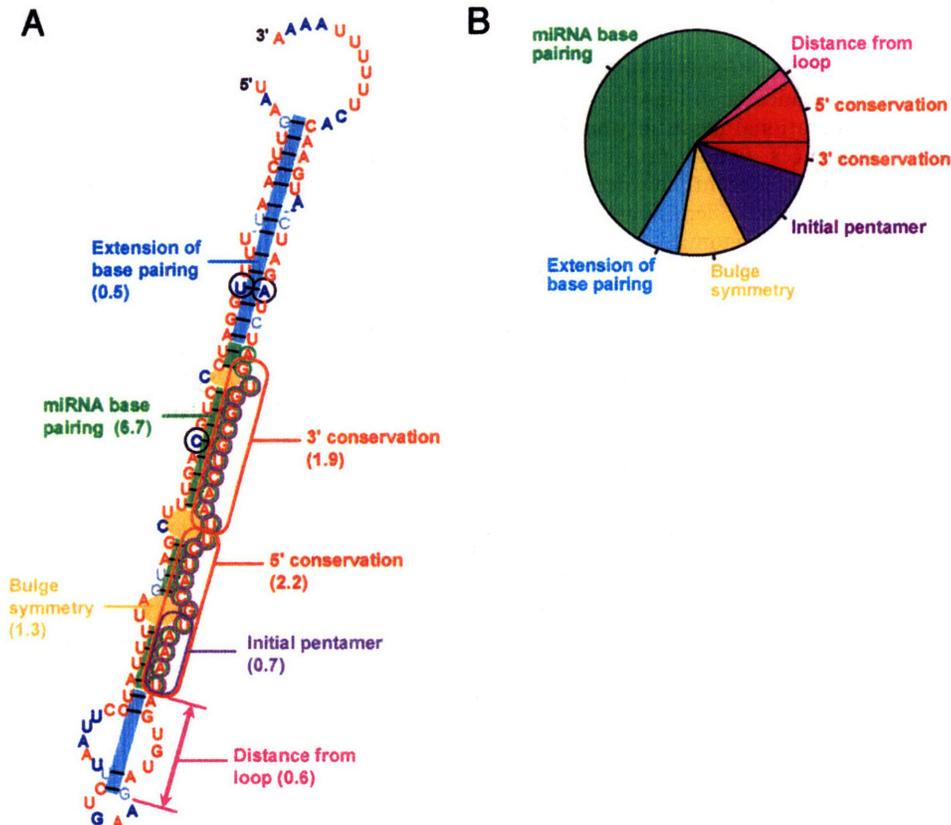


Figure 1. Criteria used by MiRscan to identify miRNA genes among aligned segments of two genomes. (A) The seven components of the MiRscan score for *mir-232* of *C. elegans/C. briggsae*. These components are annotated in the context of the MiRscan prediction for *mir-232*, with the residues of the predicted miRNA circled in purple and the residues of the validated miRNA (Table 2), circled in green. In parenthesis are the scores for each component, which were added together to give the total score of 13.9. MiRscan predictions are visualized within the consensus *C. elegans/C. briggsae* secondary structure, as generated by using ClustalW (Thompson et al. 1994) and Alidot (Hofacker and Stadler 1999). Shown is the *C. elegans* sequence with residues colored to indicate conserved sequence and pairing potential. Residues conserved in *C. briggsae* are red, residues that vary while maintaining their predicted paired or unpaired state are blue (with variant residues that maintain pairing also circled in black), and residues that maintain neither sequence nor pairing are in gray. (B) Estimated relative importance of each MiRscan criterion. Estimates were based on the relative entropy between the training set of 50 previously identified nematode miRNAs and the background set of ~36,000 potential stem loops. Because pairing and conservation were used to identify the potential stem loops, the total contributions of these types of criteria for distinguishing miRNA genes from non-protein-coding genomic sequence were underestimated. Likewise, the total contribution of the distance from the loop was underestimated because only those candidates 2–9 bp from the loop were evaluated.

than normal, possibly because of an extra 35-nt stem loop protruding from the 3' arm of the precursor stem loop (Supplementary Fig. 1). Nonetheless, miR-229 was detectable as a ~25- to 26-nt species on Northern blots, and accumulation of its presumed precursor increased in the *dcr-1* mutant, suggesting that Dicer processes this precursor despite the unusual predicted secondary structure (Supplementary Fig. 1). Furthermore, *mir-229* is only 400 bp upstream of a previously recognized miRNA gene cluster, including *mir-64*, *mir-65*, and *mir-66*. miR-229 also has significant sequence identity with the miRNAs of this cluster. We provisionally classified *mir-229* as a miRNA and a member of this *C. elegans* cluster. Greater confidence would be warranted if its unusual precursor structure were conserved in another species. A weakly homologous cluster of two potential miRNAs was found in *C. briggsae*, but neither of the predicted *C.*

briggsae homologs appeared to have an unusual precursor resembling that of miR-229.

Validation of computationally predicted miRNAs

Of the 23 newly cloned miRNAs, 20 received MiRscan scores, and these scores are indicated in yellow in Figure 2B. The other three were not scored because orthologous sequences in *C. briggsae* were not identified. A Mann-Whitney test showed that the distribution of scores for these recently cloned miRNAs was not significantly different from that of the previously cloned miRNAs. Because the recently cloned miRNAs were not known during the development of MiRscan, their high scores gave added assurance that MiRscan was not over-fitting its training set. Ten of the 23 newly cloned miRNAs were

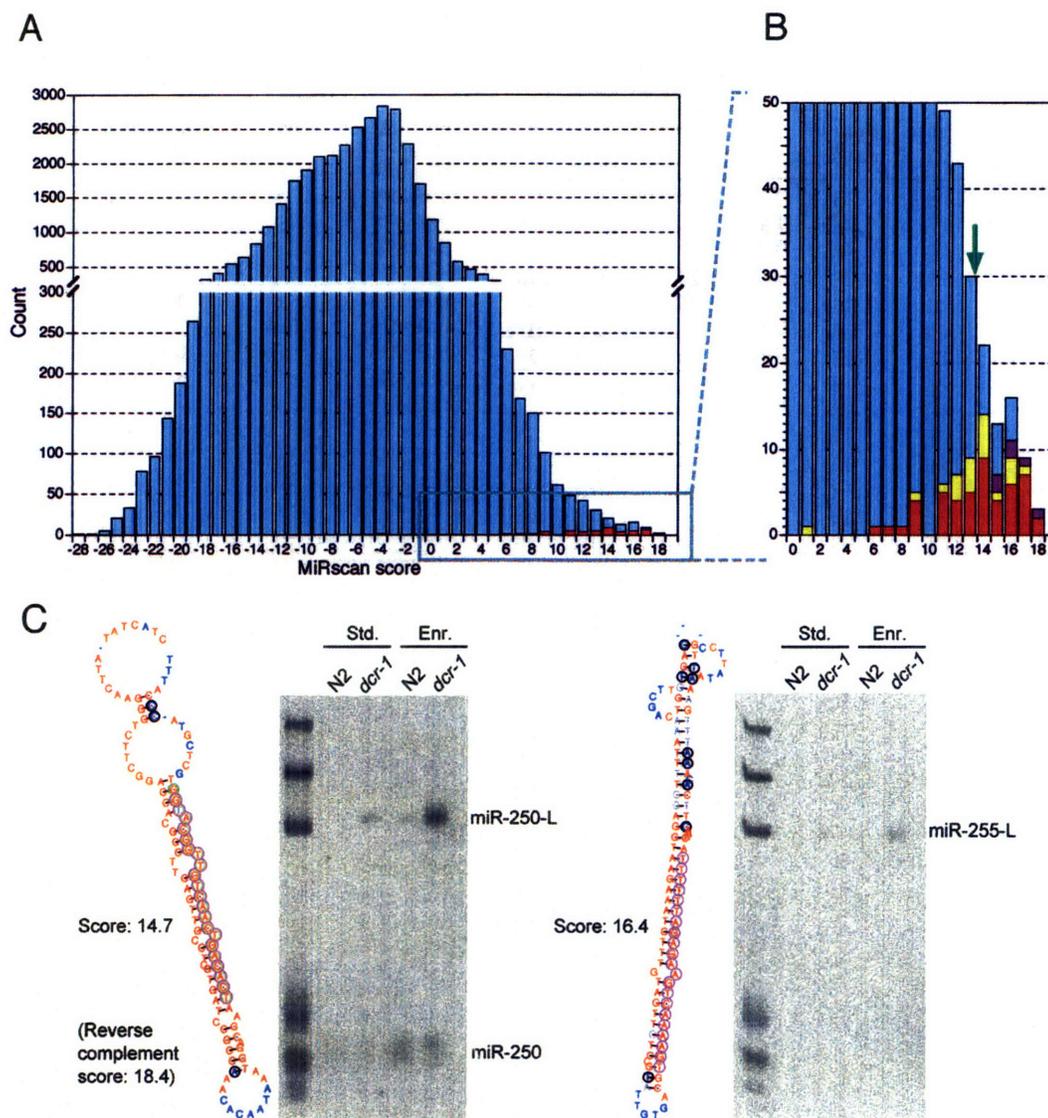


Figure 2. Computational identification of miRNA genes. (A) The distribution of MiRscan scores for 35,697 *C. elegans* sequences that potentially form stem loops and have loose conservation in *C. briggsae*. Note that the Y-axis is discontinuous so that the scores of the 50 previously reported miRNA genes that served as the training set for MiRscan can be more readily seen (red). Scores for these 50 genes were jackknifed to prevent inflation of their values because of their presence in the training set. (B) An expanded view of the high-scoring tail of the distribution. This view captures 49 of the 50 genes of the training set (red). The median score of the 58 previously reported miRNA loci that satisfy the current criteria for designation as miRNA genes (Ambros et al. 2003) is 13.9 (green arrow). Note that this median score was the midpoint between the scores of the 29th and 30th highest-scoring loci of the 50-member training set; namely, it was designated the median score after including the 8 previously reported miRNA genes that were not in the training set because they were lost during the identification of conserved hairpins, usually because they lacked sufficient *C. briggsae* homology. Scores of genes validated by cloning are indicated (yellow), as are scores of six genes that have not yet been cloned but were verified by Northern analysis (purple). (C) Examples of miRNA genes identified by MiRscan with the Northern blots that served to validate them. Stem-loops were annotated as in Figure 1A, except the DNA rather than RNA sequence is depicted. The Northern blots show analysis of RNA from either wild-type (N2) or *dcr-1* worms, isolated using either our standard protocol (Std.) or an additional polyethylene glycol precipitation step to enrich for small RNAs (Enr.). Homozygous worms of the *dcr-1* population have reduced Dicer activity, increasing the level of miRNA precursors (e.g., miR-250-L and miR-255-L), which facilitated the validation of miRNA loci, especially those for which the mature miRNA was not detected (e.g., miR-255). RNA markers (left lane) are 18, 21, 24, 60, 78, and 119 nt. The miR-250 stem loop shown received a MiRscan score of 14.7. The miR-250 reverse complement received an even greater score of 18.4, but was not detected by Northern analysis. Thus, the predicted *mir-250* gene was assigned the score of the higher-scoring, although incorrect, alternative stem loop (Table 1; Fig. 2B).

among the set of 35 high-scoring miRNA gene candidates and served to validate these 10 candidates.

The remaining 25 candidate miRNAs that had not been cloned were tested by Northern blots. RNA from

Lim et al.

Table 1. Cloning frequency and MiRscan scores of *Caenorhabditis elegans* miRNAs

miRNA	MiRscan score	Number of sequenced clones					total
		mixed stage	dauer	starved L1	<i>him-8</i>		
let-7 RNA	13.8	15	0	0	2	17	
lin-4 RNA	15.8	48	46	4	27	125	
miR-1	14.7	43	17	7	9	76	
miR-2	6.2	138	46	20	9	213	
miR-34	14.1	13	25	5	9	52	
miR-35	14.4	23	0	1	2	26	
miR-36	14.6	21	0	1	5	27	
miR-37	9.6	8	0	1	2	11	
miR-38	8.9	10	0	1	0	11	
miR-39	9.5	11	0	0	1	12	
miR-40	15.4	12	0	4	2	18	
miR-41	12.0	2	0	0	0	2	
miR-42	9.5	10	4	3	1	18	
miR-43	17.5	8	1	9	0	18	
miR-44/45	16.6/17.4	22	3	3	4	32	
miR-46	11.3	14	11	9	3	37	
miR-47	16.5	19	7	4	5	35	
miR-48	12.0	52	1	0	8	61	
miR-49	13.1	1	0	1	1	3	
miR-50	14.6	10	16	5	1	32	
miR-51	12.0	16	5	2	2	25	
miR-52	11.6	287	70	18	29	404	
miR-53	12.4	20	6	1	4	31	
miR-54	9.4	49	40	9	13	111	
miR-55	13.8	47	32	16	15	110	
miR-56	NS	40	16	9	6	71	
miR-57	12.1	31	11	8	3	53	
miR-58	17.5	181	51	27	31	290	
miR-59	18.5	1	0	0	0	1	
miR-60	14.1	20	6	3	7	36	
miR-61	13.7	8	5	1	3	17	
miR-62	15.1	4	4	6	0	14	
miR-63	NS	7	1	0	1	9	
miR-64	NS	11	4	8	3	26	
miR-65	7.4	22	7	3	2	34	
miR-66	NS	68	25	6	7	106	
miR-67	16.8	3	0	0	0	3	
miR-70	11.6	11	8	3	6	28	
miR-71	17.9	53	72	23	22	170	
miR-72	NS	49	22	10	9	90	
miR-73	11.3	13	7	1	1	22	
miR-74	17.9	35	12	6	7	60	
miR-75	12.6	14	3	2	2	21	
miR-76	14.9	1	2	6	3	12	
miR-77	14.2	17	3	0	2	22	
miR-78	NS	5	1	1	0	7	
miR-79	14.2	14	3	3	3	23	
miR-80	17.1	121	27	20	17	185	
miR-81	18.8	32	24	6	12	74	
miR-82	16.3	36	12	6	11	65	
miR-83	15.2	12	12	2	8	34	
miR-84	-3.3	12	2	1	4	19	
miR-85	17.5	10	0	0	12	22	
miR-86	16.3	46	57	30	17	150	
miR-87	16.7	1	0	0	0	1	
miR-88	-7.9					0	
miR-90	14.0	5	37	14	9	65	

(continued)

Table 1. Continued

miRNA	MiRscan score	Number of sequenced clones					total
		mixed stage	dauer	starved L1	<i>him-8</i>		
miR-124	15.7	7	16	7	5	35	
miR-228	17.5	1	13	8	3	25	
miR-229	NS	2	1	0	0	3	
miR-230	16.8	0	0	0	1	1	
miR-231	14.1	1	2	0	0	3	
miR-232	13.8	4	7	2	1	14	
miR-233	16.4	1	8	4	0	13	
miR-234	14.3	0	0	1	0	1	
miR-235	1.9	5	21	1	8	35	
miR-236	16.8	3	6	2	1	12	
miR-237	11.9	3	0	0	0	3	
miR-238	14.0	0	4	1	0	5	
miR-239a	12.7	4	0	0	1	5	
miR-239b	13.6					0	
miR-240	12.5	0	0	0	1	1	
miR-241	14.9	7	0	0	3	10	
miR-242	9.9	0	0	1	1	2	
miR-243	NS	1	0	1	0	2	
miR-244	13.4	0	2	5	0	7	
miR-245	13.8	0	1	0	0	1	
miR-246	12.8	0	0	0	1	1	
miR-247	NS	0	2	0	0	2	
miR-248	14.6	0	2	0	0	2	
miR-249	13.7	0	2	1	0	3	
miR-250	18.4					0	
miR-251	15.5					0	
miR-252	17.7					0	
miR-253	16.9					0	
miR-254	15.7					0	
miR-255	16.4					0	
Total clones		1821	851	363	388	3423	

A total of 3423 clones from logarithmically growing mixed-stage worms and worms from the indicated stages or mutant (dauer, starved L1, and *him-8*) represented 79 different miRNAs (and 80 different miRNA genes, because the miR-44/45 miRNA appears to be encoded at two loci). Genes not represented in the set of ~36,000 stem loops did not receive scores (NS). Note that the previously reported miR-68 clone is not included. This RNA was not detected on Northern blots, and neither it nor its predicted precursor appears to be conserved in another species. Accordingly, it is now classified as an endogenous siRNA. Two other *C. elegans* loci previously thought to encode miRNAs (*mir-69* and *mir-89*) also do not satisfy the current criteria for classification as miRNA genes (Ambros et al. 2003) and were not considered during the course of this study. One previously reported gene, *mir-88*, was not represented in our set of sequenced clones but is detected on Northern blots as a ~22-nt RNA (V. Ambros, pers. comm.) and thus satisfies the current criteria for classification as an miRNA gene.

dcr-1 worms was included on the blots to enhance detection of precursor hairpins. Dicer-dependent processing of ~70-nt precursors was detected for six candidates (as shown for miR-250 and miR-255; Fig. 2C), and ~22-nt miRNAs were detected for miR-250, miR-251, and miR-252. Despite prolonged exposure times and enrichment for small RNA by size fractionation, the Northern signals were generally weak, perhaps explaining why

these miRNAs were missed in the current set of 3423 sequenced miRNA clones.

To investigate whether these miRNAs eventually would have been identified after further cloning and sequencing of our cDNA library of small RNA sequences, a PCR assay was used to detect the presence of these miRNAs in the library. By using a primer specific to the 3' segment of the predicted miRNA, together with a second primer corresponding to the adapter sequence attached to the 5' terminus of all the small RNAs, the 5' segment of the miRNA was amplified, cloned, and sequenced. This procedure validated five of the six predicted miRNAs for which at least a precursor could be detected on Northern blots, including two of the candidates (miR-253 and miR-254) for which a mature ~22-nt RNA was not detected on Northern blots. In addition, it identified the 5' terminus of these five miRNAs, which is difficult to achieve with confidence when using only bioinformatics and hybridization.

Combining the cloning and expression data, 16 of the 35 computationally identified candidates were validated (10 from cloning, five from Northern blots plus the PCR assay, and one from Northern blots only, which validated the precursor but did not identify the mature miRNA). Of the remaining 19 candidates, four could be readily classified as false positives. They appear to be nonannotated larger ncRNA genes, in that probes designed to hybridize to these candidates hybridized instead to high-molecular-weight species that remained constant in the samples from *dcr-1* worms. The remaining 15 new candidates with high MiRscan scores but without any Northern signal might also be false positives, or they might be authentic miRNAs that are expressed at low levels or in only very specific cell types or circumstances. Considering the extreme case in which all the nonvalidated candidates are false positives, the minimum specificity of MiRscan for the *C. elegans/C. briggsae* analysis can be calculated as $(29 + 16)/(29 + 35)$, or 0.70, at a sensitivity level that detects half of the 58 previously known miRNAs. A summary of the miRNA genes newly identified by validating computational candidates (16 genes) or by cloning alone (13 genes) is shown in Table 2, and predicted stem-loop precursors are shown in Supplemental Material. Table 2 also includes one additional gene, *mir-239b*, which was identified based on its homology with *mir-239a* and its MiRscan score of 13.6.

Evolutionary conservation of miRNAs

The 88 *C. elegans* miRNA genes identified to this point were grouped into 48 families, each comprising one to eight genes (data not shown). Within families, sequence identity either spanned the length of the miRNAs or was predominantly at their 5' terminus. All but two of these families extended to the miRNAs of *C. briggsae*. The two families without recognizable *C. briggsae* orthologs each comprised a single miRNA (miR-78 and miR-243). Thus, nearly all (>97%) of the *C. elegans* miRNAs identified had apparent homologs in *C. briggsae*, and all but six of these *C. elegans* miRNAs (miR-72, miR-63, miR-

64, miR-66, miR-229, and miR-247) had retained at least 75% sequence identity to a *C. briggsae* ortholog. Of the 48 *C. elegans* miRNA families, 22 also had representatives among the known human miRNA genes (Fig. 3). In that these 22 families included 33 *C. elegans* genes, it appears that at least a third (33/88) of the *C. elegans* miRNA genes have homologs in humans and other vertebrates.

Developmental expression of miRNAs

The expression of 62 miRNAs during larval development was examined and compiled together with previously reported expression profiles (Lau et al. 2001) to yield a comprehensive data set for the 88 *C. elegans* miRNAs (Fig. 4). RNA from wild-type embryos, the four larval stages (L1 through L4), and young adults was probed, as was RNA from *glp-4* (*bn2*) young adults, which are severely depleted in germ cells (Beanan and Strome 1992). Nearly two thirds of the miRNAs appeared to have constitutive expression during larval development (Fig. 4A). These miRNAs might still have differential expression during embryogenesis, or they might have tissue-specific expression, as has been observed for miRNAs of larger organisms in which tissues and organs can be more readily dissected and examined (Lee and Ambros 2001; Lagos-Quintana et al. 2002; Llave et al. 2002a; Park et al. 2002; Reinhart et al. 2002).

Over one third of the miRNAs had expression patterns that changed during larval development (Fig. 4B,C), and there were examples of miRNA expression initiating at each of the four larval stages (Fig. 4B). Expression profiles for miR-48 and miR-241 (which are within 2 kb of each other in the *C. elegans* genome) were similar to those previously reported for *let-7* RNA and miR-84 (Fig. 4B; Reinhart et al. 2000; Lau et al. 2001). In fact, these four miRNAs appear to be paralogs, with all four miRNAs sharing the same first eight residues (Fig. 3). Another newly identified miRNA, miR-237, is a paralog of the other canonical stRNA, *lin-4* RNA (Fig. 3), although miR-237 exhibited an expression pattern distinct from *lin-4* RNA (Fig. 4E). The existence of these paralogs, as well as other families of miRNAs with expression initiating at the different stages of larval development, supports the idea that *lin-4* and *let-7* miRNAs are not the only stRNAs with important roles in the *C. elegans* heterochronic pathway.

Expression usually remained constant once it initiated, as has been seen for *lin-4* and *let-7* miRNA expression (Fig. 4A,B). Exceptions to this trend included the miRNAs of the *mir-35-mir-41* cluster, which were expressed transiently during embryogenesis (Lau et al. 2001); miR-247, which was expressed transiently in larval stage 3 (and dauer); and miR-248, which was most highly expressed in dauer (Fig. 4C,D). miR-234 was expressed in all stages, but expression was highest in both L1 worms (which had been starved shortly before harvest to synchronize the worm developmental staging) and dauer worms, suggesting that this miRNA might be induced as a consequence of nutrient stress.

Lim et al.

