MicroRNA Cloning and Bioinformatic Analysis

by

Earl G. Weinstein

B.A. University of Pennsylvania, 1997

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

DOCTOR OF PHILOSOPHY
AT THE
MASSACHUSETTS INSTITUTE OF TECHNOLOGY

JUNE 2002

Signature of Author:

Earl G. Weinstein
Department of Biology
(May 11, 2002)

Certified by:

David P. Bartel
Associate Professor of Biology
Thesis Supervisor

Accepted by:

Professor of Biology
Chair, Biology Graduate Committee

MicroRNA Cloning and Bioinformatic Analysis

by

Earl G. Weinstein

Submitted to the Department of Biology
on May 24, 2002 in Partial Fulfillment of the
Requirements for the Degree of Doctor of Philosophy

ABSTRACT

Part I
Two gene-regulatory noncoding RNAs (ncRNAs), let-7 RNA and lin-4 RNA, were previously discovered in the *C. elegans* genome. The *let-7* gene is conserved across a wide range of genomes, suggesting that these ncRNAs represent a wider class of gene-regulatory RNAs.

Both lin-4 and let-7 RNAs are generated from stem-loop precursor RNAs, and share a common biochemical signature, namely 5'-terminal phosphate and 3'-terminal hydroxyl groups. We refer to ncRNAs that share the characteristic size, biochemical signature, and precursor structures of let-7 and lin-4 as microRNAs (miRNAs). The size of this class of genes, and its prevalence in other genomes, are unknown. Therefore, we developed an experimental and bioinformatics strategy to identify novel miRNA genes.

We discovered a total of 75 miRNA genes in the *C. elegans* genome, and orthologues for a majority of these were computationally identified in the *C. briggsae, D. melanogaster* or *H. sapiens* genomes. Northern analysis was used to confirm and analyze the expression of these miRNAs. The data set has implications for understanding miRNA gene regulation, miRNA processing, and regulation of miRNA genes.

Part II
Directed molecular evolution has previously been applied to generate RNAs with novel structures and functions. This method works because nucleic acids can be selected, randomized, amplified and characterized using polymerase chain reaction (PCR)-based methods.

Here we present a novel method for extending directed molecular evolution to the realm of peptide selections by linking a peptide to its encoding mRNA. The method makes use of a photoactivatable bifunctional tRNA with an amide-linked amino acid. This tRNA forms a stable amide bond to a peptide, as well as a UV-induced covalent linkage to an mRNA during translation, thereby tagging a peptide with its encoding mRNA. A proof of principle selection for two different peptides indicates that this tRNA should prove useful in discovering more complex protein molecules using directed molecular evolution.

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

Acknowledgements

I thank my advisor, David Bartel, who has been a continuing source of inspiration and support throughout graduate school. One quickly realizes that the door to David's office is nearly always open, and I have spent many hours there discussing both science and more general matters. This is reflective of the unique combination of intellectual prowess and general kindness that David displays, and it is this that has made graduate school a rewarding experience.

The members of my committee, Professors Tom RajBhandary, Chris Burge and Barbara Imperiali at MIT, and Gary Ruvkun at Harvard, have provided very useful general advice as well as specific suggestions deriving from their respective fields of expertise. Many useful improvements were made to my dissertation as a result of their comments, and I would like to thank them for their time and kind advice.

I am fortunate to have a great group of friends who have provided much needed balance over the course of the past five years. I would in particular like to thank David Nadler, Sasha Opotowsky, Jason Tanz, Chanan Tigay and Noah Krasner, amongst others, for their valued friendship. I owe much of this to them.

Finally, and most importantly, I thank the members of my family, my sister Liora, and my Mom and Dad. This dissertation is the result of their unflagging support, love and kindness. I dedicate it to them.

Table of Contents

Part I: MicroRNA Cloning and Bioinformatic Analysis

Introduction

Noncoding RNAs and Biological Complexity

A common assumption in the pre-genomics era was that biological complexity would correlate with the total number of genes in a genome. However, estimates placed on the number of coding regions in recently completed genome sequences lead to a seeming paradox. In particular, phenotypic complexity does not appear to scale very well with the underlying protein-coding potential. Perhaps the most striking example of this phenomenon arises in comparing the human genome to the lower-order *Drosophila melanogaster* and *Caenorhabditis elegans* genomes. The human genome is predicted to have on the order of 30,000 genes, which is only approximately twice the number predicted for the *D. melanogaster* and *C. elegans* genomes[1-3]. Calculations for the number of genes in the human genome range up to 80,000, but even this generous estimate cannot account for what is apparently greater than a threefold increase in biological complexity in going from these two lower order species to *Homo sapiens*[4]. Furthermore, this low correlation between sheer coding potential and observed phenotypic complexity is observed down through the evolutionary tree[5]. Biological complexity need not entirely correlate with the total number of protein-coding genes in a genome if evolution replaces low complexity genes with high complexity ones and this is part of the standard view for evolutionary change at the level of the genome. However comparative genomics argues against this being the sole mechanism underlying the increasing complexity amongst higher order organisms. For instance studies using the available mouse and human genome data indicate 99% conservation between the respective genomic coding regions[6]. Again, it is difficult to imagine how the remaining 1% variation in protein-coding sequences could account for the disproportionate increase in phenotypic complexity.