Table 2. Newly identified *Caenorhabditis elegans* miRNA genes

miRNA gene	ID method	miRNA sequence	miRNA length (nt)	<i>C. briggsae</i> homology	Fold-back arm	Chr.	Distance to nearest gene	
<i>mir-124</i>	MS, C, N	UAAGGCACGCGGUGAAUGCCA	21	+++	3'	IV	within intron of C29E6.2	(s)
<i>mir-228</i>	MS, C, N	AAUGGCACUGCAUGAAUUCACGG	21–24	+++	5'	IV	0.2 kb downstream of T12E12.5	(as)
<i>mir-229</i>	C, N	AAUGACACUGGUUAUCUUUCAUCG	25–27	–	5'	III	0.4 kb upstream of <i>mir-64</i>	(s)
<i>mir-230</i>	MS, C, N	GUUUUAGUUGUGCGACCAGGAGA	23	++	3'	X	0.4 kb downstream of F13D11.3	(as)
<i>mir-231</i>	MS, C, N	UAAGCUCGUGAUCAACAGGCAGAA	23–24	++	3'	III	10.4 kb upstream of <i>lin-39</i>	(s)
<i>mir-232</i>	C, N	UAAAUGCAUCUUAACUGCGGUGA	23–24	+++	3'	IV	1.1 kb downstream of F13H10.5	(as)
<i>mir-233</i>	MS, C, N	UUGAGCAAUGCGCAUGUGCGGGA	19–23	+++	3'	X	within intron of W03G11.4	(s)
<i>mir-234</i>	MS, C, N	UUUUUGCUCGAGAAUACCCUU	21	+++	3'	II	1.5 kb downstream of Y54G11B.1	(as)
<i>mir-235</i>	C, N	UAUUGCACUCUCCCCGGCCUGA	22	+	3'	I	0.6 kb upstream of T09B4.7	(s)
<i>mir-236</i>	MS, C, N	UAAUACUGUCAGGUAUUGACGCU	21–25	+++	3'	II	0.3 kb downstream of C52E12.1	(as)
<i>mir-237</i>	C, N	UCCUGAGAAUUCUGAACAGCUU	23–24	+	5'	X	3.4 kb upstream of F22F1.2	(as)
<i>mir-238</i>	MS, C, N	UUUGUACUCCGAGUCCAUUCAGA	21–23	++	3'	III	2.0 kb upstream of <i>mir-80</i>	(s)
<i>mir-239a</i>	C, N	UUUGUACUACACAUAGGUACUGG	22–23	++	5'	X	6.0 kb upstream of C34E11.1	(s)
<i>mir-239b</i>	H	UUUGUACUACACAAAAGUACUGG	n.d.	++	5'	X	7.0 kb upstream of C34E11.1	(s)
<i>mir-240</i>	C, N	UACUGGCCCCAAAUCUUCGCU	22	++	3'	X	1.7 kb upstream of C39D10.3	(s)
<i>mir-241</i>	MS, C, N	UGAGGUAGGUGCGAGAAUUGA	21	++	5'	V	1.8 kb upstream of <i>mir-48</i>	(s)
<i>mir-242</i>	C, N	UUGCGUAGGCCUUUGCUUGCA	21	++	5'	IV	0.9 kb downstream of <i>nhf-78</i>	(as)
<i>mir-243</i>	C, N	CGGUACGAUCGCGCGGGUAUUC	22–23	–	3'	IV	1.0 kb upstream of R08C7.1	(s)
<i>mir-244</i>	C, N	UCUUUGGUUGUACAAAGUGGUUAG	23–25	+++	5'	I	1.6 kb downstream of T04D1.2	(as)
<i>mir-245</i>	C, N	AUUGGUCCCCUCCAAGUAGCUC	22	+++	3'	I	1.9 downstream of F55D12.1	(s)
<i>mir-246</i>	C, N	UUACAUGUUUCGGUAGGAGCU	22	++	3'	IV	0.4 kb downstream of ZK593.8	(s)
<i>mir-247</i>	C, N	UGACUAGAGCCUAUUCUCUUCUU	22–23	–	3'	X	1.9 kb upstream of C39E6.2	(as)
<i>mir-248</i>	MS, C, N	UACACGUGCAGGUAUACGCUCA	23	++	3'	X	within intron of AH9.3	(s)
<i>mir-249</i>	C	UCACAGGACUUUGAGCGUUGC	22–23	++	3'	X	2.7 kb upstream of Y41G9A.6	(s)
<i>mir-250</i>	MS, N, PCR	UCACAGUCAACUGUUGGCAUGG	–22	++	3'	V	0.1 kb downstream of <i>mir-61</i>	(s)
<i>mir-251</i>	MS, N, PCR	UUAAGUAGUGGUGCCGCUUUAUU	–24	+++	5'	X	0.2 kb downstream of F59F3.4	(as)
<i>mir-252</i>	MS, N, PCR	UAGUAGUAGUGCCGACAGUAAAC	–23	+++	5'	II	1.8 kb downstream of VW02B12L.4	(as)
<i>mir-253</i>	MS, D, PCR	CACACCUCACUACACUGACC	n.d.	++	5'	V	within intron of F44E7.5	(s)
<i>mir-254</i>	MS, D, PCR	UGCAAUUCUUUCGACUGUAGG	n.d.	++	3'	X	within intron of ZK455.2	(s)
<i>mir-255</i>	MS, D	–	n.d.	–	–	–	1.5 kb upstream of F08F3.9	(as)

For predicted stem-loop precursors, see Supplementary Fig. 2. Genes were identified and validated as indicated in the ID method column: MS, candidate gene had high MiRscan score (Table 1); C, miRNA was cloned and sequenced (Table 1); N, expression of the mature miRNA was detectable on Northern blots; D, the miRNA stem-loop precursor was detected on Northern blots and enriched in RNA from *dcr-1* animals, but the mature miRNA was not detected; PCR, targeted PCR amplification and sequencing detected the miRNA in a library of *C. elegans* small RNAs; H, the locus was closely homologous to that of a validated miRNA. For the miRNAs cloned and sequenced, some miRNAs were represented by clones of different lengths, due to heterogeneity at the miRNA 3' terminus. The observed range in length is indicated, and the sequence of the most abundant length is shown. For the RNAs that have not been cloned, the 5' terminus was determined by the PCR assay, but the 3' terminus was not determined. For *mir-250*, *mir-251*, and *mir-252*, the length of the miRNA sequence shown was inferred from the Northern blots; for other miRNAs not cloned, the length was not determined (n.d.). For *mir-254*, the PCR assay detected –22-nt RNAs from both sides of the fold-back, representing both the miRNA and the miRNA*. Their relative positions within the precursor suggest that the RNA from the 5' arm is 22 nt and the RNA from the 3' arm is 23 nt. The RNA from the 3' arm was chosen as the miRNA because of its similarity to the human miR-19 gene family. The miR-255 gene is known only as the precursor, a conserved stem loop with Dicer-dependent processing (Fig. 2b). Comparison to *C. briggsae* shotgun traces from the *C. briggsae* Sequencing Consortium [obtained from www.ncbi.nlm.nih.gov] revealed miRNA orthologs with 100% sequence identity (+++) and potential orthologs with >90% (++) and >75% (+) sequence identity. To indicate the genomic loci of the genes, the chromosome (Chr.), distance to nearest annotated gene, and the orientation relative to that gene, sense (s) or antisense (as), are specified.

Molecular abundance of miRNAs

The very high cloning frequency of certain miRNAs (e.g., miR-52, represented by >400 clones) raised the question as to the molecular abundance of these and other miRNA species. In addition, there was the question of whether the actual molecular abundance of miRNAs in nematodes was proportionally reflected in the numbers of clones sequenced. To address these questions, quantitative Northern blots were used to examine the molecular abundance of 12 representative miRNAs, picked so as to span the range of frequently and rarely cloned sequences and differing 3' and 5' terminal residues (Fig. 5).

To determine the molecular abundance of these 12 miRNAs in the adult worm soma, the hybridization signals for RNA from a known number of *glp-4* young adult

worms were compared with standard curves from chemically synthesized miRNAs (Fig. 5; Hutvagner and Zamore 2002). Accounting for RNA extraction yields and dividing the number of miRNA molecules per worm by the total number of cells in the worms, yielded averages of up to 50,000 molecules per cell, with the most abundant miRNAs as plentiful as the U6 snRNA of the spliceosome (Fig. 5C). These are much higher numbers than those for the typical worm mRNAs, estimated to average ~100 molecules per cell for the 5000 most highly expressed genes in the cell. [This estimate was calculated based on our yield of 20 pg total RNA per worm cell, assuming that the 5000 most highly expressed genes have mRNAs averaging 2 kb in length and represent 3% of the total RNA in an adult worm; it was consistent with estimates based on hybridization kinetics of

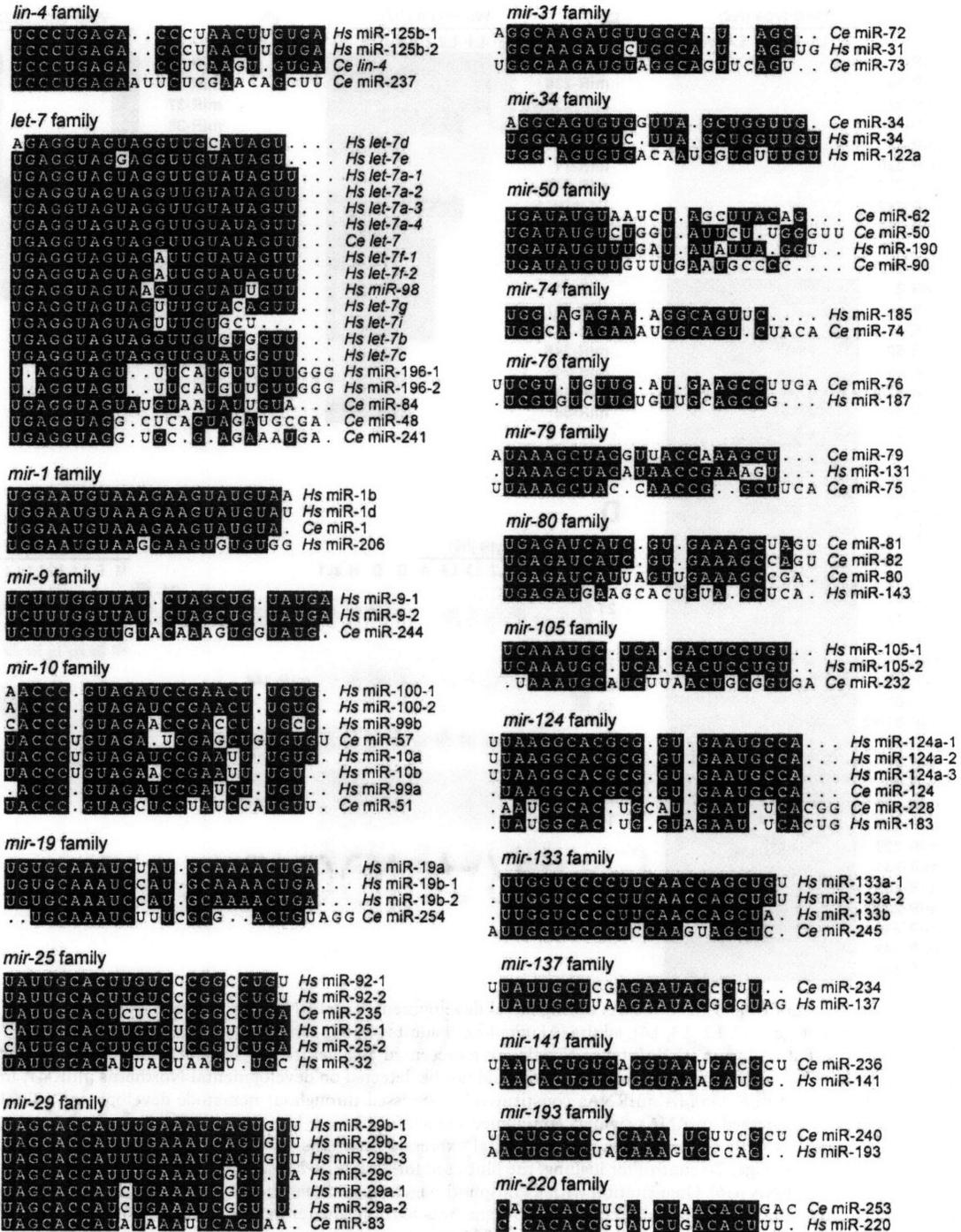


Figure 3. Alignments of *C. elegans* and human miRNA sequences that can be grouped together in families. Human miRNAs (*Hs*) are those identified in human cells (Lagos-Quintana et al. 2001; Mourelatos et al. 2002) or are orthologs of miRNAs identified in other vertebrates (Lagos-Quintana et al. 2002, 2003; Lim et al. 2003).

mRNAs from mouse tissues (Hastie and Bishop 1976.) Perhaps high concentrations of miRNAs are needed to saturate the relevant complementary sites within the target mRNAs, which might be recognized with low affinity because of the noncanonical pairs or bulges that

appear to be characteristic of the animal miRNA-target interactions.

Because these numbers represent molecular abundance averaged over all the cells of the worm, including cells that might not be expressing the miRNA, there are

Lim et al.

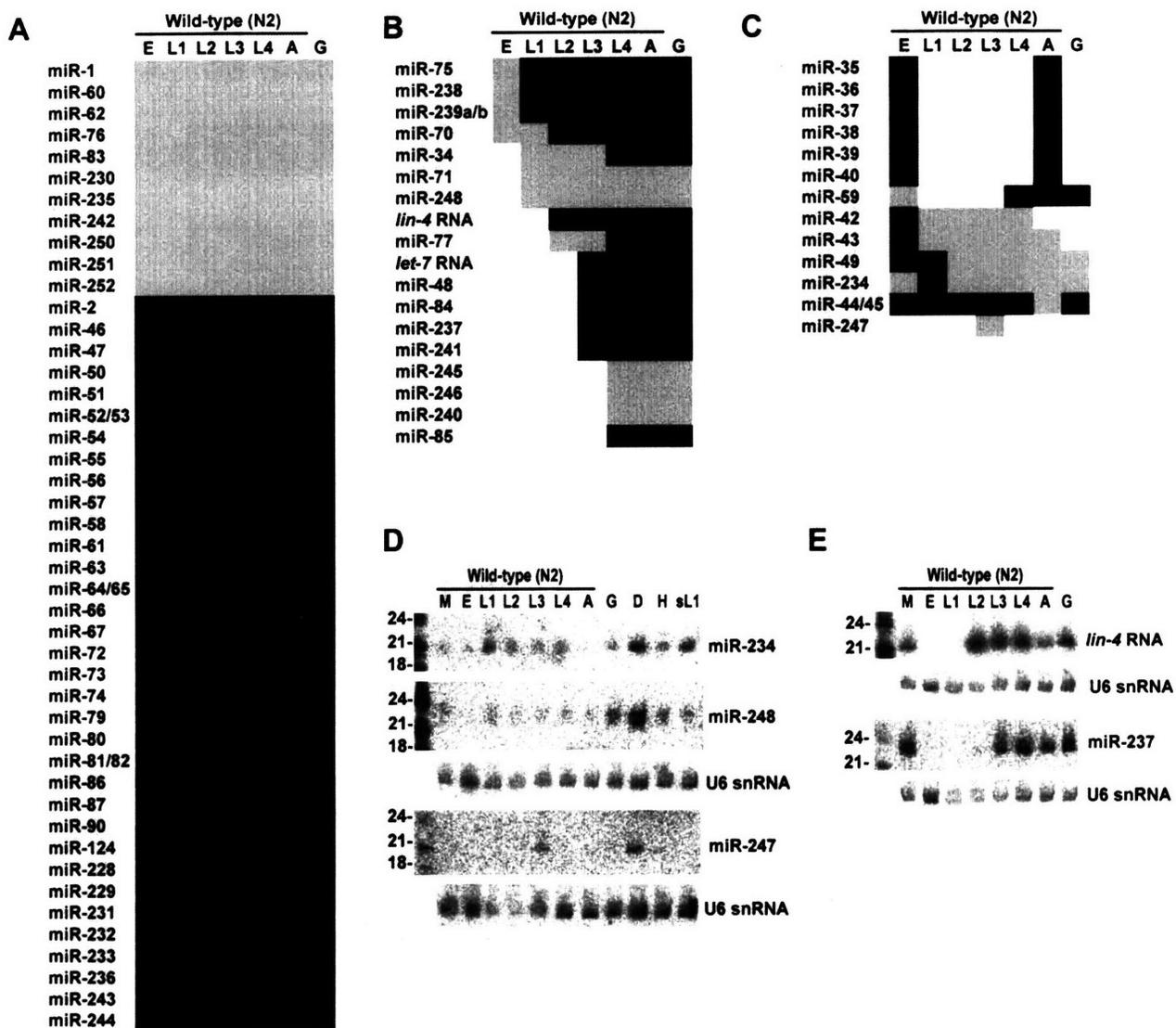


Figure 4. Expression of *C. elegans* miRNAs during larval development. Total RNA was analyzed from mixed-stage N2 worms (M), embryos (E), larval stages (L1, L2, L3, L4), adults (A), *glp-4(bn2)* adults (G), N2 dauers (D), mixed-stage *him-8(e1489)* worms (H), and N2 starvation-arrested L1 larvae (sL1). Intense signals are represented as black rectangles and faint signals are represented as gray rectangles. Of the 87 *C. elegans* miRNAs identified, 6 could not be detected on developmental Northern blots (miR-41, miR-78, miR-249, miR-253, miR-254, and miR-255). (A) miRNAs constitutively expressed throughout nematode development. (B) *lin-4* and *let-7*, and similarly expressed miRNAs, which commence expression during larval development and remain expressed through adulthood. (C) miRNAs with discontinuous developmental expression patterns. (D) Northern analysis of miRNAs with enhanced expression in the dauer stage. To control for loading, the blot used for both miR-234 and miR-248 and the blot used for miR-247 were reprobated for the U6 snRNA (U6). Quantitation with a PhosphorImager showed that the lane-to-lane variation in U6 signal was as great as threefold. Normalizing to the U6 signal, the miR-248 signal was fourfold greater in dauer than in most other stages, except for *glp-4* adults, in which it was twofold greater, whereas the miR-234 signal was highest in dauer and L1, with a signal in these stages about twofold greater than the average of the other stages. (E) Northern analysis of the *lin-4* RNA and its paralog, miR-237.

likely to be some cells that express even more molecules of the miRNA. To examine the abundance in a single cell type, HeLa RNA was probed for representative human miRNAs, yielding a similar range of molecular abundance (Fig. 5C). The high number of miRNA molecules in human cells increases the mystery as to why miRNAs had gone undetected for so long, which raises the question of whether other classes of highly expressed

ncRNAs might yet remain to be discovered. A recent large-scale analysis of full-length cDNAs from mouse indicates the possible existence of hundreds or thousands of expressed ncRNAs in vertebrates (Okazaki et al. 2002).

To address the extent to which the actual molecular abundance of miRNAs in nematodes is proportionally reflected in the numbers of clones sequenced, the abun-

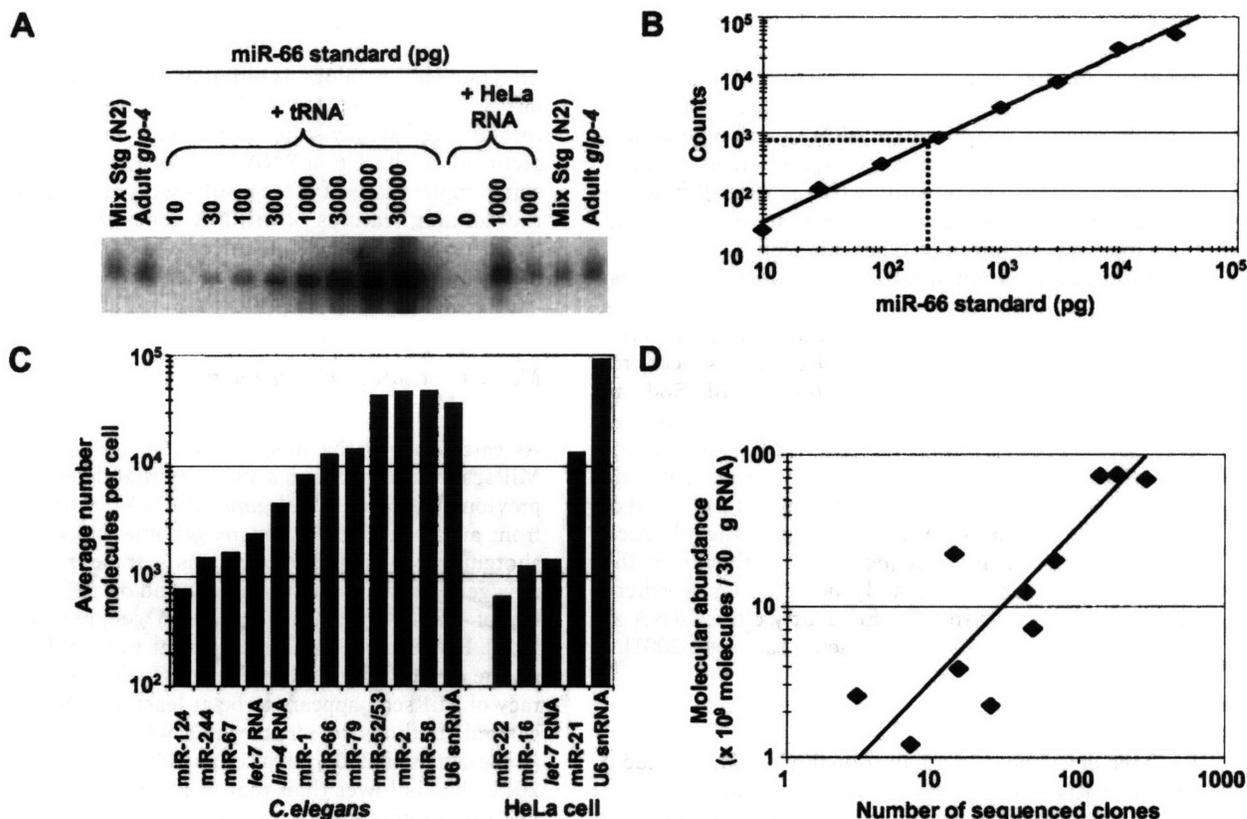


Figure 5. Quantitative analysis of miRNA expression. (A) Northern blot used to quantify the abundance of miR-66. RNA prepared from the wild-type (N2) mixed-stage worms used in cloning and from *glp-4(bn2)* young adult worms were run in duplicate with a concentration course of synthetic miRNA standard. The signal from the standard did not change when total RNA from HeLa cells replaced *E. coli* tRNA as the RNA carrier, showing that the presence of other miRNAs did not influence membrane immobilization of the miRNA or hybridization of the probe. (B) Standard curve from quantitation of miR-66 concentration course. The best fit to the data is a line represented by the equation $y = 3.3x^{0.96}$ ($R^2 = 0.99$). Interpolation of the average signal in the *glp-4* lanes indicates that the *glp-4* samples contain 240 pg of miR-66 (broken lines). (C) Molecular abundance of miRNAs and U6 snRNA. Amounts of the indicated RNA species in the *glp-4* samples were determined as shown in A and B. The average number of molecules per cell was then calculated considering the number of animals used to prepare the sample, and the yield of a radiolabeled miRNA spiked into the preparation at an early stage of RNA preparation. Analogous experiments were performed to determine the amounts of the indicated human miRNAs in HeLa RNA samples. (D) Correlation between miRNA molecular abundance and cloning frequency. The number of molecules in the mixed-stage RNA samples was determined as described for the *glp-4* samples and then plotted as a function of the number of times the miRNAs was cloned from this mixed-stage population (Table 1). The line is best fit to the data and is represented by the equation $y = 0.32x$ ($R^2 = 0.78$).

dance of the miRNA within the mixed-stage RNA preparation was compared with the number of clones generated from that preparation (Fig. 5D). The strong positive correlation observed between the molecular abundance and the number of times the miRNAs were cloned indicated that systematic biases in the cloning procedure were not major. At most, these miRNAs were over- or underrepresented fivefold in the sequenced set relative to their actual abundance as measured by quantitative Northern blots. We cannot rule out the possibility that certain miRNAs not yet cloned might be refractory to our cloning procedure, for example, because of a propensity to form secondary structures that preclude adaptor ligation reactions. Nonetheless, on the whole, the cloning frequencies can be used to approximate the molecular abundance of the miRNAs, and we have no reason to

suspect that the set of miRNAs identified by cloning differs in any substantive way, other than an overall higher steady-state expression level, from the complete set of *C. elegans* miRNAs.