The above considerations suggest that there are undetected features of the genome beyond sheer gene number that underlie much of biological complexity. Combinatorial variation arising from alternative splicing[7-9], as well as transcription factor association[10], are important examples of such mechanisms for increasing biological complexity. Here I present evidence for an extensive class of gene-regulatory noncoding RNAs, termed microRNAs (miRNAs) that likely comprises an additional part of this "dark matter" of the genome. I also delineate an integrated bioinformatics and experimental approach to rapidly defining this set of noncoding

RNAs in a given genome. Because the class of miRNAs presented here is both extensive and is thought to possess gene-regulatory functions, this work is of interest from the standpoint of network-oriented models for biological complexity that account for the asymmetry between proteome size and phenotypic complexity[11-13]. This is discussed in greater detail later.

Noncoding RNA (ncRNA) is here used to refer to any transcript that is derived from regions between open reading frames in the genome, and which is therefore not in turn translated into a functional protein molecule. ncRNAs as a class are highly underrepresented in the output from both computational and experimental screens for transcribed genes. Experimentally, expressed sequence tag (EST) databases are often enriched for coding RNAs by selecting for polyadenylated sequences, thereby missing non-adenylated ncRNA sequences[2,14]. Forward genetic approaches are also unlikely to uncover many noncoding RNAs (there are some notable exceptions as discussed below) due to their relatively small size and relatively high tolerance to base changes that would cause phenotypically-observable frameshift and nonsense mutations in coding sequences. Computationally, under representation of ncRNAs is due to the difficulty of predicting these sequences in the absence of additional information. Thus, whereas coding genes can be predicted using a combination of signal sequences, homology modeling and statistical models for codon biases, ncRNAs are far more difficult to predict[15,16]. This is primarily due to the difficulty of reliably predicting promoter elements in the absence of other sequence signals that are used for exon prediction[17]. Certain ncRNAs such as tRNAs can be reliably predicted using a combination of secondary structure information and sequence covariation across related genomes. However, secondary structure alone cannot generally be used as a reliable parameter for distinguishing transcribed ncRNAs from random non-transcribed portions of the genome[18,19]. Computationally and experimentally-derived data sets for the transcriptosome are therefore likely to under represent the number and significance of ncRNAs, and novel experimental and computational screens are needed to detect ncRNAs throughout the genome.

Interestingly, comparative genomics is already beginning to provide evidence for the prevalence of ncRNAs. A striking first fact to note, is that 98% of the total transcription output of the human genome consists of ncRNAs including introns, transfer RNAs, ribosomal RNAs and other ncRNAs[20]. Related to this is the apparent correlation between the fraction of the genome that does not code for proteins and degree of phenotypic complexity, in that 15%, 30%,

70% and 95% of prokaryotic, *S. cerevisiae*, *C. elegans* and *H. sapiens* genomes respectively consist of DNA that does not code for proteins[21]. Alignments of extended intergenic regions from the human and mouse genomes indicate that these non-protein-coding regions do indeed have functional significance. For instance, Shabalina et al. tested 100 candidate full-length intergenic pairs from the available data in the human and mouse genomes by grouping intergenic pairs using the synteny of the corresponding flanking exons[21]. They noted that the alignments consisted of extended segments of greater than 50% homology interspersed with regions that align no better than random sequence pairs. Conservation indicates that there is negative selection against evolutionary drift within these segments of noncoding sequences, suggesting a possible functional role. Interestingly, this degree of conservation amongst intergenic pairs is threefold higher than that found in comparing the *C. elegans* and *C. briggsae* genomes[21]. This further suggests that selection against sequence drift has increased in evolution, indicating an increased functional role for these noncoding sequences. It should be noted that many of these conserved sequence regions comprise regulatory elements that are not transcribed into functional ncRNAs. However, as discussed below, a number of conserved regions are transcribed into ncRNAs, and although these studies do not define specific functional roles for these ncRNAs, they do indicate that such sequences are potentially abundant and could be uncovered given the right screening methodology.

Defined ncRNA sequences with regulatory roles in prokaryotes have been known for sometime. The DsrA ncRNA for example, has been shown to regulate translation of the *rpoS* gene by competing away intramolecular secondary structure of the *rpoS* mRNA, thereby enhancing translation[22]. These ncRNAs are trans-encoded and possess sequence information that allows them to specifically recognize target genes. Screens using prokaryotic genome data, have begin to uncover additional ncRNAs. For example, Argaman et al. used known conserved prokaryotic transcription regulatory elements to predict fourteen novel prokaryotic ncRNAs. Homology comparisons and northern analysis confirmed the presence of ncRNA transcripts ranging in size from 70 to 250 bases[23]. Similar screens have been performed by first looking for sets of conserved intergenic regions, and then calculating the probability that any compensatory mutations are more consistent with conservation of secondary structure rather than coding

potential. In one study, this approach yielded a non-redundant set of approximately 20 confirmed ncRNAs with unknown functions[24,25].