Other endogenous ~22-nt RNAs of *C. elegans*

Of the 4078 *C. elegans* clones, a large majority represented authentic miRNAs (3423 clones, Table 1). The next most abundant class represented degradation fragments of larger ncRNAs, such as tRNA and rRNA (447 clones) and introns (18 clones). The remaining clones represented potential Dicer products that were not classified as miRNAs. Some corresponded to sense (18 clones) or antisense (23 clones) fragments of known or predicted mRNAs and might represent endogenous

Lim et al.

siRNAs. Others (143 clones) corresponded to regions of the genome not thought to be transcribed; these might represent another type of endogenous siRNAs, known as heterochromatic siRNAs (Reinhart and Bartel 2002). The possible roles of the potential siRNAs and heterochromatic siRNAs in regulating gene expression are still under investigation. The remaining clones were difficult to classify because they matched more than one locus, and their loci were of different types (six clones).

A fourth class of potential Dicer products (38 clones, representing 14 loci) corresponded to miRNA precursors but derived from the opposite arm of the hairpin than the more abundantly expressed miRNA, as has been reported previously for miR-56 in *C. elegans*, miR156d and miR169 in plants, and several vertebrate miRNAs (Lau et al. 2001; Lagos-Quintana et al. 2002, 2003; Mourelatos et al. 2002; Reinhart et al. 2002). Our current data add another 13 examples of this phenomenon (Fig. 6). In all of our cases, the ~22-nt RNA from one arm of the fold-back was cloned much more frequently than that from the other and was far more readily detected on Northern blots. We designated the less frequently cloned RNA as the miRNA-star (miRNA*) fragment (Lau et al. 2001).

Discussion

We have developed a computational procedure for identifying miRNA genes conserved in two genomes. By using this procedure, together with extensive sequencing of clones from libraries of small RNAs, we have now identified 87 miRNA genes in *C. elegans* (Tables 1, 2). Together with *mir-88* (Lee and Ambros 2001), which we have not yet cloned or found computationally, the number of validated *C. elegans* genes stands at 88. More than

a third of these genes have human homologs (Fig. 3), and a similar fraction, including previously unrecognized *lin-4* and *let-7* paralogs, is differentially expressed during larval development (Fig. 4). Most miRNAs accumulated to very high steady-state levels, with some at least as plentiful as the U6 snRNA (Fig. 5). Below, we discuss some implications of these results with regard to some of the defining features of miRNA genes in animals, the processing of miRNA precursors, and the number of miRNA genes remaining to be identified.

MiRscan accuracy and the defining features of miRNAs

As calculated in the Results section, the specificity of MiRscan was ≥ 0.70 at a sensitivity that detects half the previously known *C. elegans* miRNAs, when starting from an assembled *C. elegans* genome and *C. briggsae* shotgun reads. This accuracy was sufficient to identify new genes and obtain an upper bound on the total number of miRNA genes in the worm genome (described later). However, it was not sufficient to reliably identify all the conserved miRNA genes in *C. elegans*. The accuracy of MiRscan appears to be at least as high as that of general methods to identify ncRNA genes in bacteria (Argaman et al. 2001; Rivas et al. 2001; Wassarman et al. 2001), but is lower than that of algorithms designed to identify protein-coding genes or specialized programs that predict tRNAs and snoRNAs (Lowe and Eddy 1997, 1999; Burge and Karlin 1998). The relative difficulty in identifying miRNAs can be explained by the low information content inherent in their small size and lack of strong primary sequence motifs. The performance of

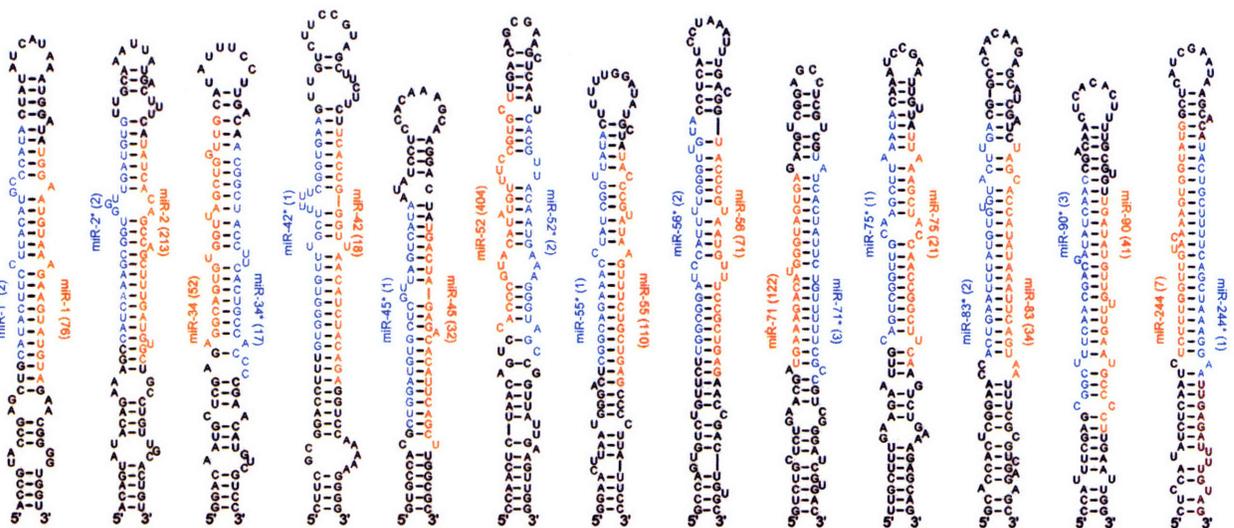


Figure 6. miRNA (red) and miRNA* (blue) sequences within the context of their predicted fold-back precursors. The number of sequenced clones is shown in parentheses. For each miRNA and miRNA*, colored residues are those for the most frequently cloned species. There was 3' heterogeneity among the sequenced clones for some miRNA*s and most miRNAs. Heterogeneity at the 5' terminus was not seen among the sequenced clones for the miRNA*s and was rare among those for the miRNAs; when it occurred, it was not observed for more than one of the many clones representing each miRNA.

MiRscan will improve with a more complete and assembled *C. briggsae* genome. We anticipate that using only those sequences conserved in a syntenic alignment of the two genomes would capture fewer of the background sequences, enabling the authentic miRNAs to be more readily distinguished from the false positives.

Improvement would also come from bringing in a third nematode genome, particularly a genome more divergent than those of *C. elegans* and *C. briggsae*. The advantage of such an additional genome is illustrated by our application of MiRscan to the identification of vertebrate miRNAs using three genomes. The version of MiRscan described here, which had been trained on the set of 50 miRNAs conserved in worms, was applied to the assembled human genome, shotgun reads of the mouse genome, and the assembled pufferfish (*Fugu*) genome (Lim et al. 2003). This analysis had a specificity of ≥ 0.71 at a sensitivity that detected three fourths of the previously known vertebrate miRNAs. The accuracy of the vertebrate analysis was therefore substantially improved over that of the *C. elegans/C. briggsae* analysis, even though the vertebrate genomes are 4–30 times larger than those of *C. elegans* and *C. briggsae*, and are expected to have a correspondingly higher number of background sequences. This improved performance can be attributed to using three genomes, as well as to the evolutionary distance between the mammalian and fish genomes, which are distant enough to reduce the number of fortuitously high scoring sequences, yet close enough to retain most of the known miRNAs.

Other improvements in the computational identification of miRNAs will come with the definition of additional sequence and structural features that specify which sequences are transcribed, processed into miRNAs, and loaded into the miRNP. With the exception of sequence conservation, the features that MiRscan currently uses to identify miRNAs (Fig. 1A) are among those that the cell also uses to specify the biogenesis of miRNAs and miRNPs. The utility of these parameters for MiRscan (Fig. 1B) is a function of both the degree to which these features are correctly modeled (or have already been used to restrict the number of miRNA candidates; see Fig. 1B legend) and their relative importance in vivo. Clearly, much of what defines a miRNA in vivo remains to be determined. Sequence elements currently unavailable for MiRscan include transcriptional promoter and termination signals. Additional sequence and structural features important for processing of the primary transcript and the hairpin precursors also remain to be identified (Lee et al. 2002).

miRNA biogenesis

The presence of miRNA* species, observed now for 14 of the *C. elegans* miRNAs (Fig. 6; Lau et al. 2001), provides evidence for the idea that Dicer processing of miRNA precursors resembles that of siRNA precursors (Hutvagner and Zamore 2002; Reinhart et al. 2002). We suspect that with more extensive sequencing of clones,

miRNA* sequences will be found for a majority of the miRNA precursors, a notion supported by the identification of additional miRNA* sequences using our PCR assay (data not shown). As observed for both *MIR156d* and *MIR169* in plants (Reinhart et al. 2002), the miRNA:miRNA* segments are typically presented within the predicted precursor, paired to each other with 2-nt 3' overhangs (Fig. 6)—a structure analogous to that of a classical siRNA duplex. This is precisely the structure that would be expected if both the miRNA and the miRNA* were excised from the same precursor molecule, and the miRNA* fragments were transient side-products of productive Dicer processing. An alternative model for miRNA biogenesis and miRNA* formation, which we do not favor but cannot rule out, is that the Dicer complex normally excises a ~22-nt RNA from only one side of a miRNA precursor but it sometimes binds the precursors in the wrong orientation and excises the wrong side. In an extreme version of the favored model, the production of the miRNA* would be required for miRNA processing and miRNP assembly; in a less extreme version, miRNA* production would be an optional off-pathway phenomenon. The idea that ~22-nt RNAs might be generally excised from both sides of the same precursor stem loop brings up the question of why the miRNAs and miRNA*s are present at such differing levels. With the exception of miR-34* (sequenced 17 times), none of the miRNA*s is represented by more than three sequenced clones. Perhaps the miRNAs are stabilized relative to their miRNA* fragments because they preferentially enter the miRNP/RISC complex. Alternatively, both the miRNA and the miRNA* might enter the complex, but the miRNA might be stabilized by interactions with its targets.

Five of the newly identified miRNAs are within annotated introns, all five in the same orientation as the predicted mRNAs. When considered together with the previously identified miRNAs found within annotated introns (Lau et al. 2001), 10 of 12 known *C. elegans* miRNAs predicted to be in introns are in the same orientation as the predicted mRNAs. This bias in orientation, also reported recently for mammalian miRNAs (Lagos-Quintana et al. 2003), suggests that some of these miRNAs are not transcribed from their own promoters but instead derive from the excised pre-mRNA introns (as are many snoRNAs), and it is easy to imagine regulatory scenarios in which the coordinate expression of a miRNA with an mRNA would be desirable.

The number of miRNA genes in *C. elegans* and other animals

In addition to providing a set of candidate miRNAs, MiRscan scoring provides a means to estimate the total number of miRNA genes in *C. elegans*. A total of 64 loci have scores greater than the median score of the 58 initially reported *C. elegans* miRNAs (Fig. 2B). Note that this set of 58 miRNAs includes not only the 50 conserved miRNAs of the training set but also the eight previously reported miRNAs that were not in our set of

Lim et al.

36,000 potential stem loops, usually because they lacked easily recognizable *C. briggsae* orthologs. Thus, the estimate calculated below takes into account the poorly conserved miRNAs without MiRscan scores. Four of the 64 high-scoring loci are known to be false positives. Thus, the upper bound on the number of miRNA genes in *C. elegans* would be $2 \times (64 - 4)$, or 120. This upper bound of ~120 genes remained stable when extrapolating from points other than the median, ranging from the top 25th–55th percentiles. For this estimate, we made the assumption that the set of all *C. elegans* miRNAs has a distribution of MiRscan scores similar to the distribution of initially reported miRNAs. Such an assumption might be called into question, particularly when considering that the initially reported miRNAs served as a training set for the development of MiRscan (even though the scores of the training-set loci have been jackknifed to prevent overfitting). However, this assumption is supported by two observations. First, the set of newly cloned miRNAs did indeed have a distribution of scores indistinguishable from that of the training set of previously reported miRNAs (Fig. 2B). Second, there is no correlation between the number of times that a miRNA has been cloned and its MiRscan score (Fig. 7). The absence of a correlation between cloning frequency and MiRscan score lessens our concern that miRNAs that are difficult to clone, including those still not present in our set of 3423 sequenced clones, might represent a population of miRNAs that are refractory to computational analysis as well.

This estimate of 120 genes is an upper bound and would decrease if additional high-scoring candidates were shown to be false positives. The extreme scenario, in which all are false positives, places the lower bound of miRNA genes near the number of validated genes, adding perhaps another five genes to account for the low-

scoring counterparts of the five computational candidates validated only by Northern and PCR, yielding a lower bound on the number of *C. elegans* miRNAs of ~93.

Our count of 105 ± 15 miRNA genes in *C. elegans* might underestimate the true count if there are miRNAs with unusual fold-back precursors that were cloned but dismissed as endogenous siRNAs or degradation fragments. To investigate this possibility, we examined the expression of each small RNA that was cloned more than once but did not appear to derive from a canonical miRNA precursor as predicted by RNAfold. Because most (72 of 88) of the authentic miRNAs identified to date were represented by multiple clones (Table 1), this analysis should uncover most of the miRNAs coming from nonconventional precursors. This broader analysis detected only a single additional miRNA, miR-229. All of the other sequences that we cloned more than once were minor degradation fragments or processing byproducts of larger ncRNAs (e.g., the 5' leader sequence of a tRNA). Thus, the number of miRNAs that derive from nonconventional precursors is not sufficient to significantly influence the miRNA gene count.

The estimated number of miRNA genes represents between 0.5% and 1% of the genes identified in the *C. elegans* genome, a fraction similar to that seen for other very large gene families with presumed regulatory roles, such as those encoding nuclear hormone receptors (270 predicted genes), C2H2 Zinc-finger proteins (157 predicted genes), and homeodomain proteins (93 predicted genes; Chervitz et al. 1998; *C. elegans* Sequencing Consortium 1998). Extending our analysis to vertebrate genomes revealed that 230 ± 30 of the human genes are miRNAs, also nearly 1% of the genes in the genome (Lim et al. 2003). The miRNA genes are also among the most abundant of the ncRNA gene families in humans, comparable in number to the genes encoding rRNAs (~650–900 genes), tRNAs (~500 genes), snRNAs (~100 genes), and snoRNAs (~100–200 genes; Lander et al. 2001). For rRNAs, tRNAs, and snRNAs, the hundreds of gene copies in the human genome represent only relatively few distinct genes, probably <100 distinct genes for all three classes combined. For the miRNAs and snoRNAs, there are many more distinct genes, and each is present in only one or a few copies.

Unlike the other large ncRNA gene families and many of the transcription-factor gene families, there is no indication that miRNAs are present in single-celled organisms such as yeast. A pilot attempt to clone miRNAs from *Schizosaccharomyces pombe* did not detect any miRNAs (Reinhart and Bartel 2002), and there is no evidence that the proteins (such as Dicer) needed for miRNA accumulation in plants and animals are present in *Saccharomyces cerevisiae*. Given the known roles of miRNAs in *C. elegans* development (Lee et al. 1993; Wightman et al. 1993; Reinhart et al. 2000) and the very probable roles of miRNAs in plant development (Rhoades et al. 2002), it is tempting to speculate that the substantial expansion of miRNA genes in animals (and the apparent loss of miRNA genes in yeast) is related to

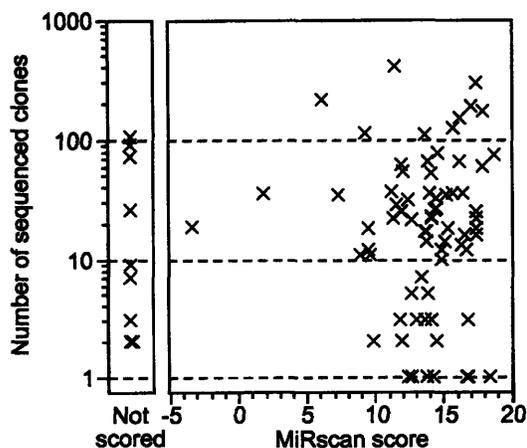


Figure 7. Plot illustrating the absence of a correlation between the MiRscan score of a cloned miRNA and the number of times that miRNA was cloned and sequenced. Nine of 80 cloned loci of Table 2 were not scored (left) because potential homologs of these genes were not identified among the available *C. briggsae* sequencing reads.

their importance in specifying cell differentiation and developmental patterning, and that the extra layer of gene regulation afforded by miRNAs was crucial for the emergence of multicellular body plans. The identification of most of the worm miRNAs and the quantitation of the number of genes remaining to be found are important steps toward understanding the evolution of this intriguing class of genes and placing them within the gene regulatory circuitry of these and other animals.

Materials and methods

Computational identification of stem loops

Potential miRNA stem loops were located by sliding a 110-nt window along both strands of the *C. elegans* genome (Worm-Base release 45, <http://www.wormbase.org>) and folding the window with the secondary structure-prediction program RNAfold (Hofacker et al. 1994) to identify predicted stem-loop structures with a minimum of 25 bp and a folding free energy of at least 25 kcal/mole ($\Delta C^\circ_{\text{folding}} \leq -25$ kcal/mole). Sequences that matched repetitive elements were discarded, as were those with skewed base compositions not observed in known miRNA stem loops and those that overlapped with annotated coding regions. Stem loops that had fewer base pairs than overlapping stem loops were also culled. *C. briggsae* sequences with at least loose sequence similarity to the remaining *C. elegans* sequences were identified among *C. briggsae* shotgun sequencing reads (November 2001 download from <http://www.ncbi.nlm.nih.gov/Traces>) using WU-BLAST with default parameters and a non-stringent cutoff of $E < 1.8$ (W. Gish, <http://blast.wustl.edu>). These *C. briggsae* sequences were folded with RNAfold to ensure that they met the minimal requirements for a hairpin structure as described above. This procedure yielded ~40,000 pairs of potential miRNA hairpins. For each pair of potential miRNA hairpins, a consensus *C. elegans/C. briggsae* structure was generated using the alidot and pfrali utilities from the Vienna RNA package (Hofacker et al. 1998; Hofacker and Stadler 1999; <http://www.tbi.univie.ac.at/~ivo/RNA>). To create RNA consensus structures, alidot and pfrali combine a Clustal alignment (Thompson et al. 1994) of a pair of sequences with either the minimum free energy structures of these sequences (alidot) derived using the Zuker algorithm (Zuker 1994) or the base pairing probability matrices of these sequences (pfrali) derived using the McCaskill algorithm (McCaskill 1990).

MiRscan

Of the ~40,000 pairs of hairpins, 35,697 had the minimal conservation and base pairing needed to receive a MiRscan score. Among this set were 50 of the 53 previously published miRNAs that were reported to be conserved between *C. elegans* and *C. briggsae* (Lau et al. 2001; Lee and Ambros 2001). [miR-53 is included as a previously reported conserved miRNA because it is nearly identical to miR-52, which has a highly conserved *C. briggsae* ortholog (Lau et al. 2001; Lee and Ambros 2001). The three conserved genes missing from the ~36,000 pairs of hairpins were *mir-56*, *mir-75*, and *mir-88*. The reverse complements of *mir-75* and *mir-88* were later observed among the ~36,000 hairpins and given scores (Table 1).] The MiRscan program was developed to discriminate these 50 known miRNA hairpins from background sequences in the set of ~36,000 hairpins. For a given 21-nt miRNA candidate, MiRscan makes use of the seven features derived from the consensus hairpin structure illus-

trated in Figure 1A: x_1 , "miRNA base pairing," the sum of the base-pairing probabilities for pairs involving the 21-nt candidate miRNA; x_2 , "extension of base pairing," the sum of the base-pairing probabilities of the pairs predicted to lie outside the 21-nt candidate miRNA but within the same helix; x_3 , "5' conservation," the number of bases conserved between *C. elegans* and *C. briggsae* within the first 10 bases of the miRNA candidate; x_4 , "3' conservation," the number of conserved bases within the last 11 bases of the miRNA candidate; x_5 , "bulge symmetry," the number of bulged or mismatched bases in the candidate miRNA minus the number of bulged or mismatched bases in the corresponding segment on the other arm of the stem loop; x_6 , "distance from loop," the number of base pairs between the loop of the stem loop and the closest end of the candidate; and x_7 , "initial pentamer," the specific bases at the first five positions at the candidate 5' terminus.

For a given feature i with a value x_i , MiRscan assigns a log-odds score

$$s_i(x_i) = \log_2 \left(\frac{f_i(x_i)}{g_i(x_i)} \right),$$

where $f_i(x_i)$ is an estimate of the frequency of feature value x_i in miRNAs derived from the training set of 50 known miRNAs, and $g_i(x_i)$ is an estimate of the frequency of feature value x_i among the background set of ~36,000 hairpin pairs. The overall score assigned to a candidate miRNA is simply the sum of the log-odds scores for the seven features:

$$S = \sum_{i=1..7} s_i(x_i).$$

To score a given hairpin, MiRscan slides a 21-nt window representing the candidate miRNA along each arm of the hairpin, assigns a score to each window, and then assigns the hairpin the score of its highest-scoring window. In order to be evaluated, a window was required to be two to nine consensus base pairs away from the terminal loop.

For features x_1 , x_3 , x_4 , x_5 , and x_6 , f_i and g_i were obtained by smoothing the empirical frequency distributions from the training and background sets, respectively, using the R statistical package (<http://lib.stat.cmu.edu/R/CRAN>) with a triangular kernel. Because x_1 and x_2 are not independent of each other, the relative contribution of x_2 was decreased by computing f_2 and g_2 separately subject to the conditions $x_1 \geq 9$ and $x_1 < 9$, in order to account for this dependence. For x_7 , a weight matrix model (WMM) was generated for the five positions at the miRNA 5' terminus. The background WMM, g_7 , was set equal to the base composition of the background sequence set. The miRNA WMM, f_7 , was derived from the position-specific base frequencies of the 50 training set sequences, using standard unit pseudo-counts and normalizing for the contributions of related miRNAs.

Because both strands of the *C. elegans* genome were analyzed, both a hairpin sequence and its reverse complement were sometimes included in the set of ~36,000 stem loops. For representation in Figure 2, in such cases both sequences were considered as a single locus that received the score of the higher scoring hairpin. Also, to prevent overscoring of the 50 known miRNA loci within the training set, each known miRNA locus was assigned a jackknife score calculated by using a training set consisting of the other 49 miRNAs. MiRscan is available for use (<http://genes.mit.edu/mirscan>).

RNA cloning and bioinformatic analyses

Small RNAs were cloned as described previously (Lau et al. 2001), using the protocol available on the Web (<http://web>).

Lim et al.

wi.mit.edu/bartel/pub). Sequencing was performed by Agen-court Bioscience. Sequences of known *C. elegans* tRNA and rRNA were removed, and the remaining clones were clustered based on the location of their match to the *C. elegans* genome (*C. elegans* Sequencing Consortium 1998), downloaded from WormBase (<http://www.wormbase.org>). Genomic loci not previously reported to encode miRNAs were examined by using the RNA-folding program RNAfold (Hofacker et al. 1994). Two sequences were folded for each locus: one included 15 nt upstream and 60 nt downstream of the most frequently cloned sequence from that locus; the other included 60 nt upstream and 15 nt downstream. Sequences for which the most stable predicted folding resembled the stem-loop precursors of previously validated miRNAs were carried forward as candidate miRNA loci. Sequences without classical stem-loop precursors were also analyzed further (see Discussion), but only one, miR-229, was classified as a miRNA. The clones classified as representing potential fragments of mRNAs (18 clones) and potential antisense fragments of mRNAs (23 clones) corresponded to predicted ORFs (as annotated in GenBank) or probable UTR segments (100 bp upstream or 200 bp downstream of the predicted ORF).

Northern

Expression of candidate miRNA loci was examined by using Northern blots and radiolabeled DNA probes (Lau et al. 2001). To maintain hybridization specificity without varying hybridization or washing conditions, the length of probes for different sequences was adjusted so that the predicted melting temperatures of the miRNA-probe duplexes did not exceed 60°C (Sugimoto et al. 1995). Probes not corresponding to the entire miRNA sequence were designed to hybridize to the 3' region of the miRNA, which is most divergent among related miRNA sequences.

PCR validation

A PCR assay was performed to detect the sequences of predicted miRNAs within a cDNA library constructed from 18- to 26-nt RNAs expressed in mixed-stage worms. This library, the same as that used for cloning (Lau et al. 2001), consisted of PCR-amplified DNA that comprised the 18- to 26-nt sequences flanked by 3'- and 5'-adaptor sequences. For each miRNA candidate, a primer specific to the predicted 3' terminus of the candidate and a primer corresponding to the 5'-adaptor sequence common to all members of the library (ATCGTAG GCACCTGAAA) were used at concentrations of 1.0 μ M and 0.1 μ M, respectively (100 μ L PCR reaction containing 5 μ L of a 400-fold dilution of the PCR reaction previously used to amplify all members of the cDNA library). The specific primer was added after the initial denaturation incubation had reached 80°C. After 20 PCR cycles, the reaction was diluted 20-fold into a fresh PCR reaction for another 20 cycles. PCR products were cloned and sequenced to both identify the 5' terminus of the miRNA and ensure that the amplified product was not a primer-dimer or other amplification artifact. Specific primers for the reactions that successfully detected candidate miRNAs were ACCATGCCAACAGTTG (miR-250), TAAGAGCGGCACCA CTAC (miR-251), TACCTGCGGCACTACTAC (miR-252), GTCAGTGTAGTGAGG (miR-253), TACAGTCGGAAAGA TTTG (miR-254), and GTGGAAATCTATGCTTC (miR-254*).

Quantitative Northern

miRNA standards (purchased from Dharmacon) were diluted to appropriate concentrations in the presence of 1.0 μ g/ μ L carrier

RNA in the form of either *E. coli* tRNA or HeLa cell total RNA. Northern analysis was performed (Lau et al. 2001), loading 30 μ g of RNA per lane, in the format shown for miR-66 (Fig. 5A). Signals were quantitated using phosphor imaging, standard curves (linear through at least three orders of magnitude, including the region of interpolation) were constructed, and absolute amounts of miRNAs per sample were determined, as illustrated for miR-66 (Fig. 5B). The average number of miRNA molecules per *glp-4* adult nematode was calculated using 19 ng as the average amount of total RNA extracted per worm. This number was determined as the average of three independent extraction trials, from known numbers of synchronized, 2-day-old adult *glp-4(bn2)* hermaphrodites, the same frozen worm population used for the quantitative Northern blots. All extractions were performed as described previously (Lau et al. 2001), except during two of the trials a radiolabeled miRNA was spiked into the preparation during worm lysis. At least 90% of this RNA was recovered, indicating near quantitative yield. Having calculated the number of each miRNA per worm, the average number of miRNAs per cell was calculated using 989 as number of cells per worm. The 989 cells per worm is based on the 959 somatic nuclei of the adult hermaphrodites plus the 30 germ nuclei of 2-day-old adult *glp-4(bn2)* animals (Sulston et al. 1983; Beanan and Strome 1992). Total RNA from known numbers of HeLa cells was determined in an analogous fashion.

Acknowledgments

We thank the *C. briggsae* Sequencing Consortium for the availability of sequencing reads, WormBase (<http://www.wormbase.org>) for annotation of the *C. elegans* genome, Compaq for computer resources, V. Ambros for communicating unpublished data, C. Mello for the *dcr-1* strain, S. Griffiths-Jones and the miRNA Gene Registry for assistance with gene names, P. Zamore for helpful comments on this manuscript, and R.F. Yeh, H. Houbaviy, and G. Ruvkun for advice and helpful discussions. Supported by grants from the NIH and the David H. Koch Cancer Research Fund (D.P.B.) and a grant from the NIH (C.B.B.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S., Griffiths-Jones, S., Matzke, M., et al. 2003. A uniform system for microRNA annotation. *RNA* 9: 277-279.
- Aravin, A.A., Naumova, N.M., Tulin, A.A., Rozovsky, Y.M., and Gvozdev, V.A. 2001. Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in *D. melanogaster* germline. *Curr. Biol.* 11: 1017-1027.
- Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., and Altuvia, S. 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* 11: 941-950.
- Beanan, M.J. and Strome, S. 1992. Characterization of a germline proliferation mutation in *C. elegans*. *Development* 116: 755-766.
- Bernstein, E., Caudy, A.A., Hammond, S.M., and Hannon, G.J. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* 409: 295-296.

- Broverman, S.A. and Meneely, P.M. 1994. Meiotic mutants that cause a polar decrease in recombination on the X chromosome in *Caenorhabditis elegans*. *Genetics* **136**: 119–127.
- Burge, C.B. and Karlin, S. 1998. Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* **8**: 346–354.
- Caudy, A.A., Myers, M., Hannon, G.J., and Hammond, S.M. 2002. Fragile X-related protein and VIG associate with the RNA interference machinery. *Genes & Dev.* **16**: 2491–2496.
- C. elegans Sequencing Consortium. 1998. Genome sequence of the nematode *C. elegans*: A platform for investigating biology. *Science* **282**: 2012–2018.
- Chervitz, S.A., Aravind, L., Sherlock, G., Ball, C.A., Koonin, E.V., Dwight, S.S., Harris, M.A., Dolinski, K., Mohr, S., Smith, T., et al. 1998. Comparison of the complete protein sets of worm and yeast: Orthology and divergence. *Science* **282**: 2022–2028.
- Djikeng, A., Shi, H., Tschudi, C., and Ullu, E. 2001. RNA interference in *Trypanosoma brucei*: Cloning of small interfering RNAs provides evidence for retroposon-derived 24–26-nucleotide RNAs. *RNA* **7**: 1522–1530.
- Elbashir, S.M., Lendeckel, W., and Tuschl, T. 2001a. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes & Dev.* **15**: 188–200.
- Elbashir, S.M., Martinez, J., Patkaniowska, A., Lendeckel, W., and Tuschl, T. 2001b. Functional anatomy of siRNAs for mediating efficient RNAi in *Drosophila melanogaster* embryo lysate. *EMBO J.* **20**: 6877–6888.
- Grishok, A., Pasquinelli, A.E., Conte, D., Li, N., Parrish, S., Ha, I., Baillie, D.L., Fire, A., Ruvkun, G., and Mello, C.C. 2001. Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control *C. elegans* developmental timing. *Cell* **106**: 23–34.
- Ha, I., Wightman, B., and Ruvkun, G. 1996. A bulged lin-4/lin-14 RNA duplex is sufficient for *Caenorhabditis elegans* lin-14 temporal gradient formation. *Genes & Dev.* **10**: 3041–3050.
- Hall, I.M., Shankaranarayana, G.D., Noma, K., Ayoub, N., Cohen, A., and Grewal, S.I. 2002. Establishment and maintenance of a heterochromatin domain. *Science* **297**: 2232–2237.
- Hamilton, A.J. and Baulcombe, D.C. 1999. A novel species of small antisense RNA in posttranscriptional gene silencing. *Science* **286**: 950–952.
- Hamilton, A., Voinnet, O., Chappell, L., and Baulcombe, D. 2002. Two classes of short interfering RNA in RNA silencing. *EMBO J.* **21**: 4671–4679.
- Hammond, S.C., Bernstein, E., Beach, D., and Hannon, G.J. 2000. An RNA-directed nuclease mediates posttranscriptional gene silencing in *Drosophila* cells. *Nature* **404**: 293–296.
- Hastie, N.D. and Bishop, J.O. 1976. The expression of three abundance classes of messenger RNA in mouse tissues. *Cell* **9**: 761–774.
- Hofacker, I.L. and Stadler, P.F. 1999. Automatic detection of conserved base pairing patterns in RNA virus genomes. *Comput. Chem.* **15**: 401–414.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, S., Tacker, M., and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte f. Chemie* **125**: 167–188.
- Hofacker, I.L., Fekete, M., Flamm, C., Huynen, M.A., Rauscher, S., Stolorz, P.E., and Stadler, P.F. 1998. Automatic detection of conserved RNA structure elements in complete RNA virus genomes. *Nucleic Acids Res.* **26**: 3825–3836.
- Hutvagner, G. and Zamore, P.D. 2002. A microRNA in a multiple-turnover RNAi enzyme complex. *Science* **297**: 2056–2060.
- 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.
- Ishizuka, A., Siomi, M.C., and Siomi, H. 2002. A *Drosophila* fragile X protein interacts with components of RNAi and ribosomal proteins. *Genes & Dev.* **16**: 2497–2508.
- Ketting, R.F., Fischer, S.E.J., Bernstein, E., Sijen, T., Hannon, G.J., and Plasterk, R.H.A. 2001. Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in *C. elegans*. *Genes & Dev.* **15**: 2654–2659.
- Klahre, U., Crete, P., Leuenberger, S.A., Iglesias, V.A., and Meins, F. 2002. High molecular weight RNAs and small interfering RNAs induce systemic posttranscriptional gene silencing in plants. *Proc. Natl. Acad. Sci.* **99**: 11981–11986.
- 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.
- Lagos-Quintana, M., Rauhut, R., Meyer, J., Borkhardt, A., and Tuschl, T. 2003. New microRNAs from mouse and human. *RNA* **9**: 175–179.
- Lai, E.C. 2002. MicroRNAs are complementary to 3'UTR motifs that mediate negative post-transcriptional regulation. *Nat. Genet.* **30**: 363–364.
- Lander E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., Fitz Hugh, W., et al. 2001. Initial sequencing and analysis of the human genome. *Nature* **409**: 860–921.
- 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., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO J.* **21**: 4663–4670.
- Lim, L.P., Glasner, M.E., Yekta, S., Burge, C.B., and Bartel, D.P. 2003. Vertebrate microRNA genes. *Science* **299**: 1540.
- Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002a. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**: 1605–1619.
- Llave, C., Xie, Z., Kasschau, K.D., and Carrington, J.C. 2002b. Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297**: 2053–2056.
- Lowe, T.M. and Eddy, S.R. 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* **25**: 955–964.
- . 1999. A computational screen for methylation guide snoRNAs in yeast. *Science* **283**: 1168–1171.
- Martinez, J., Patkaniowska, A., Urlaub, H., Luhrmann, R., and Tuschl, T. 2002. Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. *Cell* **110**: 563–574.
- McCaskill, J.S. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* **29**: 1105–1119.
- Mochizuki, K., Fine, N.A., Fujisawa, T., and Gorovsky, M.A. 2002. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell* **110**: 689–699.

Lim et al.

- 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.
- Mourelatos, Z., Dostie, J., Paushkin, S., Sharma, A., Charroux, B., Abel, L., Rappsilber, J., Mann, M., and Dreyfuss, G. 2002. miRNPs: A novel class of ribonucleoproteins containing numerous microRNAs. *Genes & Dev.* **16**: 720–728.
- Nykänen, A., Haley, B., and Zamore, P.D. 2001. ATP requirements and small interfering RNA structure in the RNA interference pathway. *Cell* **107**: 309–321.
- Okazaki, Y., Furuno, M., Kasukawa, T., Adachi, J., Bono, H., Kondo, S., Nikaido, I., Osato, N., Saito, R., Suzuki, H., et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* **420**: 563–573.
- Olsen, P.H. and Ambros, V. 1999. The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. *Dev. Biol.* **216**: 671–680.
- Park, W., Li, J., Song, R., Messing, J., and Chen, X. 2002. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12**: 1484–1495.
- Parrish, S., Fleenor, J., Xu, S., Mello, C., and Fire, A. 2000. Functional anatomy of a dsRNA trigger: Differential requirement for the two trigger strands in RNA interference. *Mol. Cell* **6**: 1077–1087.
- Pasquinelli, A.E., Reinhart, B.J., Slack, F., Martindale, M.Q., Kuroda, M., Maller, B., Srinivasan, A., Fishman, M., Hayward, D., Ball, E., et al. 2000. Conservation across animal phylogeny of the sequence and temporal regulation of the 21 nucleotide *let-7* heterochronic regulatory RNA. *Nature* **408**: 86–89.
- Pickford, A.S., Catalanotto, C., Cogoni, C., and Macino, G. 2002. Quelling in *Neurospora crassa*. *Adv. Genet.* **46**: 277–303.
- Reinhart, B.J. and Bartel, D.P. 2002. Small RNAs correspond to centromere heterochromatic repeats. *Science* **297**: 1831.
- Reinhart, B.J., Slack, F.J., Basson, M., Bettinger, J.C., Pasquinelli, A.E., 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.
- Rivas, E., Klein, R.J., Jones, T.A., and Eddy, S.R. 2001. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* **11**: 1369–1373.
- Schwarz, D.S., Hutvagner, G., Haley, B., and Zamore, P.D. 2002. Evidence that siRNAs function as guides, not primers, in the *Drosophila* and human RNAi pathways. *Mol. Cell* **10**: 537–548.
- Slack, F.J., Basson, M., Liu, Z., Ambros, V., Horvitz, H.R., and Ruvkun, G. 2000. The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. *Mol. Cell* **5**: 659–669.
- Sugimoto, N., Nakano, S., Katoh, M., Matsumura, A., Nakamura, H., Ohmichi, T., Yoneyama, M., and Sasaki, M. 1995. Thermodynamic parameters to predict stability of RNA/DNA hybrid duplexes. *Biochemistry* **34**: 11211–11216.
- Sulston, J.E., Schierenberg, E., White, J.G., and Thomson, J.N. 1983. The embryonic cell lineage of the nematode *Caenorhabditis elegans*. *Dev. Biol.* **100**: 64–119.
- 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.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4680.
- Vance, V. and Vaucheret, H. 2001. RNA silencing in plants: Defense and counterdefense. *Science* **292**: 2277–2280.
- Volpe, T., Kidner, C., Hall, I., Teng, G., Grewal, S., and Martienssen, R. 2002. Heterochromatic silencing and histone H3 lysine 9 methylation are regulated by RNA interference. *Science* **297**: 1833–1837.
- Wassarman, K.M., Repoila, F., Rosenow, C., Storz, G., and Gottesman, S. 2001. Identification of novel small RNAs using comparative genomics and microarrays. *Genes & Dev.* **15**: 1637–1651.
- 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.
- 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.
- Zilberman, D., Cao, X., and Jacobsen, S.E. 2003. ARGONAUTE4 control of locus-specific siRNA accumulation and DNA and histone methylation. *Science* **299**: 716–719.
- Zuker, M. 1994. Prediction of RNA secondary structure by energy minimization. *Methods Mol. Biol.* **25**: 267–294.

Appendix II

Patterns of flanking sequence conservation and a characteristic upstream motif for microRNA gene identification

UWE OHLER,¹ SORAYA YEKTA,^{1,2} LEE P. LIM,¹⁻³ DAVID P. BARTEL,^{1,2} and CHRISTOPHER B. BURGE¹

¹Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA

²Whitehead Institute for Biomedical Research, Cambridge, Massachusetts 02142, USA

ABSTRACT

MicroRNAs are ~22-nucleotide (nt) RNAs processed from foldback segments of endogenous transcripts. Some are known to play important gene regulatory roles during animal and plant development by pairing to the messages of protein-coding genes to direct the post-transcriptional repression of these messages. Previously, we developed a computational method called MiRscan, which scores features related to the foldbacks, and used this algorithm to identify new miRNA genes in the nematode *Caenorhabditis elegans*. In the present study, to identify sequences that might be involved in processing or transcriptional regulation of miRNAs, we aligned sequences upstream and downstream of orthologous nematode miRNA foldbacks. These alignments showed a pronounced peak in sequence conservation about 200 bp upstream of the miRNA foldback and revealed a highly significant sequence motif, with consensus CTCCGCC, that is present upstream of almost all independently transcribed nematode miRNA genes. Scoring the pattern of upstream/downstream conservation, the occurrence of this sequence motif, and orthology of host genes for intronic miRNA candidates, yielded substantial improvements in the accuracy of MiRscan. Nine new *C. elegans* miRNA gene candidates were validated using a PCR-sequencing protocol. As previously seen for bacterial RNA genes, sequence features outside of the RNA secondary structure can therefore be very useful for the computational identification of eukaryotic noncoding RNA genes. The total number of confidently identified nematode miRNAs now approaches 100. The improved analysis supports our previous assertion that miRNA gene identification is nearing completion in *C. elegans* with apparently no more than 20 miRNA genes now remaining to be identified.

Keywords: microRNA; noncoding RNA; computational gene identification; regulatory motif; transcription

INTRODUCTION

MicroRNAs (miRNAs) are a class of small noncoding RNAs that are found in a variety of eukaryotic multicellular organisms (Lee et al. 1993; Reinhart et al. 2000; Lagos-Quintana et al. 2001; Lau et al. 2001; Lee and Ambros 2001; Llave et al. 2002; Park et al. 2002; Reinhart et al. 2002). They are known to be important gene-regulatory molecules in both animals and plants (Ambros 2003; Bartel 2004). In animals, miRNAs are processed in two steps (Lee et al.

2002, 2003), from primary transcripts to ~70-nucleotide (nt) precursors by the RNase III enzyme Droscha, and from precursors to the ~22-nt single-stranded miRNAs by the RNase III enzyme Dicer. The processed miRNAs can direct post-transcriptional regulation of specific target mRNAs (Lee et al. 1993; Wightman et al. 1993; Moss et al. 1997; Reinhart et al. 2000; Lai 2002; Abrahante et al. 2003; Brennecke et al. 2003; Lewis et al. 2003; Lin et al. 2003; Yekta et al. 2004).

Noncoding RNA genes (Eddy 2001) are typically independently transcribed by one of the three RNA polymerases, for example, rRNA genes by RNA polymerase I (pol-I), most snRNA genes by RNA pol-II, and tRNA genes by RNA pol-III (Brown 2002). Alternatively, they can be cotranscribed within host genes, as is the case with most vertebrate snoRNA genes (Bachellerie et al. 2002), which are located within introns of pol-II-transcribed host genes. Most miRNA genes are located far away from any annotated genes, implying independent transcription from their own

Reprint requests to: David P. Bartel, Whitehead Institute, 9 Cambridge Center, Cambridge, MA 02142, USA; e-mail: dbartel@wi.mit.edu; fax: (617) 258-6768; and Christopher B. Burge, Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02142, USA; e-mail: cburge@mit.edu; fax: (617) 452-2936.

³**Present address:** Rosetta Inpharmatics, Merck & Co., 401 Terry Avenue N., Seattle, WA 98109, USA.

Article and publication are at <http://www.rnajournal.org/cgi/doi/10.1261/rna.5206304>.

promoters, but some lie within predicted introns of protein-coding genes (Lau et al. 2001; Lagos-Quintana et al. 2003)—for example, 22 of the 88 nematode miRNAs known at the start of this study have intronic locations. In most of these cases (80%), the introns are in the same orientation as the miRNAs, implying that the protein-coding genes might serve as host genes for coexpressed miRNAs. Therefore, in this study, we provisionally group the miRNA genes into two categories as follows: Those located in the sense strand of annotated introns are classified as cotranscribed miRNAs (although some might be independently transcribed), and all other miRNA genes, including those that are clustered in the genome in a configuration suggestive of transcription as a single polycistronic RNA (Lagos-Quintana et al. 2001; Lau et al. 2001) are classified as independently transcribed because they are unlikely to share a primary transcript with a non-miRNA host gene.

Although functional miRNA genes can be expressed by pol-II or pol-III (Zeng et al. 2002; Zeng and Cullen 2003; Chen et al. 2004), the identity of the polymerase(s) that transcribes the endogenous genes is not known. Some miRNA foldbacks are located in close genomic proximity to each other and are transcribed as polycistronic units (Lee et al. 2002; Aravin et al. 2003). The largest of these miRNA clusters extend well over a kilobase on the genome, which makes transcription of these clusters by pol-III unlikely, in that annotated nematode pol-III transcripts are only up to 300–400 bases in size (Harris et al. 2003). Likewise, the primary transcripts of some singly transcribed miRNAs often appear to be longer than typical pol-III transcripts (Lee et al. 2002). Transcriptional regulation of miRNAs is only beginning to be studied in detail (Johnson et al. 2003; Semper et al. 2003).

Computational identification of miRNAs is greatly aided by their occurrence in the context of conserved stem-loop foldbacks. Because of a more variable-sized foldback structure in plants (Reinhart et al. 2002), the prediction of plant miRNAs is more challenging, and has only recently been reported (Jones-Rhoades and Bartel 2004). Computational screens for conserved foldbacks, combined with large-scale cloning efforts, recently brought the number of identified *Caenorhabditis elegans* miRNA genes to 88 (Lim et al. 2003b). Since then, two groups have reported seven and 10 additional candidate miRNAs, respectively (Ambros et al. 2003b; Grad et al. 2003). These three independent studies give different upper-bound estimates, ranging from ~120 to 300 or more *C. elegans* miRNA genes. The number of *Drosophila* miRNA genes has been estimated at 110 (Lai et al. 2003), and about twice this number are thought to be present in vertebrates (Lim et al. 2003a). The computational approaches typically apply RNA folding methods to detect regions with potential to fold into stem-loop structures, use cross-species conservation to restrict the vast number of potential stem-loop structures found in each genome, and

score conserved foldbacks for conservation and a variety of sequence and secondary structural features.

Our goal here was to identify specific sequence features in the vicinity of independent and cotranscribed miRNAs, which might be involved in their expression, and to integrate these features into an improved version of the miRNA gene finding algorithm MiRscan (Lim et al. 2003b). In particular, we focused on (1) the pattern of conservation upstream and downstream of miRNA foldbacks; (2) specific sequence motifs adjacent to foldbacks likely to be involved in transcription or processing of miRNAs; and (3) the location of cotranscribed miRNAs in orthologous host genes. For independently transcribed miRNAs, we also examined the benefits of requiring synteny of the flanking protein-coding genes, as well as the use of whole-genome alignments. We concentrated our efforts on miRNAs in *C. elegans*, as this organism had been subject to the most comprehensive miRNA cloning effort at the time this study was begun, and the closely related nematode *Caenorhabditis briggsae* had the advantage of an assembled and preannotated genome, which has now been published (Stein et al. 2003). The presence of transcription initiation and termination sequence elements has been successfully used in computational identification of prokaryotic noncoding RNA genes (Argaman et al. 2001). Here, we demonstrate the use of features outside of the actual RNA secondary structure, such as an upstream promoter/processing motif and upstream/downstream sequence conservation, for computational discovery of noncoding RNA genes in eukaryotes.

RESULTS

Analysis of microRNA genes

Conservation upstream and downstream of miRNA genes

We assembled sets of 43 orthologous *C. elegans/C. briggsae* miRNA upstream and downstream sequences likely to contain transcriptional regulatory sequences. The Upstream Sequence Set (USS) encompasses the regions 2000 bp upstream, and the Downstream Sequence Set (DSS) encompasses the regions 1000 bp downstream of the foldbacks. For each pair of sequences from the USS and DSS data sets, the orthologous sequences were aligned with the tools DBA (Jareborg et al. 1999) and BayesBlockAligner (Zhu et al. 1998), and the resulting sets of aligned regions were merged. Downstream sequences were generally less conserved than upstream sequences, and in both directions the degree of conservation decreased with increasing distance from the foldback (Fig. 1). There was also a pronounced peak of conservation at about 200 bp upstream of the foldbacks. On average, 248 bp of the first 1000 bp upstream were aligned within conserved blocks of at least 70% se-

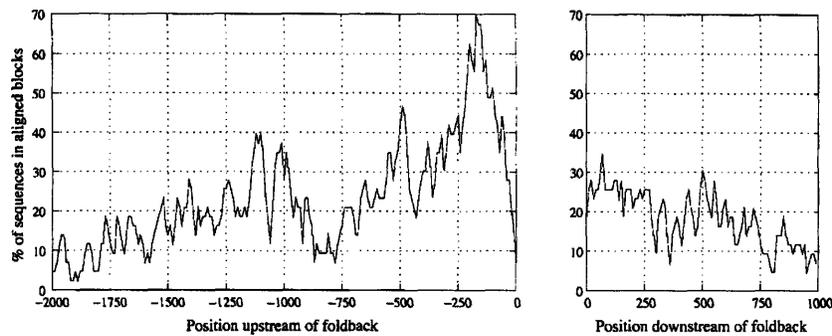


FIGURE 1. Conservation upstream and downstream of nematode microRNA foldbacks. The percentage of *C. elegans* sequences that are part of a conserved aligned block with *C. briggsae* at specific positions is plotted in bins of 10 bp. The positions are given relative to the beginning (left) or end (right) of the 110-nt segments containing the foldback. Genomic sequences were aligned using DBA and BayesBlockAligner as described in the text. Example alignments are part of the Supplementary Material (<http://genes.mit.edu/burgelab/MiRscanII>).

quence identity, compared with 146 bp of the first 1000 bp downstream.