In eukaryotes, there are also known ncRNAs with defined functions. The set of small nucleolar RNAs (snoRNAs) involved in processing and modifying ribosomal RNA, and perhaps transfer RNA, is known to be quite extensive, perhaps numbering in the hundreds[26]. ncRNAs also have important functional roles as components of the splicesosome[27], the signal recognition particle[28,29], and telomerase activity[30]. A number of groups have recently conducted screens for additional novel ncRNAs using newly available genome data. Olivas et al. designed a genomic screen for ncRNAs in *S. cerevisiae*, the first eukaryotic genome to be fully sequenced[31]. They tested 10 candidate regions near strong consensus polIII promoter sites, and 59 intergenic regions of greater than 2 kb, for expression using northern analysis. Two of the sixteen RNAs that were shown to have expression had no known function or any obvious coding potential. Huttenhofer et al. used a more direct approach to detecting ncRNAs that allowed them to obtain ncRNA sequence data[32]. CTP tailing was used to generate a cDNA library from size-selected mouse RNA in the range of 50-500 bases. This type of library is unbiased towards coding RNAs, as there is no initial selection for polyA signals. This approach yielded a large set of novel snoRNAs, together with another set of expressed ncRNAs of unknown function.

These types of screens, together with the set of known, well-characterized ncRNAs mentioned above, indicate that ncRNAs represent an extensive and potentially functionally diverse class of genes. Indeed, the discovery of *lin-4* and *let-7*, two genes that encode ncRNAs involved in a specific regulatory network in *C. elegans*, represents one of the most interesting examples of the use of ncRNAs as networking molecules in genetic circuits[33-35]. In particular, these two ncRNAs are the first example of small (~21 nucleotide) sequences that can specifically regulate target genes. The subsequent discovery that the *let-7* gene is highly conserved in the genomes of possibly all bilaterally symmetric animals, led many to speculate that these two ncRNAs might be just a small part of a larger, more general class[36]. I turn now to a description of *let-7* and *lin-4* in the context of this specific regulatory network as it provided the "jumping off" point for much of the work presented here.

Lin-4 and Let-7: Two ncRNAs Involved in a Gene Regulatory Network

Developmental pathways involve a tightly regulated series of events at both the temporal and spatial levels, and much of this regulation involves coordinated gene expression and repression. Because specific genes are either turned on or off depending on their location and point in time, if ncRNAs do have a broad role in modulating gene regulatory networks, this role should become apparent in studying developmental pathways. Indeed, two of the most well studied ncRNAs were initially discovered using traditional forward genetics in the *C. elegans* heterochronic developmental pathway .

Development in *C. elegans* involves four post-embryonic stages labeled L1-L4 in which precursor undifferentiated blast cells divide into groups of cells with specific lineages[37-40]. Each of these lineages then expresses specific types of cell fates, and this cell fate expression is temporally regulated to coincide with specific embryonic stages. Genes that are involved in this type of temporal, as opposed to spatial, organization, are referred to as heterochronic genes, and a number of them have been genetically identified[38].

Examples of protein-coding heterochronic genes include *lin-14, lin-28, lin-41, lin-42* and *lin-29*. Gain of function (gf) and loss of function (lf) mutations of these heterochronic genes all have specific phenotypes that effect temporal transitions between the L1-L4 stages (Figure 1A). Specifically, stage-specific events are either precociously expressed at an earlier stage than normal, or executed and then reiterated in later stages than normal. Thus, *lin-14 (lf)* causes precocious expression of cell fates normally reserved for the L2 stage[41]. Conversely, *lin-14 (gf)* leads to reiteration of L1 cell fates during later than normal stages. In this example, therefore, *lin-14* seems to act as a negative switch that controls progression from the L1 stage to later developmental stages. Indeed LIN-14, a known nuclear protein, is observed to decline throughout L1, suggesting that LIN-14 acts as a transcription factor to negatively regulate genes involved in specifying L2-specific cell fates[27,42]. The *lin-14* gene must somehow in turn be specifically downregulated at the end of L1 to allow normal progression to the later L2 stage. Applying similar reasoning, *lin-28* appears to act as a negative switch governing transitions from the L2 stage to the L3 stage. Thus, *lin-28 (lf)* show precocious expression of L3-specific cell fates, whereas *lin-28 (gf)* execute and reiterate L2-specific cell fates at later stages. Again, *lin-28* must somehow be downregulated to allow progression from the L2 to L3 stages.

lin-14 and *lin-28* act as switches, therefore, to turn on and off genes responsible for executing specific cell fates during the ordered progression from the L1 to the L3 early developmental stages. These switches must somehow in turn be temporally regulated. A similar picture emerges for the temporal regulation of the transition from later developmental stages to the final adult stage. Thus, *lin-29 (lf)* reiterate later developmental stages and fail to proceed to the adult stage[43]. Furthermore, *lin-29* is epistatic to all known heterochronic genes with precocious phenotypes, indicating that it acts downstream of these genes and is the final determinant in the switch to the adult stage[39]. Regulation of *lin-29* must therefore be tightly controlled to prevent precocious expression of adult cell fates. An additional heterochronic gene, *lin-41*, plays a role in temporal regulation of *lin-29*. In *lin-41 (lf)*, normal L1-L3 development is observed followed by precocious adult terminal differentiation. Conversely, *lin-41* overexpression inhibits the transition to the adult stage. Furthermore, the LIN-41 protein appears to decline just as LIN-29, a zinc-finger transcription factor begins to appear. This implies that *lin-41* acts as a negative regulator of *lin-29* to ensure that terminal differentiation takes place after the L1-L4 stages are properly executed[39]. As with *lin-14* and *lin-28*, *lin-41* downregulation must be temporally regulated to ensure that *lin-29* is not prematurely activated.