Identification of a conserved upstream element

Next, we searched for conserved upstream sequence motifs, which might be involved in miRNA transcription or processing. Algorithms for the identification of conserved motifs can be grouped into enumerative and alignment approaches (Ohler and Niemann 2001). The ST algorithm, based on an approach described by Sinha and Tompa (2000), is an enumerative word-based algorithm that identifies statistically over-represented oligomers in a target set of sequences when compared with a background model. With this algorithm, we searched for over-represented words in the *C. elegans* sequence blocks conserved with *C. briggsae*, using a background model derived from the whole 2-kb upstream regions. The two significant distinct motifs that were found had the consensus sequences CTCCGCC (motif A) and GCGTGGCS (motif B; S = C or G). Motif A was highly significant, frequently occurring, and well conserved. By comparison, motif B had a much lower score and was less frequent (Fig. 2A).

We repeated this search with the alignment-based motif-finding tool MEME (Bailey and Elkan 1995), choosing the “zero-or-one-occurrence” alignment mode, which identifies motifs present in some, but not necessarily all of the sequences. MEME reported a motif essentially identical to motif A as the strongest hit, either when searching only in the conserved sequence blocks or in the complete USS (Fig. 2B). A highly similar motif was identified in the *C. briggsae* sequences. In both *C. elegans* and *C. briggsae* sequences, the motif was preferentially located <500 bp upstream of the foldback (Fig. 2C). The motif was found on both strands, with a ~2:1 preference for the forward strand. In most cases, the location of the best match in *C. elegans* (on either the

forward or reverse strand) was similar to that in *C. briggsae* (in 25/43 cases, the locations relative to the hairpin differed by <250 bp). Motif B (Fig. 2A) was not identified by MEME.

Finally, we asked whether motif A is also frequently found upstream of non-miRNA genes. The ST algorithm did not identify a similar motif in conserved sequence blocks upstream of 74 orthologous *C. elegans* and *C. briggsae* protein-coding genes (Webb et al. 2002). Also, no similar motif was found by MEME in sequences upstream of the 36 annotated *C. elegans* pol-II-transcribed snRNA genes (WormBase release 100), in the intronic sequences upstream of the 13 cotranscribed miRNA genes, or in the sequences upstream of

the 13 protein-coding genes with cotranscribed intronic miRNA genes. These observations indicated that occurrence of motif A is a useful marker of independently transcribed miRNA genes.

Upstream elements in mammals and insects

An investigation of the regions upstream of 59 orthologous human/mouse orthologous miRNAs likewise identified an over-represented motif, CCCWCCC (ST algorithm Z-score 11.1; control background score 5.7; W = A or T), which was present 98 times in conserved blocks of 45 upstream regions. A second motif, ATGCAT, was present 18 times in 14 regions. Analysis of a set of 31 upstream sequences of independently transcribed *Drosophila melanogaster* miRNAs (Aravin et al. 2003) with the ST algorithm again yielded ATGCAT as an over-represented motif, with an exact match in 13 sequences. We also scanned the 1000-bp upstream regions of these *Drosophila* miRNA genes for motifs enriched in core promoters of protein-coding genes (Ohler et al. 2002), but did not detect a consistent preference for any of the known motifs.

Analysis of downstream sequences and foldbacks

Next, we investigated whether candidate termination signals could be identified by the approach described above. A search for over-represented oligonucleotides in the conserved blocks of the DSS using the ST algorithm did not identify a single statistically significant motif. Because the alignment algorithms require colinearity of sequences, conserved motifs might be missed if their positions were poorly conserved. Applying MEME to the complete DSS identified the motifs TTTT[TG]GAAA in *C. elegans* (E-value 1.7e-5) and TTTYGAAA in *C. briggsae* (E-value 2.2e-6). Although instances of these motifs were found in all of the downstream sequences, there was no apparent positional conser-

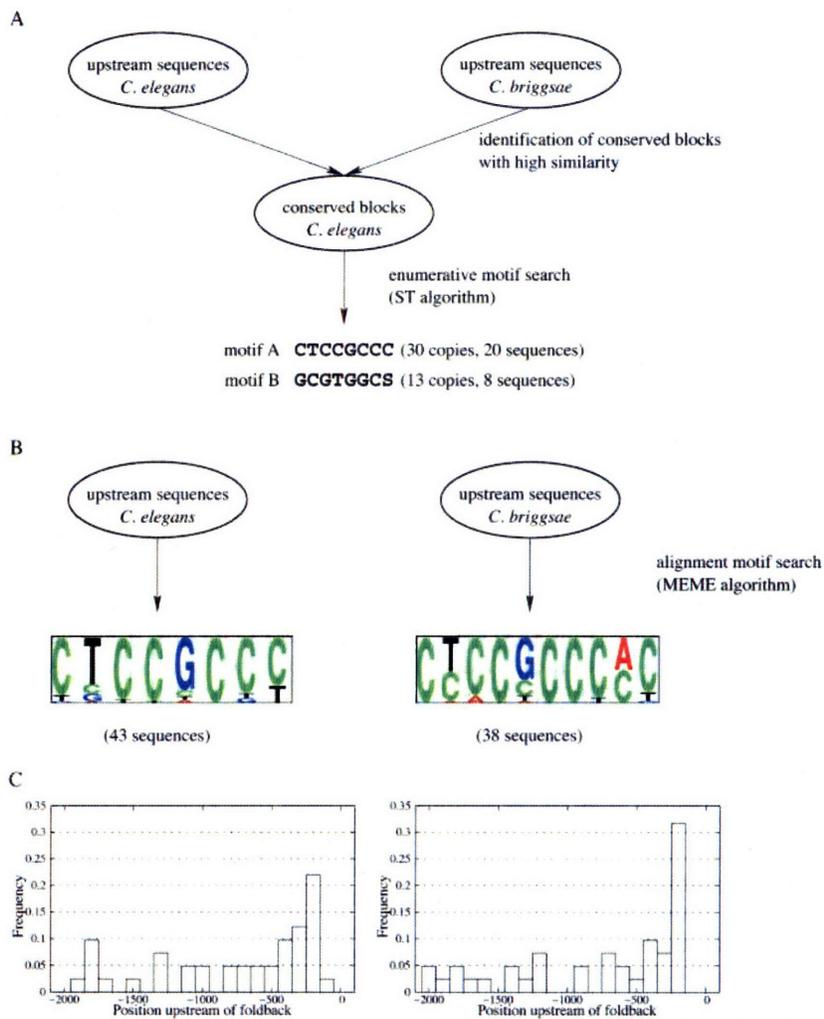


FIGURE 2. Identification of conserved upstream sequence elements. (A) Enumerative search for over-represented 8-mers within conserved upstream regions. Next to each consensus sequence is the number of instances of this sequence in conserved *C. elegans* blocks allowing for zero or one mismatch to the consensus or its reverse complement, and the number of distinct upstream sequences containing these instances. The Z-score of the consensus motif A was 29.0, the score of motif B was 14.7. As a control, a search in equally sized, randomly generated sequences delivered a Z-score of 11.2. (B) Application of the MEME local alignment algorithm to the complete 2000-bp upstream sequence sets. Shown are the pictograms (<http://genes.mit.edu/pictogram.html>) computed from the sequences that were used in the alignment by MEME for *C. elegans* (E-value of $3.0e-24$) and *C. briggsae* (E-value of $1.5e-37$). Both methods identify a highly similar motif as the most significant one. (C) Histograms of the locations of the best hit per sequence to the motifs given in B, in bins of 100 bp.

vation of the best hit in orthologous loci (data not shown). A similar motif was described previously in *C. elegans* introns (Fig. 2 in Lim and Burge 2001), and we also observed similar motifs downstream of protein-coding genes (data not shown). The common occurrence of this motif in introns argues against a role in transcriptional termination. Together, its relatively low statistical significance and ubiquitous distribution suggested that presence of this motif would not be a useful discriminatory feature for miRNA gene finding.

The polyadenylation-related motif AAAWTRAAA (Brown 2002) was the most significant motif computationally identified downstream of *C. elegans* protein-coding genes using MEME. No similar motif was identified in the sequences downstream of independently transcribed miRNAs. Therefore, although a subset of miRNA primary transcripts could be polyadenylated, polyadenylation does not appear to be a general feature of *C. elegans* miRNA transcripts. The absence of an identifiable polyadenylation signal does not rule out the possibility of pol-II-driven transcription, because other RNA genes, such as yeast snoRNAs, are derived from nonpolyadenylated pol-II transcripts (Steinmetz et al. 2001).

Finally, we examined the sequences around *C. elegans* miRNA foldbacks to search for candidate elements involved in the recognition and processing of the foldback from the primary transcript. As known foldbacks in polycistronic clusters are located immediately adjacent to one another, we restricted the search to ± 15 bases around the start and end of the foldback. No significant motif was identified, suggesting that the processing of the foldbacks is driven more by their secondary structure than by any conserved sequence. This conclusion is consistent with recent biochemical studies of pri-miRNA recognition and processing (Lee et al. 2003).

Improvement in microRNA gene finding

Previous approaches for the computational identification of miRNA genes have focused only on the stem-loop portion of the genes (Ambros et al. 2003b; Grad et al. 2003; Lai et al. 2003; Lim et al. 2003a,b). Our previous efforts started with conserved 110-nt genomic segments that were predicted to form stem-loops and did not fully overlap with protein-coding regions (Lim et al. 2003a,b). After passing an initial threshold on secondary structure similarity, the foldbacks were ranked using the program MiRscan. MiRscan evaluates miRNA candidates by sliding a 21-nt window along each arm of the foldback and assigning log-odds scores for seven features: base pairing of the candidate to the other arm of the stem, base pairing in the remainder of the stem–

loop structure, conservation of the 5' and 3' halves of the candidate miRNA, distance of the 21-nt window from the terminal loop, symmetry of the internal loops and bulges, and the sequence of the initial pentamer (Lim et al. 2003b). Overlapping 110-nt segments from both strands were then merged, and the higher scoring candidates were carried forward.

The observed upstream sequence motif and the patterns of sequence conservation flanking the stem-loop portion of the miRNA genes motivated us to develop an improved miRNA gene-finding algorithm, which we call MiRscanII. From here on, the previous version will be referred to as MiRscanI when needed for clarity. For the identification of independently transcribed *C. elegans* miRNA genes, we included three additional features as follows: (1) the score of the best hit to the *C. elegans* motif A within 1000 bp upstream of the predicted stem-loop; (2) the percentage of sequence contained in conserved blocks with >80% identity in the 1000 bp upstream of the stem-loop; and (3) the percentage of sequence contained in conserved blocks within 1000 bp downstream of the stem-loop. Log-odds scores for these features were derived from the MiRscanI training set of 50 conserved nematode miRNAs (Lim et al. 2003b), and these scores were simply added to the MiRscanI log-odds scores to give MiRscanII scores. The scores range from -3.3 to +2.0 bits for feature 1, -2.0 to +1.6 bits for feature 2, and -1.4 to +0.9 bits for feature 3.

MicroRNA candidates located on the sense strand of introns in protein-coding genes were not scored with these new features, but were instead filtered on the basis of their conserved genomic context. We observed that 11 of 13 known miRNAs in this group were located in introns of orthologous host genes as annotated in the Ensembl database (Clamp et al. 2003). For one of the two exceptions, the *C. briggsae* miRNA was located just downstream of the annotated orthologous gene, and in the remaining case, no ortholog was annotated. Thus, we kept only those foldbacks that were situated within, or at most 5000 bp from the *C. briggsae* ortholog, or for which no ortholog was annotated.

We included four additional filtering steps to eliminate the following types of unlikely candidates that had been scored in our previous effort: (1) candidate stem-loops that were located within extremely short intergenic regions between genes transcribed in opposite directions (<100 nt to each gene); (2) candidates on the antisense strand

of an intron, where one end is too close to a splice site, leaving insufficient room for promoter or terminator sequences; (3) independent candidates with no upstream or downstream conservation whatsoever; and (4) candidates that overlapped an exon by >50 bp. Previously, all foldbacks were kept if they overlapped at all with noncoding sequence.

The third filter was the only one for which a known miRNA gene (*mir-238*) was lost. A possible explanation is that the *C. briggsae* ortholog assigned by BLAST in our procedure was not the true ortholog. The fourth criterion eliminated a surprisingly large number of candidates (~7000), implying that many exons overlap conserved secondary structures. The minimal overlap of 50 bp ensures that at least one arm of a miRNA stem-loop is located within an intron, and there is one case (*mir-62*) where one arm of a known miRNA stem-loop overlaps with an exon of a nearby gene (T07C5.1) on the sense strand in both species. Assuming that this portion of the pre-mRNA is not alternatively spliced, *mir-62* processing would be expected to compete with splicing, producing either the coding sequence or the miRNA foldback.

A flowchart of the filtering and rescoring of candidate foldbacks is shown in Figure 3. To allow a direct comparison, the same set of sequence windows was used as in our previous study. Of ~43,000 foldbacks obtained from align-

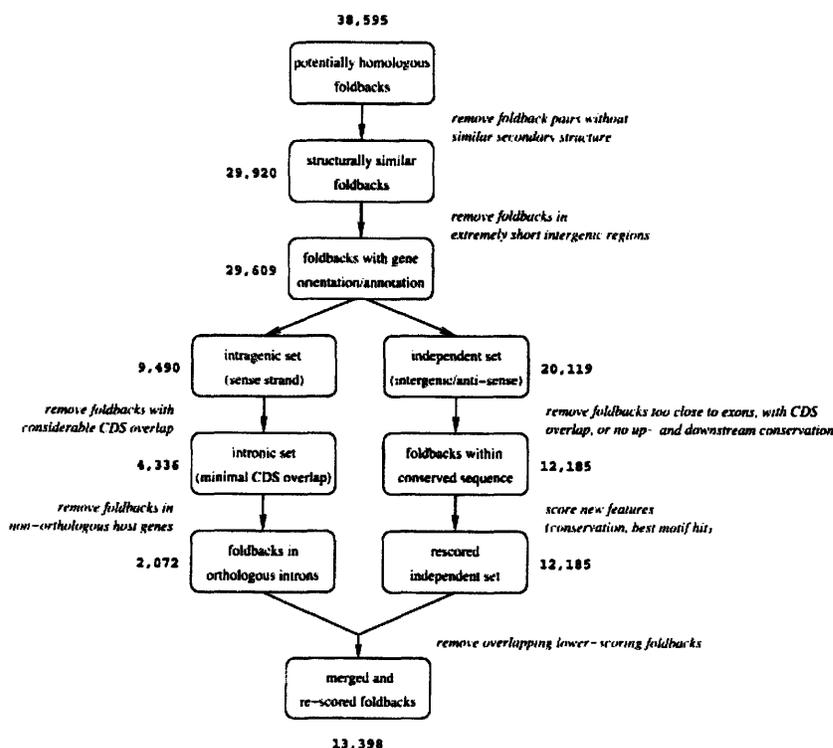


FIGURE 3. Flowchart of filtering and rescoring of candidate foldbacks with MiRscanII. Input was the set of conserved foldbacks that had received scores by MiRscanI. The numbers show how many candidates passed each step.

ments of the *C. elegans* genome with *C. briggsae* shotgun reads, ~35,700 passed the initial threshold on secondary structure similarity in MiRscanI, and were merged to ~28,000 nonoverlapping sequence windows. We realigned these ~43,000 foldbacks to the assembled *C. briggsae* genome sequence, recovering ~38,600 alignments. Of these, ~29,900 passed the secondary structure filter. All of the miRNAs previously scored by MiRscanI, as well as all previously tested candidates, were in this smaller set. After the additional filtering steps described above, the set of ~38,600 candidates was narrowed down to a mere ~13,400, as compared with ~28,000 previously (Fig. 3).

Compared with the previous analysis, the mode of the MiRscanII score distribution shifted from -4 to -9, and the

score range expanded from [-28,18] to [-30,23] (Fig. 4; Lim et al. 2003b). Of the 86 miRNAs cloned and/or detected by Northern in our previous study, 77 are scored by MiRscanII. The average score of these miRNAs increased by 0.9 bits when adding the new features, whereas the average score of all ~13,400 foldbacks decreased by 1.3 bits. In total, 73 miRNAs scored higher than nine bits, whereas four received low or negative scores. The remaining nine were not scored, either because a *C. briggsae* homolog was not identified by our automated methods or did not pass the folding free energy threshold (eight genes), or because flanking conservation was lacking (one gene).

The additional filters combined with the additional scoring features appear to have substantially increased the speci-

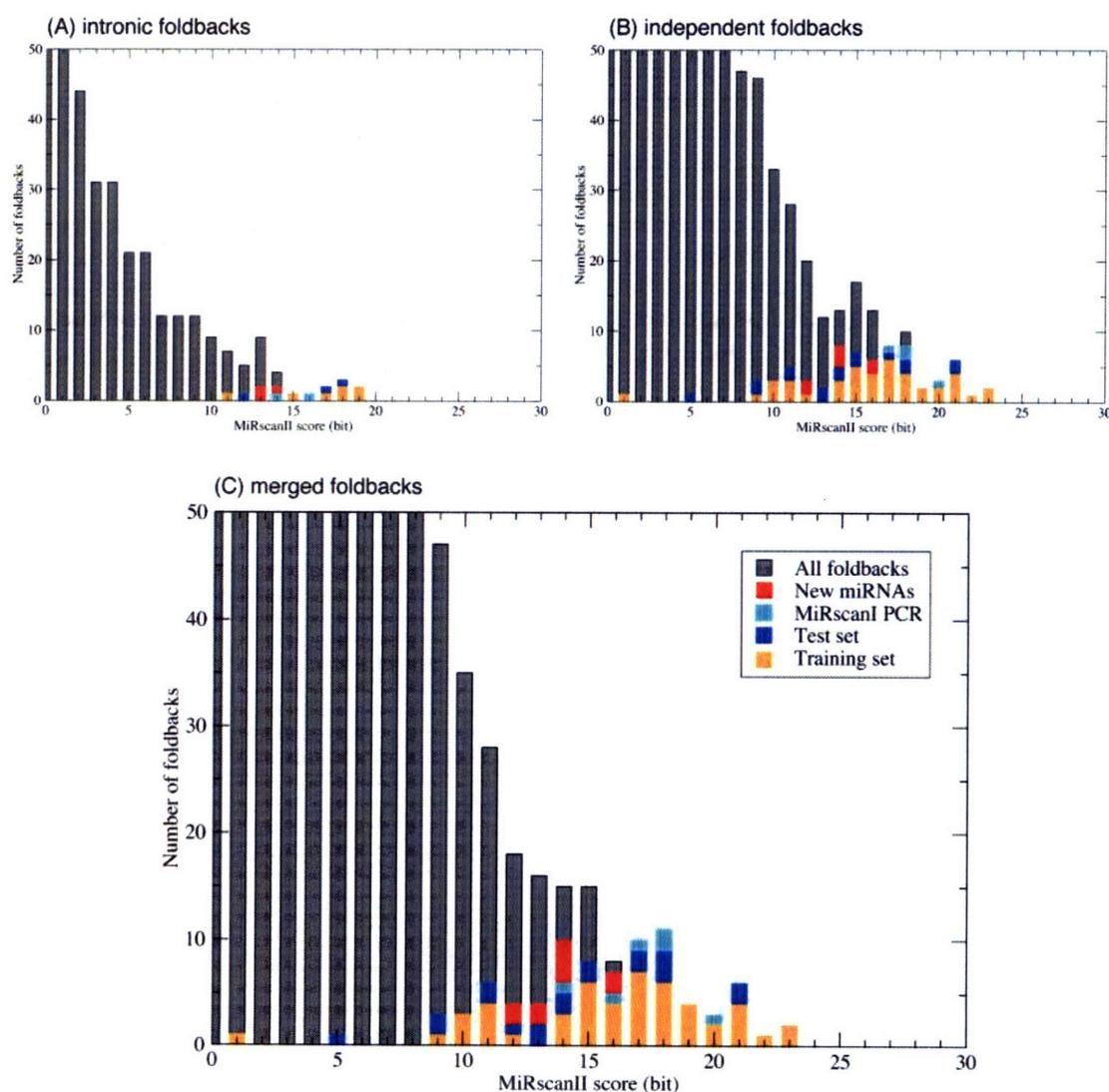


FIGURE 4. Histograms of MiRscanII scores greater than zero (nonsynthetic analysis). (A) Intronic foldbacks. (B) Independent foldbacks. (C) Merged set of 13,398 foldbacks. The training set (orange), test set (dark blue), previously verified MiRscanI predictions (light blue), and newly verified MiRscanII predictions (red) are marked in color. The score distributions were truncated at 50 foldbacks on the y axis. The scores of one miRNA gene in the training set (*mir-59*) was negative, and thus is not shown. Each bin covers a score range of one bit, e.g., the bin labeled 15 includes candidates with scores between 15 and 16 bits.

ficity of MiRscanII. In the MiRscanI analysis, 35 candidate miRNAs scored higher than the median score of cloned *C. elegans* miRNAs at the time, but 19 of these could not be confirmed by additional cloning or Northern blots (Lim et al. 2003b). Only seven of the 19 were left in the MiRscanII set of 13,400 rescored candidates, suggesting that the other 12 candidates are not, in fact, miRNAs. A total of 42 of 88 known miRNAs score higher than any unconfirmed sequence.

Because miRNAs are sometimes found within clusters, the relative position of miRNA candidates can provide a means of computationally identifying new genes, as was first shown in the prediction of *mir-39* and *mir-65*, two genes that were subsequently validated by expression analysis and/or cDNA cloning (Lau et al. 2001; Lim et al. 2003b). With this in mind, we scanned all candidates for their proximity to other candidates and retained those that were <1000 bp from each other, requiring a minimum score of five bits for all and eight bits for at least one candidate in such a potential cluster set. This simple algorithm recovered all known *C. elegans* miRNA clusters, and identified two additional potential clusters with two members each.

Experimental verification of new candidate miRNAs

Because MiRscanII more clearly distinguished previously identified miRNAs from other candidates, it was practical to examine new candidates with scores below the median of the test set. All unverified predictions that scored higher than the 43rd percentile of the test set miRNAs (12.7 bits) were subject to experimental screening. This set consisted of 35 new candidates plus six candidates that had not been detected in the previous attempt to validate computational

candidates by Northern blotting (Lim et al. 2003b). One of the clusters, which resulted from the cluster analysis described above, was part of this set due to high scores. The other cluster was additionally included, giving a total set of 43 candidates that were subject to experimental testing by PCR and subsequent cloning and sequencing to confirm the identity of the amplified product.

With this approach, we verified 10 miRNA candidates (Table 1), two of which, in retrospect, had been previously identified. One corresponded to miR-259, recently reported in (Ambros et al. 2003b) based on the perfect conservation of the miRNA in *C. briggsae* combined with its detectable expression on Northern blots. Our PCR-sequencing validation defined the terminus of this miRNA, which turns out to be shifted by 3 nt from the previously proposed position. The second previously identified miRNA in the set of 10 new validations was miR-239b, which had been previously proposed to be a homolog of miR-239a, but not experimentally verified (Lim et al. 2003b). Interestingly, sequencing of the PCR product from the miR-239b amplification revealed that the miRNA had a different 5' terminus than that seen for all four of the miR-239a clones. It was one nucleotide shorter on the 5' end, providing evidence that a second *mir-239* gene was indeed expressed and that the primer was preferentially hybridizing to miR-239b, rather than to miR-239a. Had we not seen this difference in the 5' termini of the miR-239 RNAs, it would have been difficult to argue against the possibility of primer cross-hybridization. Among the eight confirmed miRNAs that had not been previously proposed were two clustered candidates. The other clustered pair of candidates, which had scores lower than the 43rd percentile, was not validated by our PCR-sequencing protocol. Two of the eight newly identified

TABLE 1. Experimentally verified *C. elegans* miRNA candidates

miRNA	Sequence	Chr	Location	Arm	Sim
miR-353	caauugccauguguugguauu	I	intron of D1007.12 (s)	5'	+
miR-354	accuuguuuugucugcuccu	I	intron of Y105E8A.16 (s)	3'	+++
miR-355	uuuguuuuagccugagcuaug	II	1 kb ds of T27D12.3	5'	+++
miR-356	uugagcaacgcaacaaauc	III	intron of ZK652.2 (s)	5'	++
miR-357	uaauugccagucguugcagga	V	0.6 kb us of C10B5.1	3'	+
miR-358	caauugguaucccugucaagg	V	0.9 kb us of C10B5.1	3'	+
miR-359	ucacuggucuuucucugacga	X	0.5 kb ds of Y41G9A.6	3'	+
miR-360	ugaccgaaaucccguucacaa	X	0.5 kb us of Y23B4A.2	3'	+++
miR-392	uaucacgaucacgugugaug	X	1.0 kb us of F54B11.5	3'	+
miR-239b	uuguacuacacaaaaguacug	X	7.0 kb us of C34E11.1	5'	++
miR-259	aaauccuauccuauucuggua	V	1.2 kb us of F25D1.4	5'	+++
<i>lcy-6</i> miRNA	uuuuguagagacgcauuucg	V	0.5 kb us of C32C4.3	3'	++

The first nine rows show newly identified genes, the last three rows show the revised sequences for the successfully cloned, but previously described miRNAs miR-239b, miR-259, and *lcy-6*. The miRNAs are shown as 21-nt RNAs, but their actual length is generally not known because the PCR assay and sequencing validation determined the 5' but not 3' termini. The exception is the *lcy-6* miRNA for which the 21-nt length was deduced from the 5' terminus of the miRNA* and assuming Drosha processing leaving a 2-nt 3' overhang. For miR-358 and miR-360, some of the observed clones showed 5' ends shifted by 2 nt toward the 3' end. "Arm" denotes the side of the foldback on which the miRNA is located. The level of similarity (sim) with the miRNAs in the *C. briggsae* foldbacks are shown as +++ (100%), ++ (>90%), and + (>75%). For predicted stem-loops, see Supplementary Material at <http://genes.mit.edu/burgelab/MiRscanII>. (us) Upstream; (ds) downstream; (s) sense

miRNAs appear to be distant paralogs of previously identified *C. elegans* miRNAs; miR-357 and miR-356 have 5' homology with miR-232 and miR-233, respectively.

DISCUSSION

Conserved and nonconserved miRNAs—the limitations of current computational approaches

Like other computational miRNA gene finders, MiRscan misses genes that lack detectable homologs in related species. The observation that clear *C. briggsae* homologs were not readily found for 12 genes known at the start of this study (eight genes without MiRscanII scores, and four genes with low scores; Lau et al. 2001; Lim et al. 2003b) does not imply that these 12 miRNAs lack homologs in *C. briggsae*. Our previous analysis showed that 10 of these 12 miRNAs were related to other *C. elegans* miRNAs, which, in turn, had easily identifiable orthologs in *C. briggsae*, leaving only two miRNAs without an identifiable homolog (Lim et al. 2003b). Nonetheless, because of extensive divergence within certain of the families of paralogous genes, some of the *C. elegans* genes are not matched with their proper *C. briggsae* homologs in the BLAST searches of our automated analysis. A manual investigation of syntenic regions illustrated this limitation of our automated search for homologous miRNA stem-loops. Among the four related miRNA genes in a cluster on chromosome III, one (*mir-65*) had been matched to a homolog in *C. briggsae*, whereas three (*mir-64*, *mir-66*, and *mir-229*) had not been. A closer look at the syntenic locus in *C. briggsae* revealed two additional foldbacks flanking the previously identified *mir-65* ortholog. The putative miRNAs of these *C. briggsae* foldbacks matched residues 2–15 and 2–12 of the *C. elegans* miR-64 and miR-66 miRNAs. The other interesting case concerned *mir-72*, which was among the foldbacks with negative MiRscanII scores, and for which no orthologous foldback in *C. briggsae* had been previously reported. Inspection of the *C. elegans* locus showed that an alternative foldback structure, which placed miR-72 on the 5' instead of the 3' arm, was energetically more favorable than the structure proposed previously (Lau et al. 2001). An analysis of the syntenic *C. briggsae* region revealed a homologous foldback that resembled the revised *C. elegans mir-72* foldback, except that it had an extra stem protruding from near the terminal loop of the *C. briggsae* structure (see Supplementary Fig. 2 at the MiRscanII Web site, <http://genes.mit.edu/burgelab/MiRscanII>). This extra stem is reminiscent of that seen in the *C. elegans mir-229* foldback (see Supplementary Fig. 1 at <http://genes.mit.edu/burgelab/MiRscanII>; Ambros et al. 2003b; Lim et al. 2003b).

Computational miRNA prediction in other animals (Lai et al. 2003; Lim et al. 2003a) has utilized whole-genome alignments (WGAs) of related species to restrict the search space for conserved foldbacks. At the time MiRscan was developed, the *C. briggsae* genome was only available in the

form of short sequence reads, so there was no choice but to use BLAST searches of the reads to identify homologous foldbacks. To enable a direct comparison between the old and new versions, we decided to start with these foldbacks and realign them to the *C. briggsae* contigs. We therefore used the annotation of orthologous protein-coding genes to restrict the number of initially determined candidates, instead of starting from WGAs. The conservation of intronic miRNAs in orthologous host genes turned out to be a useful step for filtering of potential candidates (Fig. 3). We also explored the filtering of independently transcribed candidates in a similar manner (data not shown). First, we determined the *C. briggsae* orthologs of the closest flanking protein-coding genes. If both of these were located in the same *C. briggsae* contig, but the *C. briggsae* best match to the foldback under consideration mapped to a different contig, the foldback was eliminated. Of the 12,185 independent foldbacks (cf. Fig. 3), 45% did not pass this test, thus greatly reducing the number of candidates. Among those not passing the filter were four previously tested candidates that could not be verified by PCR and sequencing. However, this filter also eliminated five known miRNAs, including one of the newly identified ones, for which the corresponding *C. briggsae* sequence was not part of the same contig as the closest orthologous protein-coding genes. We checked all BLAST hits of these five miRNAs to the *C. briggsae* genome above the E-value threshold of 1.8 in more detail. In one case, the sequence in the syntenic *C. briggsae* location between the protein-coding genes also showed weak similarity to the foldback and might have been part of a WGA. The other four genes had no detectable similarity in the syntenic locus and would have been missed by a WGA, illustrating potential pitfalls of this approach, which can be confounded either by misassemblies, unusual rearrangements, or the selective loss of paralogs in different species. Lai et al. (2003) also reported that one of the first 24 *Drosophila* miRNAs to have been cloned was not part of the fly WGA, even though it was detectable by BLAST.

The above analysis suggested that a strict requirement for synteny would lower the sensitivity of the analysis, but we expected that demanding synteny would still be useful for increasing its specificity and might even lead to identification of a few additional miRNAs in cases for which a fortuitous BLAST hit to a nonsyntenic locus obscured the identification of the orthologous foldback pairs. The recent publication of the *C. briggsae* genome (Stein et al. 2003) contained a collection of 4837 syntenic blocks, that is, regions of long-range colinearity between the genomes of *C. elegans* and *C. briggsae*, allowing us to reconsider the synteny analysis in an alternative fashion. In total, these blocks covered 84.6% of the *C. elegans* and 80.8% of the *C. briggsae* genome. We repeated the complete MiRscanII analysis, this time restricting BLAST to match potentially homologous hairpins within the syntenic blocks only. We used the *C. elegans* sequence from release 77 of WormBase, because the

synteny coordinates were given with respect to this release. Of the 88 *C. elegans* miRNA foldbacks from Lim et al. (2003b) and the additional nine that we newly verified, 92 were contained in the syntenic blocks, and 81 were part of the final set of foldback pairs scored by MiRscanII, compared with 87 when we did not require synteny. This analysis yielded 17 foldback pairs with scores higher than 12.7 bits (our cutoff for experimental validation) that were not previously considered. The scores of five of these foldback pairs differed only slightly from the nonsyntenic analysis; their scores were pushed above the 12.7-bit threshold because of slight score fluctuations resulting from an independent analysis using a different genome assembly. When these 17 candidates were subject to experimental verification using our PCR-sequencing assay, only one, miR-392, was verified (Table 1). Even this one case did not result from the use of syntenic alignments; instead, sequence differences in the *C. elegans* genome versions used for the original MiRscan and the syntenic analysis led to an improvement in the *mir-392* foldback score.

Overall, demanding synteny for independently transcribed candidates provided essentially no improvement in MiRscan efficacy, and decreased the sensitivity of our approach without leading to the identification of any new genes missed by simply using the top BLAST hit in the genome, irrespective of its location. Nonetheless, considering synteny would likely provide substantial benefit to computational approaches with lower inherent specificity or to the application of MiRscan to more complex genomes.

A consideration of other recently reported miRNAs

Two other publications (Ambros et al. 2003b; Grad et al. 2003) have recently reported newly identified nematode miRNAs. Ideally, these could serve as additional independent test sets to assess the sensitivity of MiRscanII. Of the seven miRNAs uniquely reported in Ambros et al. (2003b), we found one (miR-259) by computational analysis and PCR sequencing. The other six (miR-256, miR-257, miR-258, miR-260, miR-261, and miR-262) are reportedly not conserved across species, and thus, were not in the initial set of foldbacks scored by MiRscanII. For so many of these newly reported genes to lack homologs in *C. briggsae* was unexpected, because *C. briggsae* homologs could be identified for all but two of the first 80 miRNAs cloned from *C. elegans* (Lau et al. 2001; Lee and Ambros 2001; Lim et al. 2003b). One possibility is that some have homologs, but these happen to fall in portions of the *C. briggsae* genome that have not yet been sequenced. To assess how MiRscanII would score these six miRNAs in the event that a homolog was eventually found, we applied the program to pairs of identical *C. elegans* sequences, assuming the best possible scenario of perfect conservation. Still, only two candidates scored above our experimental cutoff of 12.7 bits. This observation indicated that, conservation aside, most of these uniquely reported miRNAs have features that are atypical of

classical miRNAs. These features include an unusually long distance between the miRNA and the loop and less base pairing flanking the miRNA. Although not considered when originally formulating the criteria for miRNA annotation (Ambros et al. 2003a), base pairing flanking the miRNA is now known to be important for the nuclear processing of human primary miRNA transcripts by the enzyme Drosha (Lee et al. 2003). Because Drosha is conserved in nematodes and other metazoa, similar pairing is likely to be required in *C. elegans*.

The cloning effort that identified miR-256, miR-257, miR-258, miR-260, miR-261, and miR-262 also identified 33 unique tiny noncoding RNAs (tncRNAs), which differ from miRNAs in that they are not evolutionarily conserved, do not have the potential to be derived from miRNA-like precursors, and often begin with a G (Ambros et al. 2003b). With their lack of *C. briggsae* conservation and their atypical hairpin structures, a case could be made that most of these six uniquely reported miRNAs are instead tncRNAs, that is, they comprise the few tncRNAs that happened to have fortuitous potential pairing to flanking genomic sequence that was sufficient to satisfy the guidelines at that time for classification as miRNAs. Most of these six RNAs are also similar to the tncRNAs in another important aspect; their expression requires particular proteins of the RNAi pathway not generally needed for miRNA expression, further indicating that most of these six would be more accurately classified as tncRNAs (V. Ambros, pers. comm.).

None of the validated MiRscanII candidates matched the 10 miRNAs uniquely reported by Grad et al. (2003), which were assigned names *cp-miR-264* to *cp-miR-273*, where *cp* stands for computationally predicted. With the exception of *cp-mir-268*, none of the *cp*-miRNA foldbacks have easily identified *C. briggsae* orthologs. Two (*cp-mir-264* and *cp-mir-272*) have atypical foldbacks, as revealed by their poor MiRscan scores when compared against themselves. The eight remaining *cp*-miRNAs were initially found as homology candidates, that is, *C. elegans* hairpins that had segments with loose sequence similarity to previously known mature animal miRNAs, usually miRNAs of *C. elegans*. One possibility is that these foldbacks are distant paralogs of *C. elegans* miRNAs, not all of which might be conserved between species. Another possibility is that some of these foldbacks are in fact not miRNA genes, even though their authenticity was supported by a PCR assay (Grad et al. 2003). The PCR verification protocol used was less stringent than ours because it used the complete miRNA 21mer as a primer and lacked an additional sequence-verification step. Without this additional step, we would have counted an additional 10 of our 43 candidates as new miRNAs because they resulted in clear bands of the right size (35–45 nt). However, they did not pass the subsequent sequence-verification test. *cp-miR-268*, which received a score above our cutoff for experimental validation, was one of our candi-

dates with a PCR band that did not pass the sequence-verification test. Further supporting the idea that some cp-miRNAs are not authentic paralogs is the observation that in three cases (cp-miR-267, cp-miR-268, and cp-miR-271), the presumed mature miRNA resides on the opposite arm of the foldback when compared with the presumed paralog (miR-55, miR-73, and miR-35, respectively). Of the five remaining foldbacks, *cp-mir-266* and *cp-mir-273* look the most promising, in that each has additional sequence similarity with its presumed paralog (*mir-72* and *mir-56*, respectively) that falls outside of the mature miRNA in a pattern that might be expected for authentic paralogs. In addition, cp-miR-269 can be regarded as a paralog of cp-miR-266, as they differ by only three nucleotides.

The recent discovery of the *lsy-6* miRNA gene (Johnston and Hobert 2003), which appears to be expressed in only eight cells of the adult nematode, raises the question as to whether our strategy of computational prediction and large-scale cloning might lack the sensitivity to detect this and similar cases. The reported *lsy-6* foldback pair scored 9.91 bits with MiRscanII, including a positive contribution of the upstream motif described in this study. Our computational pipeline also included the opposite strand of the *lsy-6* locus, which scored slightly better (10.27 bits), including a negative contribution of motif A, because the orientation was incorrect. This score was at the 29th percentile of our test set, and therefore not high enough to be included in the set targeted for experimental verification. To determine whether we would have been able to validate the *lsy-6* gene if we had tested candidates down to the 29th percentile, we applied our PCR-sequencing assay and detected the *lsy-6* miRNA, showing that this assay is sufficiently sensitive to detect a miRNA expressed in only a few cells of the animal. The assay also detected the *lsy-6* miRNA* arising from the opposite arm of the hairpin and presumably present at even lower abundance in our library of small RNAs. These RNAs had not been detected previously, and the sequencing of their 5' termini performed in the course of the assay enabled us to define the mature *lsy-6* miRNA (Table 1). In summary, *lsy-6* is one of the anticipated miRNAs with a score somewhat below our current cutoff for experimental tests, but not otherwise unusual, and can be readily detected by the PCR-sequencing assay despite its restricted expression.

The estimated number of miRNA genes in *C. elegans*

Starting from MiRscanI predictions, we previously estimated that there were at least 93, but no more than ~120 miRNA genes in *C. elegans* (Lim et al. 2003b). The identification of additional miRNA genes, together with the increased specificity of MiRscanII, allows us to revisit these estimates. The 88 miRNA loci listed in our previous study and the 11 miRNA genes of Table 1 not present in the previous list add up to 99 unique loci. Nineteen of these were not among our 3423 sequenced miRNA clones (Lim et al. 2003b), and instead, were primarily identified by experi-

mentally verifying MiRscan predictions. We attempted to validate only those MiRscan candidates with scores above the 43rd percentile of the miRNAs in our test set, and all of these 19, with the exception of *lsy-6*, scored higher than the threshold. It is therefore reasonable to assume that these 18 miRNA genes include no more than 57% of the miRNA genes not represented among our 3423 clones. This implies that at least another 12 genes resembling the *lsy-6* miRNA have escaped our detection or validation efforts, because they either have no MiRscan scores or low scores. Thus, the current analysis enables the estimated lower limit on the number of miRNA genes in *C. elegans* to be revised upward to 99 + 12, or 111.

An upper limit of ~120 *C. elegans* miRNA genes was originally estimated by considering the number of MiRscanI candidates (validated genes together with nonvalidated candidates) that had scores exceeding the median score of the cloned miRNAs (Lim et al. 2003b). Because the cloned miRNAs included miRNAs without recognizable *C. briggsae* homologs, this calculation took into account poorly conserved miRNAs without MiRscan scores. Furthermore, the absence of a correlation between the number of times an miRNA was cloned and its MiRscan score argued against the idea that there might be a disproportionate number of *C. elegans* genes that have escaped detection because they are both difficult to clone and difficult to identify computationally (Lim et al. 2003b). Our confidence in this upper limit increases with the improved specificity of the current analysis. For instance, there is now reason to suspect that eight of the unvalidated candidates used to calculate this upper bound of ~120 are false positives, in that these eight had too much exon overlap to be considered in the current analysis. However, we do not attempt to revise the estimate on the upper bound of *C. elegans* miRNA genes because of the danger of some overfitting in the current analysis. For example, the more complicated and bifurcating set of filters and scoring schemes of the current analysis (Fig. 3) made it less amenable to jackknifing, a procedure implemented earlier so that the scores of genes from the training set could be considered when estimating the upper bound on the number of genes (Lim et al. 2003b). Because the status of many of the miRNAs uniquely reported by Ambros et al. (2003b) and Grad et al. (2003) is in doubt, we did not consider these candidates when estimating the lower and upper bounds of gene numbers in *C. elegans*. Thus, our estimate of ~110 to ~120 miRNA genes in *C. elegans* would have to be revised upward if future experiments overturn the idea that most of these candidates are not authentic miRNAs. Finally, the MiRscan pipeline to detect conserved foldback pairs excluded foldbacks with extreme GC- or AT-content, and filtered out sets of highly repetitive foldbacks, the members of which overlapped with RepeatMasked sequences (Lim et al. 2003b). We are currently investigating the extent to which such foldbacks potentially harbor noncoding RNA products.

Analysis of conserved upstream sequence elements

Our analysis of sequences upstream of independently transcribed nematode miRNAs identified a conserved sequence element, motif A with consensus CTCCGCCC, which is highly specific and useful for miRNA gene identification. The transcription factor database TRANSFAC (version 6.0 public; Matys et al. 2003) contains only a handful examples of nematode transcription factors, and none of them matched motif A. A literature search also failed to turn up any previously reported similar nematode sequence motifs. At this point, it is open as to whether motif A is a transcription-factor binding site, whether it is a signal that directs an miRNA processing enzyme to the miRNA genes, or whether its function is possibly related to both of these alternatives. Recent studies have shown that there is considerable coupling between transcription initiation and mRNA processing, in which transcription factors assist in the direction of splicing factors to the nascent transcript (Maniatis and Reed 2002). One can easily envision an analogous scenario for efficient recruitment of factors responsible for recognition and processing of miRNA stem-loops.

We also identified a common enriched sequence element in vertebrates, CCCWCCC, which was different from that found in nematodes. A second enriched sequence element, ATGCAT, occurred in only a subset of vertebrate upstream sequences and was also found in a subset of *Drosophila* upstream sequences. According to TRANSFAC, Sp-1 and POU1F1, respectively, are likely candidates for transcription factors that bind to these motifs. Sp-1 is a ubiquitous transcription factor, which has been shown to activate transcription. The occurrence of multiple instances of the first motif is consistent with binding by Sp-1, which often binds to several sites per regulatory region (Courey et al. 1989). POU1F1 is a growth hormone factor that contains one POU and one homeobox domain and also acts as a transcriptional activator (Lefevre et al. 1987). POU1F1 is not conserved in *Drosophila*, but other members of the same family of POU-homeobox-containing transcription factors with potentially similar binding preferences are present.

In none of the organisms under consideration—nematodes, arthropods, and vertebrates—were we able to identify strong motifs reminiscent of known eukaryotic core promoter sequence elements such as the TATA box. Even in the case of *Drosophila*, where a recent study has extended the set of motifs prevalent in core promoters, and reliable computational tools for pol-II transcription start site prediction are available (Ohler et al. 2002), no clear picture emerges at this point. Therefore, miRNA promoters do not share a common layout, but instead appear to be highly variable, as is characteristic of protein-coding gene promoters. In another parallel to protein-coding genes, a recent study showed that sequence elements as far as 1000 bp or more

upstream are required for specific activation of the *let-7* gene (Johnson et al. 2003).

In summary, our efforts showed that features distinct from RNA primary and secondary structure, such as upstream and downstream conservation and an upstream sequence motif, lead to a considerable improvement in gene-prediction accuracy for an important family of noncoding RNAs. Our improved method enabled us to identify nine new miRNA genes that had gone undetected, despite previous computation and large-scale cloning efforts. The set of known conserved nematode miRNAs is now approaching completeness, which should aid efforts to identify their target genes and to understand their roles in the *C. elegans* regulatory circuitry.

MATERIALS AND METHODS

Data sets

We constructed sets of orthologous upstream and downstream regions of independently transcribed miRNAs from a total set of 88 nematode miRNA genes (Lim et al. 2003b). First, we identified *C. elegans* miRNAs located in intergenic regions or on the anti-sense strand of introns, that are therefore likely to be transcribed independently of nearby protein-coding genes (WormBase annotation release 83). Next, we aligned the ~22-nt miRNA sequences to the assembled *C. briggsae* genome (July 2002) with BLAST (Altschul et al. 1997), retaining only those with >90% identity, that is, with no more than two mismatches. This stringent requirement should exclude the possibility of aligning upstream regions of related but nonorthologous miRNA genes. We then extracted up to 2000 bp upstream of both *C. elegans* and *C. briggsae* fold-backs for the Upstream Sequence Set (USS), and up to 1000 bp downstream for the Downstream Sequence Set (DSS), excluding overlaps with annotated *C. elegans* genes. For miRNAs in clusters, only the regions upstream of the first miRNA were included in the USS, and only the regions downstream of the last miRNA were included in the DSS, leaving 43 miRNA pairs. For three *C. elegans* genes (*mir-45*, *mir-77*, and *mir-90*), two sequences in *C. briggsae* met all of the above requirements, and both were included in the analysis.

For training and evaluation of the revised model, we started from the same set of 88 miRNA genes. We used a training set of 50 sequences as described in our previous study (Lim et al. 2003b), excluding *mir-88* with an unknown processed miRNA sequence. The 24 miRNA genes newly cloned in the same study were kept as an independent test set. miRNAs that had not been cloned, but had been identified only by experimental validation of computational predictions, were excluded from both the training and test sets. Three miRNAs in the test set were not scored, because our automated procedure did not find an orthologous candidate fold-back.