The first indication that ncRNAs play a role in this gene regulatory network came from work on the *lin-4* heterochronic gene. *lin-4* has the opposite phenotype of the *lin-14* gene that negatively regulates the L1/L2 transition. Thus, *lin-4 (0)* have the same phenotype as *lin-14 (gf)*, namely reiteration of L1 cell fates at later stages[33]. Conversely, overexpression of *lin-4* leads to precocious expression of L2 cell fates as in *lin-14 (lf)*[27]. This, combined with the observation that *lin-4* expression coincides with the beginning of a rapid decline in LIN-14 implies that *lin-4* is also involved in controlling the L1/L2 transition but in an opposite manner to the *lin-14* gene. In particular, *lin-4* has the hallmarks of a negative regulator of *lin-14*. Thus, *lin-4* expression appears to promote transition to the L2 stage by relieving the LIN-14 block on progression to the L2 stage.

Definitive evidence for a link between *lin-4* and *lin-14* came when Lee et al. discovered that *lin-4* encodes a small twenty-two-nucleotide ncRNA[34]. The *lin-4* ncRNA sequence is complementary to sites located within the *lin-14* mRNA. These sites are specifically located in the 3' untranslated region (3' UTR) of the *lin-14* mRNA, and correspond to sequence regions

mutated in *lin-14 (lf)* in which transition to the L2 stage is precocious and unregulated[44]. This suggests that the *lin-4* complementary sites within the 3' UTR have functional significance as target sites for binding of *lin-4*. Further implicating the 3' UTR as the target for the *lin-4* ncRNA, fusion of an unrelated reporter gene to the 3' UTR of *lin-14* leads to a pattern of temporal expression of the unrelated gene that matches the temporally regulated expression of *lin-14*[34]. This regulation is *lin-4* dependent and is not observed when the fusion construct is introduced into a *lin-4*(-) background. Finally, *lin-4* complementary sites are conserved in the related *C. briggsae* genome[34,45] Taken together, this indicates that *lin-4* encodes a novel ncRNA that negatively regulates *lin-14* at a post-transcriptional level by binding to complementary sites within the 3' UTR of *lin-14* mRNA. Binding of the *lin-4* ncRNA turns off the *lin-14* repressional switch, thereby initiating the transition to the L2 stage. Regulation of *lin-28* by *lin-4* was later shown to involve a similar mechanism[46].

The subsequent discovery of *let-7,* a second ncRNA that regulates later stages of the developmental pathway, suggested that *lin-4* and *let-7* might be members of a larger class of ncRNAs[35]. *let-7 (lf)* reiterate earlier cell fates during the adult stage, whereas *let-7* overexpression leads to precocious expression of cell fates normally limited to the adult stage. This suggests a role for *let-7* in the regulation of the L4/adult transition. Like *lin-4, let-7* encodes a 21-nucleotide ncRNA. However, *let-7* complementary sites are found in the 3' UTRs of a different set of heterochronic genes, namely *lin-14, lin-28*, *lin-41*, and *daf-12*, although the functional significance of the *lin-14* and *lin-28* sites is unknown since these genes are involved in developmental regulation prior to *let-7* expression. The function of these sites is therefore best understood in the context of *lin-41.* Precocious expression of *lin-41* can suppress *let-7 (0)* suggesting that *let-7* negatively regulates *lin-41.* This is consistent with the fact that *lin-29* expression in L4 is dependent upon proper *let-7* expression and *lin-29* is in turn negatively regulated by *lin-41.* Thus, *let-7* seems to act as a positive regulator of *lin-29* by negatively regulating *lin-41* through binding to complementary sites within the *lin-41* 3' UTR. Indeed, *let-7* is itself temporally regulated with highest expression in L4 just as *lin-29* expression begins.

Beyond the fact that both *lin-4* and *let-7* regulate their respective genes at a post-transcriptional level by binding to the 3' UTR of their target mRNAs, little is known about the precise biochemical mechanisms underlying this regulation. The total number of stRNA

complementary sites may be important as *lin-14* has seven potential *lin-4* target sites in its 3' UTR[34,45] while *lin-28* has just one[46]. Interestingly, Wightman et al. showed that in the presence of *lin-4*, levels of *lin-14* mRNA remain constant while levels of the LIN-14 protein go down, suggesting a role for *lin-4* in regulating *lin-14* expression at the translational level[45] (Figure 1B). Furthermore, even in the presence of *lin-4*, *lin-14* mRNA remains associated with polyribosomes, suggesting that *lin-4* represses expression at the level of translation at a point after the initiation step[42].

Subsequent to the discovery of *let-7* in *C. elegans*, *let-7* orthologues with perfect conservation were identified in a wide range of bilaterally symmetric species[36]. Expression of these orthologues was also confirmed, and *let-7* complementary sites were identified in two *lin-41* orthologues. A mammalian *lin-4* orthologue in the mouse genome has also recently been identified[47]. The wide-ranging homology of *let-7*, and possibly *lin-4* as well, suggests that these ncRNAs, termed small temporal RNAs (stRNAs) because of their temporally regulated expression profiles, might be part of a broader and widely utilized class of gene-regulatory ncRNAs.

stRNAs and RNA Interference

One particularly suggestive feature of the conserved *let-7* orthologues pointed to an intriguing connection with the RNA interference pathway, and provided a number of very useful observations for the work presented here. In particular, in the genomes of *C. elegans, D. melanogaster,* and *H. sapiens*, there is a distinctive predicted stem-loop secondary structure that forms when *let-7* is folded with surrounding genomic sequence[36]. In all of the orthologues, the 21 nucleotide *let-7* sequence is located within the precursor structure near the loop portion of the stem-loop. The *let-7* sequence is also base-paired to the opposite side of the foldback with a limited number of mismatches and bulges, and in all the orthologous structures the stRNA sequence is located on the 5' side of the stem-loop. These stem-loop foldback structures appear then to be precursor structures that contain structural and perhaps sequence information that leads to the small ncRNA being processed out of the precursor into its active 21-nucleotide form.

This hypothesized precursor processing mechanism bears a striking resemblance to the initial steps of the RNA interference (RNAi) pathway, in which small amounts of double

stranded RNA (dsRNA) introduced into a cell can specifically knock down genes bearing sequence regions homologous to that of the exogenously introduced dsRNA[48]. In the RNAi pathway, small ncRNAs in the same 21-23 size range as *let-7* and *lin-4* are processed from precursor dsRNAs and are sufficient to knock down gene expression[49,50]. The size of the small ncRNAs involved in RNAi and their role in gene regulation bear a striking resemblance to the *lin-4* and *let-7* stRNAs. Indeed, genetic evidence indicates that both the RNAi and stRNA pathways share common components. In particular, mutations in *dcr-1,* the *C. elegans* homolog of the enzyme Dicer that is known to be important in processing the active small ncRNAs from precursor dsRNA in the RNAi pathway, have similar phenotypes to *let-7* and *lin-4* mutations[51,52]. Direct biochemical evidence that Dicer is involved in processing stRNAs from precursor foldbacks came from knocking down the Dicer protein in *D. melanogaster,* which leads to accumulation of the 72-nucleotide *let-7* precursor stem-loop structure[53] . Similarly, inactive *dcr-1* in *C. elegans* leads to accumulation of the *let-7* and *C. elegans*-specific *lin-4* stRNA precursors[51]. Conversely, in the presence of functional Dicer, *let-7* is processed into small active stRNAs containing a 5'-terminal monophosphate and a 3'-terminal hydroxyl group[53]. Both this biochemical signature, and the small RNA size, are identical to what is observed in Dicer processing of precursor dsRNA in the first step of the RNAi pathway during which Dicer processes out short 21-23 nucleotide dsRNA molecules containing two base pair overhangs at their 3' ends and 5'-terminal monophosphate and 3'-terminal hydroxyl groups[50]. A further connection between the two pathways is suggested by the fact that members of the related Argonaute and RDE-1 protein families seem to be necessary for *lin-4/let-7* processing and RNAi respectively[51].

RNAi and *let-7/lin-4* regulation of heterochronic genes are therefore both examples of post-transcriptional gene regulation sharing a common initial step involving Dicer processing from longer precursor RNAs. There are, however, a number of differences between the two types of ncRNAs, suggesting that they belong to two different classes of gene-regulatory ncRNAs. At the broadest level, RNAi and stRNAs seem to have evolved for two very different purposes. In particular, RNAi seems to function primarily in silencing exogenously introduced genes such as those from viruses and transgenes[54,55], and in repressing transposon mobilization[56,57], although there is one known instance of siRNA regulation of an endogenous

gene[58]. In contrast, both *let-7* and *lin-4* are involved in regulating expression of endogenous genes. Mechanistically, the short interfering RNAs (siRNAs) produced in RNAi, and the *let-7* and *lin-4* stRNAs have very different effects. In particular, siRNAs exist as short double-stranded ncRNAs that function as guide RNAs in an RNA-induced silencing complex (RISC). The RISC is guided to target genes containing homologous sequence to the guide siRNA and in turn degrades the target gene into 21-23 nucleotide fragments using a component nuclease activity[49,59]. In contrast, the two well-characterized stRNAs function as single-stranded molecules by binding to the 3' UTR of a target gene, and silencing expression without degrading the target mRNA. Structurally, siRNAs are processed from perfectly complementary dsRNAs and in turn require long uninterrupted stretches of perfect complementary in regions of their target genes to induce mRNA degradation[60]. stRNAs are processed from precursor stem-loop foldback structures containing mismatches and bulges in the stRNA region and throughout the stem-loop, and the short functional stRNAs in turn form base-paired structures containing mismatches and bulges, with sequences in the 3' UTR of target genes[51]. In fact, a bulged C residue is found in four out of the seven proposed *lin-4* target sites in the *lin-14* 3' UTR, and this mismatch is conserved in the 3' UTR target sites of the homologous *C. briggsae* gene[39,45]. A functional role for this mismatch is suggested by the fact that reporter genes fused to a 3' UTR containing the mismatch are repressed by *lin-4* in a similar manner to *lin-4* repression of *lin-14*, whereas this is not the case when the mismatched C residue is replaced by the matching nucleotide[61].

In summary, *let-7* and *lin-4* are both short noncoding RNAs that are processed from precursor foldback structures by the enzyme Dicer, and require members of the RDE-1/Argonaute family to function. The mature single-stranded short RNAs then regulate gene expression through a post-transcriptional mechanism involving binding to specific semi-complementary target sequences in the 3' untranslated region (UTR) of a target gene. This regulation is distinct from RNAi in that the mRNA levels remain constant even after *let-7* or *lin-4* binding. Instead, the protein levels decrease, and there is evidence that regulation occurs at the level of translation at a point after the initiation step.