A set of 59 sequence pairs upstream of orthologous human-mouse miRNA genes (Lim et al. 2003a) were chosen in the same fashion as described for nematode miRNAs. Finally, 31 sequences upstream of independently transcribed *D. melanogaster* miRNAs according to the above criteria were taken from Aravin et al. (2003).

Alignment of upstream and downstream regions

We aligned the orthologous sequence pairs with the probabilistic sequence alignment tools BayesBlockAligner (BBA; Zhu et al. 1998) and Dynamic Block Aligner (DBA; Jareborg et al. 1999). Both programs have been specifically designed to identify short, highly conserved blocks in an alignment of two sequences, a pattern that can be expected in promoter sequences where transcription-factor binding sites are surrounded by stretches of nonconserved sequence. They perform a global alignment of two sequences, effectively ignoring stretches of unalignable sequences.

DBA uses a pair-hidden Markov model and computes the optimal alignment under a model of several match states corresponding to four different levels of conservation (with an average identity of 65%, 75%, 85%, and 95%). It requires colinearity of the two sequences, but allows for gaps within the conserved blocks. We retained blocks with at least 70% identity for the identification of motifs, and at least 80% for the feature computation in miRNA gene finding. The following parameter settings were used: block open probability 0.03, block close probability 0.98, gap probability 0.01.

BBA samples from the set of all possible alignments, covering a range of different substitution matrices and numbers of blocks. The output is the posterior probability that a specific position in one sequence is contained in an ungapped conserved sequence block with any position in the other sequence. In principle, these blocks are not required to be colinear. We considered all positions with posterior probability of at least 0.4 to be in an aligned conserved sequence block. We used PAM matrices from PAM5 to PAM30 in steps of 5 and base blocksize of 20.

In the case of multiple orthologs in *C. briggsae*, we merged the aligned blocks in *C. elegans* from all pairwise alignments. To avoid missing modestly conserved segments, we merged the output of both programs for the motif identification task. Because DBA and BBA deliver largely similar results, and the time complexity of the BBA algorithm is much higher, we restricted ourselves to DBA for the alignments scored by MiRscanII.

Two approaches for motif finding

We used an efficient implementation of the algorithm described by Sinha and Tompa (2000), here called the ST algorithm, which identifies statistically over-represented oligomers in a target set of sequences when compared with a background Markov chain model (H. Köstler, G. Stemmer, and U. Ohler, unpubl.). The algorithm uses a third-order Markov chain as a model for the background sequences and corrects for self-overlapping and complementary motifs. The motifs are composed of the standard A,C,G,T characters, but may also contain up to two ambiguous characters (N, S, W, R, Y). We retained all motifs with Z scores higher than a threshold obtained by a search in sequence sets of identical size, generated randomly with the same background distribution. We post-processed the resulting list of often highly similar significant oligonucleotides to determine how many distinct motifs were present. Details of this strategy to obtain motifs from lists of over-represented words have been given for a similar application elsewhere (Fairbrother et al. 2002).

We also used the probabilistic local alignment tool MEME (Bailey and Elkan 1995), with standard single-nucleotide frequencies

as background, motif length 5–10 bases, and “zero or one occurrence” mode. MEME motif E-values refer to the expected number of motifs of the same width with equal or higher likelihood in the same number of random sequences with the same nucleotide composition as the considered set of sequences.

Parameter estimation for additional features scored in MiRscanII

We derived log-odds scores for the upstream and downstream features in the following way: (1) 1 kb upstream and downstream of the foldback window of 110 bp—or less, if an annotated exon was closer—were aligned with DBA. (2) From these blocks, we obtained the percentage of nucleotides contained in blocks of 80% or more sequence identity, and used these values as features representing upstream and downstream sequence conservation. (3) For the foldbacks that had passed the initial filter of containing at least some conservation (see Fig. 3), a discrete distribution was obtained by binning the feature values in intervals of five percentage points. (4) As the foreground distribution for true miRNAs was restricted to a small set of values, we took two measures to prevent overfitting to the scarce data and to allow for reasonable scores for foldbacks that might have features just outside the range of observed values. First, the discrete distributions of both foreground and background were smoothed with two iterations of a mean filter of width 3 bins, with 0.75 weight for the central value and 0.125 for the values to the left and right. By doing so, we spread a small amount of probability to unseen values adjacent to the range of observed values. As an example, if we saw 20%–40% conservation in the foreground sequences, this filter would extend the range of positive foreground values to 10%–50%. Next, we truncated the foreground and background distributions at the last foreground value with positive probability on both ends of the range. The background values at the low and high cutoffs were set to the sum of all bins below or above the cutoff, respectively. In our example, we would set the background value at the 10% bin to the sum of all values below and up to 10%, and the 50% bin to the sum of all values equal or higher than 50%. Thus, we do not rely on arbitrary scores for feature values in the range where we do not see any positive probabilities even after smoothing. (5) From these modified distributions, log-odds scores were computed as the base 2 logarithm of the ratio of foreground to background probability.

To judge the presence of the promoter motif, we used the tool patser-v3d (Hertz and Stormo 1999) to compute the score of the best hit within the 1-kb upstream sequence on either strand. From these values, discrete distributions for foreground and background were obtained using bins of 5 bits, and these distributions were smoothed and converted to log-odds scores as above. We also reapplied the above smoothing procedure for the set of seven features used by MiRscan, and used these slightly different parameter sets instead of the original ones.

PCR-sequencing assay

A PCR-sequence assay identical to the one described in Lim et al. (2003b) was performed to detect the sequences of predicted miRNAs within a cDNA library constructed from 18 to 26 nt RNAs. This library was the same as the one used for cloning (Lau et al.

2001). As specific primers, 17-nt-long sequences complementary to the 3' ends of the predicted miRNAs were used, sometimes shifted by one or two nucleotides to prevent overlap with the primer to the generic 5' adapter sequences in the library. In some cases, the algorithm might correctly identify a miRNA foldback, but predicts the wrong strand or the wrong side of the foldback as the location of the mature miRNA. To account for this possibility, a second primer was also tested, corresponding to the second highest score from either the other side of the foldback or the other strand of the sequence.

Following PCR amplification, the products were cloned and sequenced to ensure that no primer-dimers were obtained, and to verify that the nucleotides between the two primers indeed matched the corresponding genomic sequence. This step also identified the 5' end of the miRNA; along with the greater sensitivity, this is a second advantage of this validation method compared with Northern blotting.

Primers for the successful reactions were as follows:

GCAATAATACCAACACA (miR-353),
 AGGAGCAGCAACAAACA (miR-354),
 ATTTGTTTCGCGTTGCTC (miR-355),
 CGAACTCCTGCAACGAC (miR-356),
 TGAGACCTTGACAGGGA (miR-357),
 CGTCAGAGAAAAGACCAG (miR-358),
 TTGTGAACGGGATTACG (miR-359),
 AGCTCAGGCTAAAACAA (miR-360),
 TCATCACACGTGATCGA (miR-392),
 CCAGTACTTTTGTGTAG (miR-239b),
 ACCAGATTAGGATGAGA (miR-259),
 ATGATTTTGATACTAGA (*lcy-6* miRNA),
 CATCGAAATGCGTCTCA (*lcy-6* miRNA*).

In all cases but miR-360 and *lcy-6*, the algorithm correctly identified the strand of the mature miRNA. In these two cases, the difference to the second highest score from the reverse strand was <0.4 bits.

Additional data files

Additional data files containing the following supplementary information are available through the Burge Lab Web site, <http://genes.mit.edu/burgelab/MiRscanII>: Supplementary Figure 1, foldback structures of newly identified miRNA genes; Supplementary Figure 2, foldback structure of the revised mir-72 locus; Supplementary Figure 3, examples of upstream alignments.

ACKNOWLEDGMENTS

We thank Harald Köstler and Georg Stemmer for work on the ST algorithm, Colleen T. Webb for sharing the data set of orthologous protein coding genes, Matthew W. Rhoades for help with the set of conserved mammalian miRNAs, the Genome Sequencing Center at the Washington University School of Medicine, St. Louis, and the Sanger Institute, Cambridge, UK, for sharing the assembled *C. briggsae* genome sequence before publication, N.C. Lau for providing the cDNA library of small *C. elegans* RNAs, and Victor Ambros for sharing unpublished results. This work was supported by grants from the NIH (C.B.B. and D.P.B.) and the Searle Scholars Program (C.B.B.).

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

Received October 14, 2003; accepted January 13, 2004; additional material accepted June 17, 2004.

REFERENCES

- Abrahante, J.E., Daul, A.L., Li, M., Volk, L.M., Tennessen, J.M., Miller, E.A., and Rougvie, A.E. 2003. The *Caenorhabditis elegans* hunchback-like gene *lin-57/hbl-1* controls developmental time and is regulated by microRNAs. *Dev. Cell* 4: 625–637.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.
- Ambros, V. 2003. MicroRNA pathways in flies and worms: Growth, death, fat, stress, and timing. *Cell* 113: 673–676.
- 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. 2003a. A uniform system for microRNA annotation. *RNA* 9: 277–279.
- Ambros, V., Lee, R.C., Lavanway, A., Williams, P.T., and Jewell, D. 2003b. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr. Biol.* 13: 807–818.
- Aravin, A.A., Lagos-Quintana, M., Yalcin, A., Zavolan, M., Marks, D., Snyder, B., Gaasterland, T., Meyer, J., and Tuschl, T. 2003. The small RNA profile during *Drosophila melanogaster* development. *Dev. Cell* 5: 337–350.
- Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E.G., Margalit, H., and Altuvia, S. 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* 11: 941–950.
- Bachelierie, J.-P., Cavaillea, J., and Hüttenhofer, A. 2002. The expanding snoRNA world. *Biochimie* 84: 775–790.
- Bailey, T.L. and Elkan, C. 1995. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach. Learn.* 21: 51–83.
- Bartel, D.P. 2004. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell* 116: 281–297.
- 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.
- Brown, T.A. 2002. *Genomes II*. Wiley, New York.
- Chen, C.Z., Li, L., Lodish, H.F., and Bartel, D.P. 2004. MicroRNAs modulate hematopoietic lineage differentiation. *Science* 303: 83–86.
- Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V., et al. 2003. Ensembl 2002: Accommodating comparative genomics. *Nucleic Acids Res.* 31: 38–42.
- Courey, A.J., Holtzman, D.A., Jackson, S.P., and Tjian, R. 1989. Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. *Cell* 59: 827–836.
- Eddy, S.R. 2001. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.* 2: 919–929.
- Fairbrother, W.G., Yeh, R.-F., Sharp, P.A., and Burge, C.B. 2002. Predictive identification of exonic splicing enhancers in human genes. *Science* 297: 1007–1013.
- Grad, Y., Aach, J., Hayes, G.D., Reinhart, B.J., Church, G.M., Ruvkun, G., and Kim, J. 2003. Computational and experimental identification of *C. elegans* microRNAs. *Mol. Cell* 11: 1253–1263.
- Harris, T.W., Lee, R., Schwartz, E., Bradnam, K., Lawson, D., Chen, W., Blasier, D., Kenny, E., Cunningham, F., Kishore, R., et al. 2003.

- WormBase: A cross-species database for comparative genomics. *Nucleic Acids Res.* **31**: 133–137.
- Hertz, G.Z. and Stormo, G.D. 1999. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**: 563–577.
- Jareborg, N., Birney, E., and Durbin, R. 1999. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9**: 815–824.
- Johnson, S.M., Lin, S.Y., and Slack, F.J. 2003. The time of appearance of the *C. elegans* *let-7* microRNA is transcriptionally controlled utilizing a temporal regulatory element in its promoter. *Dev. Biol.* **259**: 364–379.
- Johnston, R.J. and Hobert, O. 2003. A microRNA controlling left/right neuronal asymmetry in *Caenorhabditis elegans*. *Nature* **426**: 845–849.
- Jones-Rhoades, M.W. and Bartel, D.P. 2004. Computational identification of plant microRNAs and their targets, including a stress-induced miRNA. *Mol. Cell.* **14**: 787–799.
- 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.
- 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., Tomancak, P., Williams, R.W., and Rubin, G.M. 2003. Computational identification of *Drosophila* microRNA genes. *Genome Biol.* **4**: R42.
- 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., Jeon, K., Lee, J.T., Kim, S., and Kim, V.N. 2002. MicroRNA maturation: Stepwise processing and subcellular localization. *EMBO J.* **21**: 277–284.
- Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., Lee, J., Provost, P., Radmark, O., Kim, S., et al. 2003. The nuclear RNase III drosha initiates microRNA processing. *Nature* **425**: 415–419.
- Lefevre, C., Imagawa, M., Dana, S., Grindlay, J., Bodner, M., and Karin, M. 1987. Tissue-specific expression of the human growth hormone gene is conferred in part by the binding of a specific *trans*-acting factor. *EMBO J.* **6**: 971–981.
- 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. and Burge, C.B. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci.* **98**: 11193–11198.
- 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., 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.
- Lin, S.Y., Johnson, S.M., Abraham, M., Vella, M.C., Pasquinelli, A., Gamberi, C., Gottlieb, E., and Slack, F.J. 2003. The *C. elegans* hunchback homolog, *hbl-1*, controls temporal patterning and is a probable microRNA target. *Dev. Cell* **4**: 639–650.
- Llave, C., Kasschau, K.D., Rector, M.A., and Carrington, J.C. 2002. Endogenous and silencing-associated small RNAs in plants. *Plant Cell* **14**: 1605–1619.
- Maniatis, T. and Reed, R. 2002. An extensive network of coupling among gene expression machines. *Nature* **416**: 499–506.
- Matys, V., Fricke, E., Geffers, R., Gossling, E., Hanbuck, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., et al. 2003. TRANSFAC: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* **31**: 374–378.
- 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. and Niemann, H. 2001. Identification and analysis of eukaryotic promoters: Recent computational approaches. *Trends Genet.* **17**: 56–60.
- Ohler, U., Liao, G.-C., Niemann, H., and Rubin, G.M. 2002. Computational analysis of core promoters in the *Drosophila* genome. *Genome Biol.* **3**: RESEARCH0087.1–12.
- Park, W., Li, J., Song, R., Messing, J., and Chen, X. 2002. CARPEL FACTORY, a Dicer homolog, and HEN1, a novel protein, act in microRNA metabolism in *Arabidopsis thaliana*. *Curr. Biol.* **12**: 1484–1495.
- 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.
- Sempere, L.F., Sokol, N.S., Dubrovsky, E.B., Berger, E.M., and Ambros, V. 2003. Temporal regulation of microRNA expression in *Drosophila melanogaster* mediated by hormonal signals and broad-complex gene activity. *Dev. Biol.* **259**: 9–18.
- Sinha, S. and Tompa, M. 2000. A statistical method for finding transcription factor binding sites. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **Vol. 8** 344–354.
- Stein, L.D., Bao, Z., Blasiar, D., Blumenthal, T., Brent, M.R., Chen, N., Chenwalla, A., Clarke, L., Clee, C., Coghlan, A., et al. 2003. The genome sequence of *Caenorhabditis briggsae*: A platform for comparative genomics. *PLoS Biol.* **1**: E45.
- Steinmetz, E.J., Conrad, N.K., Brow, D.A., and Corden, J.L. 2001. RNA-binding protein Nrd1 directs poly(A)-independent 3'-end formation of RNA polymerase II transcripts. *Nature* **413**: 327–331.
- Webb, C.T., Shabalina, S.A., Ogurtsov, A.Y., and Kondrashov, A.S. 2002. Analysis of similarity within 142 pairs of orthologous intergenic regions of *Caenorhabditis elegans* and *Caenorhabditis briggsae*. *Nucleic Acids Res.* **30**: 1233–1239.
- 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.
- Yekta, S., Shih, I-H., and Bartel, D.P. 2004. MicroRNA-directed cleavage of HOXB8 mRNA. *Science* **304**: 594–596.
- Zeng, Y. and Cullen, B.R. 2003. Sequence requirements for microRNA processing and function in human cells. *RNA* **9**: 112–123.
- Zeng, Y., Wagner, E.J., and Cullen, B.R. 2002. Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. *Mol. Cell* **9**: 1327–1333.
- Zhu, J., Liu, J.S., and Lawrence, C.E. 1998. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**: 25–39.

Appendix III

The microRNA *miR-196* acts upstream of *Hoxb8* and *Shh* in limb development

Eran Hornstein¹, Jennifer H. Mansfield¹, Soraya Yekta², Jimmy Kuang-Hsien Hu¹, Brian D. Harfe³, Michael T. McManus⁴, Scott Baskerville², David P. Bartel² & Clifford J. Tabin¹

MicroRNAs (miRNAs) are an abundant class of gene regulatory molecules (reviewed in refs 1, 2). Although computational work indicates that miRNAs repress more than a third of human genes³, their roles in vertebrate development are only now beginning to be determined. Here we show that *miR-196* acts upstream of *Hoxb8* and Sonic hedgehog (*Shh*) *in vivo* in the context of limb development, thereby identifying a previously observed but uncharacterized inhibitory activity that operates specifically in the hindlimb. Our data indicate that *miR-196* functions in a fail-safe mechanism to assure the fidelity of expression domains that are primarily regulated at the transcriptional level, supporting the idea that many vertebrate miRNAs may function as a secondary level of gene regulation.

Sonic hedgehog (*Shh*) is a key signal mediating anteroposterior polarity in both the fore- and hindlimb buds⁴. Retinoic acid (RA) signalling is required for *Shh* expression in the forelimb and the hindlimb^{5–8}. The transcription factor *Hoxb8* seems to mediate the induction of *Shh* by RA in the forelimb in that *Hoxb8* is upregulated as an immediate-early response to ectopic RA administered to the chick forelimb bud⁷, and ectopic *Hoxb8* expression in the anterior of the forelimb of a transgenic mouse leads to *Shh* expression⁹. Ectopic RA does not lead to *Hoxb8* induction in the hindlimb bud, however, owing to the presence of an unknown hindlimb-specific inhibitory activity¹⁰.

Reasoning that the unknown hindlimb inhibitory activity¹⁰ might be mediated by a small silencing RNA, we blocked miRNA processing by using a conditional knockout allele of *Dicer*, a key enzyme required for producing functional miRNAs from their precursors^{11,12}. *Dicer* activity can be specifically removed from the limb buds by using a conditional allele¹³ and a limb-specific *Prx1::cre* construct¹⁴ (Supplementary Fig. 1a), which recombine floxed alleles efficiently in the limb mesenchyme (Supplementary Fig. 1b). To test whether the inhibition of *Hoxb8* induction by RA in hindlimb buds is relieved by the removal of *Dicer* activity, hindlimbs from *Dicer*^{Δfloxed/Δfloxed} and wild-type mice at embryonic day 11.5 (E11.5) were cultured in the presence of RA. As in chick limbs, the presence of RA led to a marked upregulation of *Hoxb8* messenger RNA in the forelimb tissue of both wild-type and mutant animals (Fig. 1a, b), but not in wild-type hindlimbs (Fig. 1c). In *Dicer*^{Δfloxed/Δfloxed} hindlimbs, however, RA induced the expression of *Hoxb8* (Fig. 1d). As previously shown¹³, loss of *Dicer* activity does not affect the expression of other known patterning genes in the developing limb bud (Supplementary Fig. 1c). Thus, the previously uncharacterized inhibitory activity¹⁰ is lost in the absence of *Dicer*.

Dicer is crucial for the processing of hundreds of miRNAs and many siRNAs. To identify specific candidate miRNAs that could be

responsible for the hindlimb-specific inhibitory activity downstream of *Dicer*, we used microarray analysis¹⁵. Of the miRNAs that are expressed in the limb primordia, 12 were at least twofold more abundant in either the forelimb or the hindlimb bud (Fig. 2a and Supplementary Table 1). The most differentially expressed miRNA in the screen was *miR-196*, with an expression signal in the hindlimb exceeding by 20-fold that in the forelimb (Fig. 2a and Supplementary Table 1). Differential *miR-196* expression was verified in northern blot analyses of RNA isolated from forelimbs and hindlimbs of both chick and mouse (Fig. 2b) and was also consistent with the expression domain suggested by a transgenic reporter study¹⁶. Intriguingly, *Hoxb8* mRNA is a known target of *miR-196 in vivo*^{16,17}. Therefore, we investigated whether *miR-196* might be the unknown hindlimb-specific activity preventing *Hoxb8* induction by RA.

First, to establish that *Hoxb8* is indeed an *in vivo* target of *miR-196* in the hindlimb, we carried out a modified 5' rapid amplification of

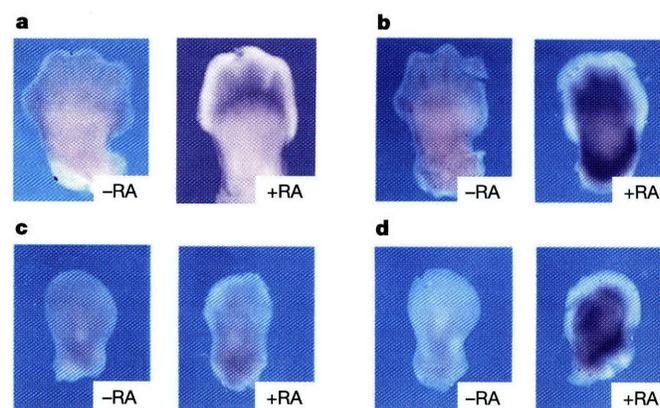


Figure 1 | Activity downstream of *Dicer* inhibits RA-induced expression of *Hoxb8* in mouse hindlimbs. **a**, E11.5 *Dicer*^{floxed/+}; *Prx1::Cre* (wild-type) forelimbs were cultured without RA (–RA), leading to no detection of *Hoxb8* ($n = 6/6$), or with 100 nM RA for 12 h (+RA), leading to induction of *Hoxb8* ($n = 6/6$). Expression of *Hoxb8* was detected by means of whole-mount *in situ* hybridization. **b**, E11.5 *Dicer*^{floxed}/*Dicer*^{floxed}; *Prx1::Cre* (*Dicer* knockout) forelimbs subjected to the same treatment similarly resulted in induction of *Hoxb8* only in the presence of RA ($n = 6/6$ negative, without RA; 6/6 positive, with RA). **c**, Hindlimbs from the mice in **a** were cultured similarly, and RA failed to induce *Hoxb8* expression ($n = 8/8$ with RA; 8/8 without RA). **d**, Hindlimbs from the *Dicer* knockout mice in **b** were cultured similarly. Complete deletion of *Dicer* did not result in induction of *Hoxb8* in untreated hindlimbs ($n = 5/5$), but it enabled the accumulation of *Hoxb8* transcripts in RA-treated hindlimb mesenchyme ($n = 6/6$).

¹Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, Massachusetts 02115, USA. ²Whitehead Institute for Biomedical Research, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, Massachusetts 02142, USA. ³Department of Molecular Genetics and Microbiology, University of Florida College of Medicine, Gainesville, Florida 32610, USA. ⁴Department of Microbiology and Immunology, Diabetes Center, University of California at San Francisco, San Francisco, California 94143, USA.

complementary DNA ends (RACE) protocol, commonly used as an assay for miRNA-directed mRNA cleavage^{17,18}. By sequencing the 5' RACE products, we could determine whether any amplified *Hoxb8* degradation products were cleaved precisely at the predicted *miR-196*-binding site. We could easily observe *miR-196*-directed *Hoxb8* cleavage in the wild-type hindlimb, whereas *Hoxb8* cleavage in the forelimb tissue was barely seen (Fig. 3a, b). These data indicate that *Hoxb8* is indeed both transcribed, at a level detectable by polymerase chain reaction (PCR), and cleaved *in vivo* in the hindlimb.