It should be noted that UTR-mediated gene regulation is not without precedent. For example, both the *hunchback* and *caudal* mRNAs appear to be regulated post-transcriptionally

through elements in their 3' UTRs involved in binding Nanos and Pumilio proteins in the case of *hunchback* and Bicoid protein in the case of *caudal*[62]. Yet, *let-7* and *lin-4* are the first cases where small noncoding RNA sequences have been shown to mediate gene-specific regulation in eukaryotes, using a mechanism distinct from RNAi. Furthermore, the complete conservation of *let-7* across the genomes of bilaterally symmetric animals, together with the known Dicer homologues, suggests that the machinery underlying this gene-regulatory mechanism is widespread.

Summary of the Current Work

The size of this class of gene-regulatory ncRNAs, and the extent of their functional roles and mechanisms remains an open question. I am interested, therefore, in an integrated experimental and bioinformatics strategy for rapidly defining the total set of stRNA-like genes in a given genome. This set should include all the small ncRNAs in the genome that are processed by Dicer but are not siRNAs and that possess *let-7*-like precursor foldback structures. We refer to these ncRNAs as microRNAs (miRNAs), with the understanding that they are quite small but that not all of them will be temporally regulated like stRNAs. Briefly, the current strategy utilizes the characteristic 5'-terminal phosphate and 3'-terminal hydroxyl groups generated by Dicer in processing miRNAs and siRNAs, to directionally clone and sequence large sets of noncoding sequences containing candidate miRNAs. This is followed by the use of a suite of custom developed bioinformatics tools that effectively locates novel ncRNAs with the unique signatures of miRNAs within the sequencing data.

Using this strategy, we uncovered a total of 75 novel microRNA genes in the *C. elegans* genome. For the majority of these genes, we were able to locate orthologues in other genomes and found that one gene is, like *let-7*, perfectly conserved across multiple genomes. Clustering of related sequences within this set indicates that there are also groups of paralagous sequences with strong consensus motifs at their 5' ends. For instance, we found a set of four *let-7* paralogs and a group of seven related miRNAs that are proximally located within the genome. The large set of genes contains trends that have significance from the standpoint of understanding regulation of microRNA genes, microRNA target recognition, and microRNA processing from precursor foldback structures. Results from application of this strategy to the *C. elegans* genome

indicate that miRNAs comprise an extensive and evolutionarily-widespread class of gene-regulatory noncoding sequences.

References

1. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
2. Ewing, B. & Green, P. Analysis of expressed sequence tags indicates 35,000 human genes. *Nat Genet* **25**, 232-234 (2000).
3. Aparicio, S. A. How to count ... human genes. *Nat Genet* **25**, 129-130 (2000).
4. Wright, F. A. *et al.* A draft annotation and overview of the human genome. *Genome Biol* **2**, RESEARCH0025 (2001).
5. Rubin, G. M. *et al.* Comparative genomics of the eukaryotes. *Science* **287**, 2204-2215 (2000).
6. Stormo, G. D. Gene-finding approaches for eukaryotes. *Genome Res* **10**, 394-397 (2000).
7. Smith, C. W. & Valcarcel, J. Alternative pre-mRNA splicing: the logic of combinatorial control. *Trends Biochem Sci* **25**, 381-388 (2000).
8. Croft, L. *et al.* ISIS, the intron information system, reveals the high frequency of alternative splicing in the human genome. *Nat Genet* **24**, 340-341 (2000).
9. Mironov, A. A., Fickett, J. W. & Gelfand, M. S. Frequent alternative splicing of human genes. *Genome Res* **9**, 1288-1293 (1999).
10. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**, 238-241 (1996).
11. Hasty, J., McMillen, D., Isaacs, F. & Collins, J. J. Computational studies of gene regulatory networks: in numero molecular biology. *Nat Rev Genet* **2**, 268-279 (2001).
12. Duboule, D. & Wilkins, A. S. The evolution of 'bricolage'. *Trends Genet* **14**, 54-59 (1998).
13. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-52 (1999).
14. Marra, M. A., Hillier, L. & Waterston, R. H. Expressed sequence tags--ESTablishing bridges between genomes. *Trends Genet* **14**, 4-7 (1998).
15. Claverie, J. M. Computational methods for the identification of genes in vertebrate genomic sequences. *Hum Mol Genet* **6**, 1735-1744 (1997).
16. Burge, C. & Karlin, S. Prediction of complete gene structures in human genomic DNA. *J Mol Biol* **268**, 78-94 (1997).
17. Burge, C. B. & Karlin, S. Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8**, 346-354 (1998).
18. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**, 955-964 (1997).
19. Rivas, E. & Eddy, S. R. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* **16**, 583-605 (2000).
20. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep* **2**, 986-991 (2001).
21. Shabalina, S. A., Ogurtsov, A. Y., Kondrashov, V. A. & Kondrashov, A. S. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet* **17**, 373-376 (2001).
22. Wassarman, K. M., Zhang, A. & Storz, G. Small RNAs in Escherichia coli. *Trends Microbiol* **7**, 37-45 (1999).
23. Argaman, L. *et al.* Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Curr Biol* **11**, 941-950 (2001).

24. Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Curr Biol* **11**, 1369-1373 (2001).
25. Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev* **15**, 1637-1651 (2001).
26. Eliceiri, G. L. Small nucleolar RNAs. *Cell Mol Life Sci* **56**, 22-31 (1999).
27. Feinbaum, R. & Ambros, V. The timing of lin-4 RNA accumulation controls the timing of postembryonic developmental events in Caenorhabditis elegans. *Dev Biol* **210**, 87-95 (1999).
28. Lewin, R. Surprising discovery with a small RNA. *Science* **218**, 777-778 (1982).
29. Walter, P. & Blobel, G. Signal recognition particle contains a 7S RNA essential for protein translocation across the endoplasmic reticulum. *Nature* **299**, 691-698 (1982).
30. Prescott, J. & Blackburn, E. H. Telomerase RNA mutations in Saccharomyces cerevisiae alter telomerase action and reveal nonprocessivity in vivo and in vitro. *Genes Dev* **11**, 528-540 (1997).
31. Olivas, W. M., Muhlrad, D. & Parker, R. Analysis of the yeast genome: identification of new non-coding and small ORF-containing RNAs. *Nucleic Acids Res* **25**, 4619-4625 (1997).
32. Huttenhofer, A. *et al.* RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *Embo J* **20**, 2943-2953 (2001).
33. Chalfie, M., Horvitz, H. R. & Sulston, J. E. Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell* **24**, 59-69 (1981).
34. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).
35. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901-906 (2000).
36. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86-89 (2000).
37. Rougvie, A. E. Control of developmental timing in animals. *Nat Rev Genet* **2**, 690-701 (2001).
38. Ambros, V. Control of developmental timing in Caenorhabditis elegans. *Curr Opin Genet Dev* **10**, 428-433 (2000).
39. Slack, F. & Ruvkun, G. Temporal pattern formation by heterochronic genes. *Annu Rev Genet* **31**, 611-634 (1997).
40. Ruvkun, G. & Hobert, O. The taxonomy of developmental control in Caenorhabditis elegans. *Science* **282**, 2033-2041 (1998).
41. Ambros, V. & Horvitz, H. R. Heterochronic mutants of the nematode Caenorhabditis elegans. *Science* **226**, 409-416 (1984).
42. Olsen, P. H. & Ambros, V. 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 (1999).
43. Ambros, V. & Moss, E. G. Heterochronic genes and the temporal control of C. elegans development. *Trends Genet* **10**, 123-127 (1994).
44. Seggerson, K., Tang, L. & Moss, E. G. Two Genetic Circuits Repress the Caenorhabditis elegans Heterochronic Gene lin-28 after Translation Initiation. *Dev Biol* **243**, 215-225 (2002).
45. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855-862 (1993).

46. Moss, E. G., Lee, R. C. & Ambros, V. 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 (1997).
47. Lagos-Quintana, M. *et al.* Identification of tissue-specific microRNAs from mouse. *Curr Biol* **In Press.** (2002).
48. Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811 (1998).
49. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25-33 (2000).
50. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15**, 188-200 (2001).
51. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell* **106**, 23-34 (2001).
52. Knight, S. W. & Bass, B. L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* **293**, 2269-2271 (2001).
53. Hutvagner, G. *et al.* A cellular function for the RNA-interference enzyme Dicer in the maturation of the let-7 small temporal RNA. *Science* **293**, 834-838 (2001).
54. Matzke, M., Matzke, A. J. & Kooter, J. M. RNA: guiding gene silencing. *Science* **293**, 1080-1083 (2001).
55. Voinnet, O., Pinto, Y. M. & Baulcombe, D. C. Suppression of gene silencing: a general strategy used by diverse DNA and RNA viruses of plants. *Proc Natl Acad Sci U S A* **96**, 14147-14152 (1999).
56. Wu-Scharf, D., Jeong, B., Zhang, C. & Cerutti, H. Transgene and transposon silencing in Chlamydomonas reinhardtii by a DEAH-box RNA helicase. *Science* **290**, 1159-1162 (2000).
57. Jensen, S., Gassama, M. P. & Heidmann, T. Cosuppression of I transposon activity in Drosophila by I-containing sense and antisense transgenes. *Genetics* **153**, 1767-1774 (1999).
58. Aravin, A. A. *et al.* Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr Biol* **11**, 1017-1027 (2001).
59. Hammond, S. M., Bernstein, E., Beach, D. & Hannon, G. J. An RNA-directed nuclease mediates post-transcriptional gene silencing in Drosophila cells. *Nature* **404**, 293-296 (2000).
60. Parrish, S., Fleenor, J., Xu, S., Mello, C. & Fire, A. Functional anatomy of a dsRNA trigger: differential requirement for the two trigger strands in RNA interference. *Mol Cell* **6**, 1077-1087 (2000).
61. Ha, I., Wightman, B. & Ruvkun, G. A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes Dev* **10**, 3041-3050 (1996).
62. Bloom, T. Patterning the Drosophila embryo. *Curr Biol* **6**, 6-8 (1996).