In wild-type chick embryo, after 2.5 d of incubation *Hoxb8* is expressed in the neural tube and somites. *Hoxb8* is also expressed in the forelimb field, where it functions in inducing *Shh* during the early limb field stages (Fig. 3c). To test whether *miR-196* activity could attenuate *Hoxb8* expression at the early limb field (stage 16), we used a replication-competent viral expression system (RCAS). Our analysis showed that 26 h after *in ovo* injection of the virus RCAS::*miR-196*, *Hoxb8* expression was reduced throughout the embryo and, in particular, endogenous expression of *Hoxb8* in the forelimb field was markedly repressed (Fig. 3d).

We next addressed whether *miR-196* could be responsible for the inability of ectopic RA to induce *Hoxb8* in the hindlimb¹⁰. We implanted RA-soaked beads into wild-type chick forelimbs, which induced *Hoxb8* within 4 h (Fig. 3e). By contrast, parallel implantations failed (or were only marginally able) to induce *Hoxb8* in forelimb buds ectopically expressing *miR-196* (Fig. 3f). Misexpression of *miR-196* in the forelimb thus creates a situation that is reminiscent of wild-type hindlimb, in which endogenously high expression of *miR-196* leads to observable degradation of endogenous *Hoxb8* and correlates with an inability of RA to induce ectopic *Hoxb8*.

The *miR-196*-sensitivity of *Hoxb8* thus provides a compelling explanation for the inability of RA to induce *Hoxb8* in the hindlimb. In previous studies^{7,10}, RA and *Hoxb8* were placed upstream of *Shh* expression in the forelimb and, indeed, blocking endogenous RA activity resulted in a significant, albeit incomplete, downregulation of endogenous *Shh* expression^{7,10}. If the *miR-196*-sensitivity of *Hoxb8* expression were truly involved in mediating RA-induced expression of *Shh* in the forelimb bud, then *Shh* expression itself should be downregulated on the introduction of *miR-196* into the forelimb. Indeed, when chick embryos were analysed 2 d after viral misexpression of *miR-196* in the right limb field, endogenous *Shh* was consistently downregulated (Fig. 4a, compare with 4b). Other genes, not described to be downstream of *Hoxb8* in the limb mesenchyme,

were not affected by misexpression of *miR-196*, suggesting that this was a specific effect (Supplementary Fig. 2). To quantify the effect of *miR-196* on *Shh* levels, we infected chick embryos as above and assayed them 2 d later by quantitative real-time PCR. *Shh* expression was decreased in the *miR-196*-infected forelimb to roughly a third of the level seen in wild-type limbs (Fig. 4c).

We also checked whether ectopic misexpression of *miR-196* would block RA-induced ectopic expression of *Shh*. When RA-soaked beads were implanted into wild-type chick forelimb for 36 h, an anterior domain of ectopic *Shh* was induced⁴ (Fig. 4d); however, in *miR-196*-infected limbs, *Shh* expression was blocked or diminished and more diffuse (Fig. 4e). Although *Shh* was repressed by *miR-196* misexpression in the forelimb, the expression of *Shh* in the hindlimb was not affected by the same manipulation (Fig. 4f, g). This difference highlights the rather unexpected conclusion that independent pathways control *Shh* expression in the forelimb and the hindlimb (Fig. 4h), which may be explained by a dual role for Hox genes in specifying forelimb versus hindlimb identity and in regulating *Shh* expression. After *Hoxb8* and other related Hox genes evolved to specify forelimb-specific morphology, a different, *Hoxb8*-independent, mechanism of regulating *Shh* downstream of RA had to evolve for the hindlimb.

Despite the evidence presented here and elsewhere⁹ supporting a role for *Hoxb8* in regulating *Shh* in the forelimb, it has been reported that even the removal of all three Hox8 paralogues has no effect on limb formation¹⁹, suggesting that this gene has possible redundancy with other Hox genes. In this respect, *Hoxa7* is also expressed in the posterior of the forelimb bud and is induced by RA^{20,21}. Moreover, we found that, like *Hoxb8*, *Hoxa7* is expressed in a forelimb-specific fashion (Supplementary Fig. 3). Intriguingly, *Hoxa7* is also a predicted target of *miR-196*, with several conserved matches to the 5' portion of the miRNA known as the 'seed'²². We did not observe changes in *Hoxa7* mRNA in response to *miR-196* misexpression (data

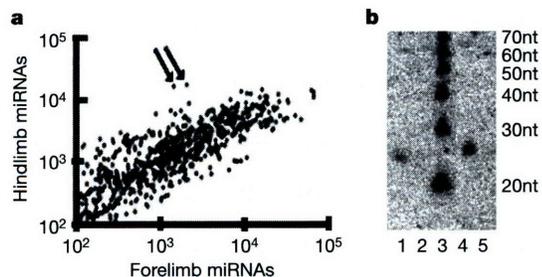


Figure 2 | Hindlimb-specific expression of *miR-196*. **a**, Representation of miRNA array analysis, comparing the expression of individual miRNAs (dots) in E10.5 mouse forelimb and hindlimb buds (in arbitrary units). Abundance of an individual miRNA in the hind- and forelimb is shown by its relative position along the logarithmically scaled y and x axes, respectively. Arrows indicate features corresponding to *miR-196*. **b**, Northern blot hybridization detected *miR-196* in extracts from hindlimbs of E10.5 mouse and stage-22 chick (lanes 1 and 4, respectively) but not in mouse and chick forelimb buds (lanes 2 and 5, respectively). Data are representative of four independent samples. The lengths of DNA oligomers (lane 3) used as size markers are specified next to the blot in nucleotides (nt).

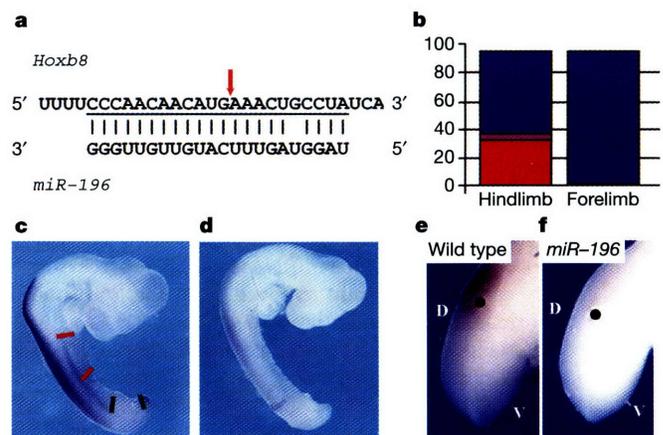


Figure 3 | *miR-196* downregulates *Hoxb8* accumulation. **a**, Sequence of the 3' UTR of *Hoxb8* complements *miR-196*. An arrow indicates the 5' end of the primary cleavage product. **b**, 5' RACE analysis in hindlimb and forelimb. Of the 96 hindlimb clones sequenced, 33 yielded a sequence consistent with *miR-196*-directed cleavage (red); four were also truncated *Hoxb8* clones, but cleavage was outside the miRNA-binding site (pink); and 59 were sequences unrelated to *Hoxb8* (blue). In the forelimb, no clones were consistent with *miR-196*-directed cleavage. **c**, By whole-mount *in situ* hybridization with a *Hoxb8* probe, an expression domain of *Hoxb8* was detected in the forelimb field (red bars), but not in the hindlimb field (green bars), of a stage-16 chick embryo ($n = 8/8$). **d**, Early pan infection with RCAS::*miR-196* resulted in downregulation of *Hoxb8* ($n = 6/6$). **e**, An RA-soaked bead implanted into the anterior aspect of a stage-22 wild-type forelimb induced *Hoxb8* expression ($n = 8/10$). **f**, Only marginal induction of *Hoxb8* expression was detected on implantation of an RA-soaked bead in a forelimb infected with RCAS::*miR-196* ($n = 6/8$). Anterior view; D, dorsal; V, ventral.

not shown), however, indicating that if *miR-196* is repressing *Hoxa7*, it is reducing *Hoxa7* protein without substantially destabilizing the *Hoxa7* transcript. Such a mechanism would be consistent with the results of a heterologous reporter assay showing that a *Hoxa7* untranslated region (UTR) fragment containing the *miR-196* seed matches predominantly mediates *miR-196*-dependent repression through the reduction of protein rather than mRNA levels¹⁷.

The experiments described here indicate that *miR-196* may be an *in vivo* inhibitor of *Hoxb8* in the hindlimb, and thereby may be responsible for the inability of ectopic RA to induce *Hoxb8* in the hindlimb. Low *Hoxb8* expression and *miR-196*-directed degradation was detected in the naive hindlimb bud by 5' RACE, indicating that *miR-196* activity is a component of *Hoxb8* regulation in the unmanipulated limb. Notably, however, loss of miRNA activity in the Dicer-deficient hindlimb did not, in itself, result in *Hoxb8* induction, suggesting that the primary level of regulation of forelimb-specific *Hoxb8* expression is transcriptional and independent of small regulatory RNAs.

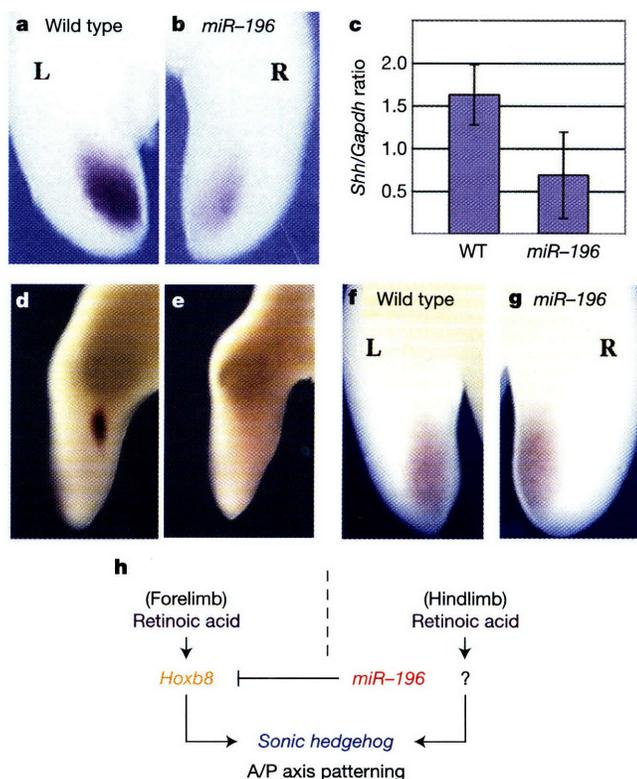


Figure 4 | *miR-196* downregulates *Shh* in the chick forelimb. **a**, Expression of *Shh* in the left (L) limb of stage-23 embryo. Posterior view ($n = 20/20$). **b**, In the right forelimb (R) of the same embryo, endogenous *Shh* expression was diminished 2 d after infection with *RCAS::miR-196* ($n = 18/20$). **c**, Three untreated sample tubes, each containing four forelimbs (stage 23), and three corresponding sample tubes with *RCAS::miR-196*-infected limbs were subjected to real-time PCR quantification of *Shh* mRNA. Three replicate runs were done on each sample tube. Blue bars represent the expression of *Shh*, normalized to *Gapdh*, in wild-type limbs (mean \pm s.d., 1.62 ± 0.35) and limbs infected with *RCAS::miR-196* (0.68 ± 0.51). The difference in the mean value between the *miR-196*-infected sample and the untreated control was significant (one-tailed *t*-test, $P = 0.029$). **d**, One and a half days after an RA-soaked bead (1 mg ml^{-1}) was implanted into the anterior aspect of stage-20 forelimbs, an ectopic *Shh* expression domain was detected by whole-mount *in situ* hybridization ($n = 6/6$). **e**, *RCAS::miR196* infection inhibited the ectopic expression of *Shh* in the anterior ($n = 5/8$). **f**, **g**, *Shh* expression was comparable in the left uninfected hindlimb (**f**) and the right *RCAS::miR196*-infected hindlimb (**g**) of chick embryos ($n = 20/20$). **h**, Model of the epistatic relations among *miR-196*, RA, *Hoxb8* and *Shh* in the developing limbs. A/P, anterior–posterior.

Thus, in normal limb development, the role of *miR-196* seems to be to safeguard against inappropriate Hox activity in the hindlimb. This conclusion fits well with the report that the genes that are downregulated when a miRNA is delivered to human cells are preferentially those that are expressed at low levels in tissues that normally express the miRNA²³. It thus seems that a chief role of some miRNAs in vertebrate development may be to prevent inappropriate activity of genes in domains where they are already repressed transcriptionally. Some miRNAs have been experimentally implicated to have roles in other facets of vertebrate development, including *miR-181* in haematopoiesis²⁴, *miR-430* in brain morphogenesis²⁵ and *miR-1* in heart development²⁶. In contrast to our findings, *miR-1* and its target *hand2* are predominantly expressed in the same cells, enabling *miR-1* to have a key role in regulating the switch between cardiomyocyte differentiation and proliferation²⁶. Together, these two studies indicate that these intriguing regulators of gene activity can take on diverse roles in coordinating vertebrate developmental and physiological processes.

METHODS

Mice and organ culture. Mice were housed and handled in accordance with protocols approved by the Institutional Animal Care and Use Committee of Harvard Medical School. Male mice carrying one copy of the *Prx1::Cre* allele and one *Dicer^{flxed}* allele were crossed to *Dicer^{flxed/flxed}* females. Cre recombinase, driven by the *prx1* enhancer, excises a required region in the RNase IIIb domain to yield a nonfunctional *Dicer* allele in limb buds¹³. Timed-pregnant females were killed at E11.5, embryos were dissected, and limbs were separately cultured in hanging drops. After 12 h of incubation in DMEM medium supplemented with 10% fetal calf serum, penicillin and streptomycin with or without 100 nM all-trans RA (Sigma), limbs were fixed in 4% paraformaldehyde for 4 h and processed for *Hoxb8* *in situ* hybridization.

MicroRNA–cDNA probe and expression array hybridization. Total RNA was isolated from E10.5 mouse fore- and hindlimbs with Trizol (Invitrogen) according to the manufacturer's instructions. Small RNAs were size-fractionated, ligated to adaptor oligonucleotides, reverse-transcribed and amplified. Labelled probes (Cy5 for the hindlimb sample and Cy3 for the forelimb sample) were hybridized to an expression array as described¹⁵. After hybridization, the array was scanned (Genepix pro 4000b; Axon) and analysed. Along with the vertebrate spots on the array, spots for all known *Caenorhabditis elegans* miRNAs are printed, most of which should not be hybridized to a vertebrate probe. Thus, background was set at a score equal to 95% that of the spots from the *C. elegans* section of the array¹⁵.

5' RACE of *Hoxb8*. Total RNA was obtained from a pool of 30 E10.5–11 mouse hind- and forelimbs and was subjected to modified 5' RACE as described¹⁷ with the following primers: 5'-CCATAAAGCAATTCACAGATACAGG-3' and 5'-GGTTGCGAGGAAAGATG-3'.

Generation of *RCAS::miR-196*. A 500-bp fragment of genomic DNA surrounding the chicken *miR-196-1* locus (chromosome 27, *HoxB* cluster) was amplified by PCR. An *ApaI* site was appended to the 5' end and an *EcoRI* site was appended to the 3' end by using the following primers (restriction sites are in parentheses): 5'-AATTCC(GGGCCC)CTCTATTTGTCAACTATTTGTAACG-3' and 5'-G(GAATTC)GCATTTTGGCCTCCGAGAGG-3'. The PCR fragment was then cloned, by means of the *ApaI* and *EcoRI* sites, downstream of the RNA polymerase III U6 promoter, into a pBS-U6 plasmid. The whole U6 promoter and *miR-196* genomic DNA were then excised with *ClaI* and cloned into the *RCAS* virus. *RCAS::miR-196* viral particles at a titre of 10^{10} particles per ml were collected from the medium of transfected chicken embryonic fibroblasts. Proper transcription and processing of mature *miR-196-1* was confirmed by northern blots of total RNA extracted from chicken embryonic fibroblasts (data not shown).

Chicken embryo manipulations and *in situ* hybridization. Fertilized eggs were obtained from SPAFAS and incubated at 37°C, and the embryos were staged according to ref. 27. Eggs were incubated up to stage 7–8 and then the whole embryo was targeted by multiple injections of *RCAS::miR-196*. Alternatively, at stage 12–13 the coelomic cavity was targeted to infect the lateral plate mesoderm. Resin beads were soaked in 100 nM all-trans RA in dimethylsulphoxide for 1 h and then implanted into the anterior of stage-22 chick forelimbs for a further 4 h, as described¹⁰, except that AG-1X8 beads (Bio-Rad) were used. Alternatively, RA-soaked AG-1X2 beads ($1 \text{ mg ml}^{-1} = 300 \text{ mM}$) were implanted into stage-20 limbs that were allowed to develop *in ovo* for 36 h more⁴. Embryos were then collected and fixed in 4% paraformaldehyde overnight. Whole-mount *in situ* hybridization and probes have been described^{4,28}. The *Hoxa7* probe was

amplified directly by PCR from chicken genomic DNA and transcribed, without subcloning, by using the following primers: 5'-ACCTACCCCGCTACCAGAC-3' and 5'-TGTAATACGACTCACTATAGGGCCCTCTTCTCATCTTCTTCCA-3'.

Quantitative real-time PCR for chick *Shh*. Three untreated sample tubes, each containing four stage-23 forelimbs, and three corresponding sample tubes with *miR-196*-infected limbs were subject to quantification of *Shh* mRNA. Three replicate runs were done on each sample tube with a Lightcycler 2000 (Roche) using SYBER Green DNA Master Mix (Roche) and the following primers: GAPDH-5', 5'-CGGAGTCAACGGATT-3'; GAPDH-3', 5'-ATAACACGCTTA GCACC-3'; *Shh*-5', 5'-TGCTAGGGATCGGTGGATAG-3'; *Shh*-3', 5'-ACAA GTCAGCCCAGAGGAGA-3'. A 'no RT' control was done in parallel (data not shown). One-tailed *t*-test determined the significance of the difference in the mean value between the *miR-196*-infected sample and the untreated control.

Received 24 June; accepted 10 August 2005.

- Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
- Ambros, V. The functions of animal microRNAs. *Nature* **431**, 350–355 (2004).
- Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15–20 (2005).
- Riddle, R. D., Johnson, R. L., Laufer, E. & Tabin, C. Sonic hedgehog mediates the polarizing activity of the ZPA. *Cell* **75**, 1401–1416 (1993).
- Stratford, T., Horton, C. & Maden, M. Retinoic acid is required for the initiation of outgrowth in the chick limb bud. *Curr. Biol.* **6**, 1124–1133 (1996).
- Helms, J. A., Kim, C. H., Eichele, G. & Thaller, C. Retinoic acid signalling is required during early chick limb development. *Development* **122**, 1385–1394 (1996).
- Lu, H. C., Revelli, J. P., Goering, L., Thaller, C. & Eichele, G. Retinoid signalling is required for the establishment of a ZPA and for the expression of *Hoxb-8*, a mediator of ZPA formation. *Development* **124**, 1643–1651 (1997).
- Stratford, T., Logan, C., Zile, M. & Maden, M. Abnormal anteroposterior and dorsoventral patterning of the limb bud in the absence of retinoids. *Mech. Dev.* **81**, 115–125 (1999).
- Charite, J., de Graaff, W., Shen, S. & Deschamps, J. Ectopic expression of *Hoxb-8* causes duplication of the ZPA in the forelimb and homeotic transformation of axial structures. *Cell* **78**, 589–601 (1994).
- Stratford, T. H., Kostakopoulou, K. & Maden, M. *Hoxb-8* has a role in establishing early anterior-posterior polarity in chick forelimb but not hindlimb. *Development* **124**, 4225–4234 (1997).
- He, L. & Hannon, G. J. MicroRNAs: small RNAs with a big role in gene regulation. *Nature Rev. Genet.* **5**, 522–531 (2004).
- Meister, G. & Tuschl, T. Mechanisms of gene silencing by double-stranded RNA. *Nature* **431**, 343–349 (2004).
- Harfe, B. D., McManus, M. T., Mansfield, J. H., Hornstein, E. & Tabin, C. J. The RNaseIII enzyme Dicer is required for morphogenesis but not patterning of the vertebrate limb. *Proc. Natl Acad. Sci. USA* **102**, 10898–10903 (2005).
- Logan, M. et al. Expression of Cre Recombinase in the developing mouse limb bud driven by a *Pxrl* enhancer. *Genesis* **33**, 77–80 (2002).
- Baskerville, S. & Bartel, D. P. Microarray profiling of microRNAs reveals frequent coexpression with neighbouring microRNAs and host genes. *RNA* **11**, 241–247 (2005) 13.
- Mansfield, J. H. et al. MicroRNA-responsive 'sensor' transgenes uncover Hox-like and other developmentally regulated patterns of vertebrate microRNA expression. *Nature Genet.* **36**, 1079–1083 (2004).
- Yekta, S., Shih, I. H. & Bartel, D. P. MicroRNA-directed cleavage of *HOXB8* mRNA. *Science* **304**, 594–596 (2004).
- Llave, C., Xie, Z., Kasschau, K. D. & Carrington, J. C. Cleavage of Scarecrow-like mRNA targets directed by a class of *Arabidopsis* miRNA. *Science* **297**, 2053–2056 (2002).
- van den Akker, E. et al. Axial skeletal patterning in mice lacking all paralogous group 8 Hox genes. *Development* **128**, 1911–1921 (2001).
- Kim, M. H. et al. Retinoic acid response element in *HOXA-7* regulatory region affects the rate, not the formation of anterior boundary expression. *Int. J. Dev. Biol.* **46**, 325–328 (2002).
- Min, W. et al. 307-bp fragment in *HOXA7* upstream sequence is sufficient for anterior boundary formation. *DNA Cell Biol.* **17**, 293–299 (1998).
- Lewis, B. P., Shih, I. H., Jones-Rhoades, M. W., Bartel, D. P. & Burge, C. B. Prediction of mammalian microRNA targets. *Cell* **115**, 787–798 (2003).
- Lim, L. P. et al. Microarray analysis shows that some microRNAs downregulate large numbers of target mRNAs. *Nature* **433**, 769–773 (2005).
- Chen, C. Z., Li, L., Lodish, H. F. & Bartel, D. P. MicroRNAs modulate hematopoietic lineage differentiation. *Science* **303**, 83–86 (2004).
- Giraldez, A. J. et al. MicroRNAs regulate brain morphogenesis in zebrafish. *Science* **308**, 833–838 (2005).
- Zhao, Y., Samal, E. & Srivastava, D. Serum response factor regulates a muscle-specific microRNA that targets *Hand2* during cardiogenesis. *Nature* **436**, 214–220 (2005).
- Hamburger, V. & Hamilton, H. L. A series of normal stages in the development of the chick embryo. *J. Morphol.* **88**, 49–82 (1951).
- Nelson, C. E. et al. Analysis of *Hox* gene expression in the chick limb bud. *Development* **122**, 1449–1466 (1996).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank E. McGlenn and S. Nissim for critically reading the manuscript. This work was funded by grants from the NIH (to C.J.T. and to D.P.B.). E.H. was supported by a 'Dorot' fellowship and 'Bikura' award. J.H.M. is supported by a Kirchstein postdoctoral fellowship.

Author Information Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to C.J.T. (tabin@genetics.med.harvard.edu).

© 2005  nature publishing group

To order reprints, please contact:

Americas: Tel 212 726 9278; Fax 212 679 0843; author-reprints@nature.com

Europe/UK/ROW: Tel +44 (0)20 7833 4000; Fax +44 (0)20 7843 4500; author-reprints@nature.com

Japan & Korea: Tel +81 3 3267 8751; Fax +81 3 3267 8746; reprints@naturejpn.com