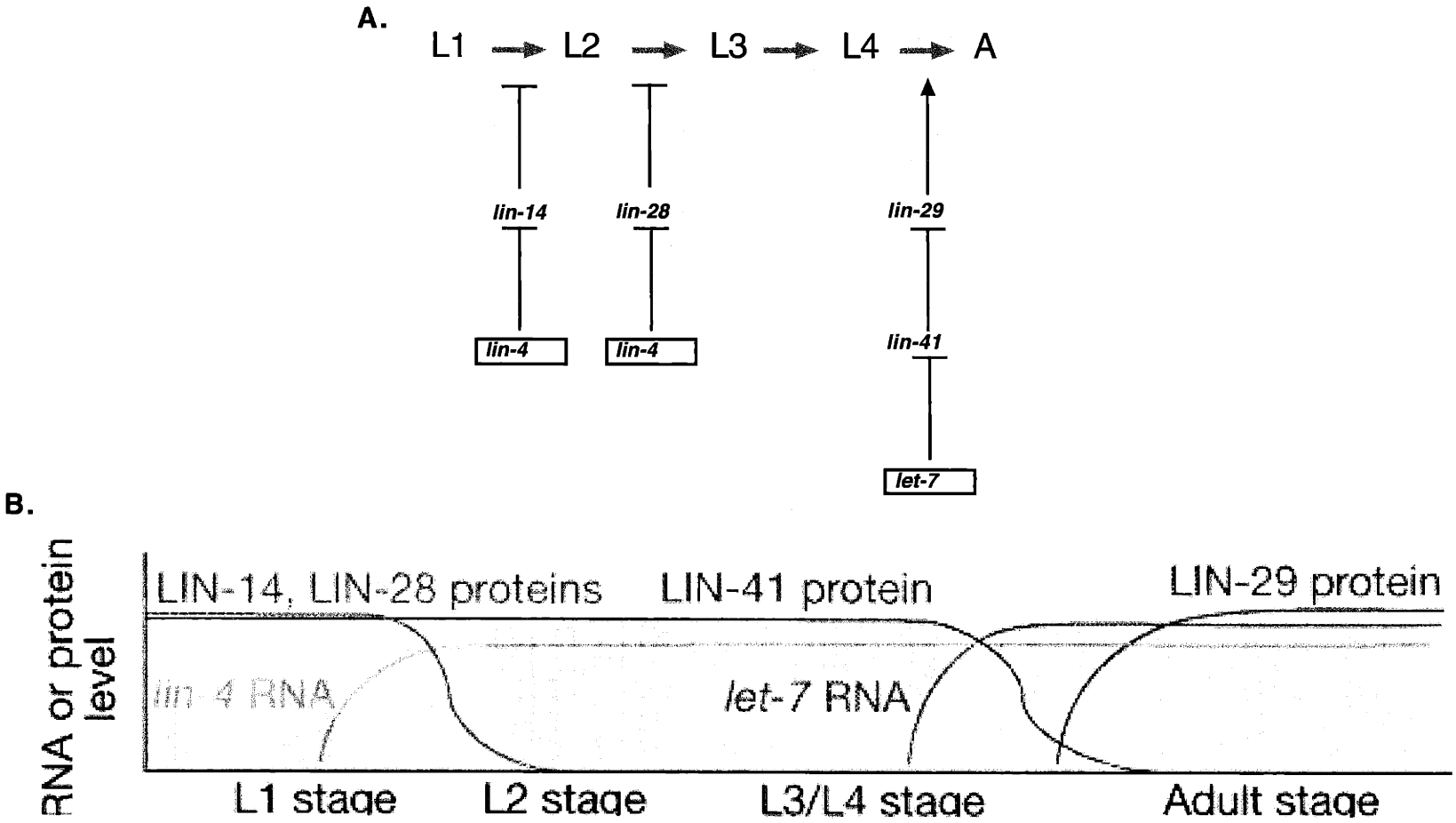


Figure 1. Two small ncRNAs, *lin-4* and *let-7*, regulate heterochronic gene expression.
(A) Heterochronic genes involved in post-embryonic development regulate transitions between larval stages L1-L4 and then to the adult stage. Two small ncRNAs, *lin-4* and *let-7*, downregulate heterochronc genes in this pathway as shown.

(B) Protein levels of genes regulated by *lin-4* and *let-7* decline as lin-4 and let-7 RNA levels rise, whereas mRNA levels remain constant (taken from Nature, 403:6772 pg. 901).

Large-Scale Sequencing of *C. elegans* MicroRNA Genes

The work presented in this chapter was a collaborative effort between myself and Nelson Lau. Specifically, I developed the software tools, cloned the microRNA sequences, probed the Northerns and undertook all of the downstream bioinformatics analysis. Nelson synthesized the adenylated adapter oligos needed for the directional cloning, prepared the *C. elegans* RNA from which the microRNA genes were cloned and prepared the Northerns.

ABSTRACT: A large-scale microRNA cloning and sequencing project using the *C. elegans* genome is presented. 18 novel microRNA genes were discovered, and expression for 17 of these was detected by Northern analysis. A majority of the sequences are constitutively expressed during development, and three of the sequences show developmental stage-specific expression. One of these three sequences is downregulated early in development contrary to the usual trend where a microRNA shows constant expression once turned on during larval development. The set of *C. elegans* microRNA sequences now includes four let-7 paralogs and a second perfectly conserved microRNA sequence. In eight cases, the small RNA sequence corresponding to the opposite side of the microRNA precursor foldback was cloned. Although some of the sequences are represented by hundreds of clones, eight of the miRNAs in the data set are still represented by a single clone suggesting that there are more microRNA genes to be discovered.

Introduction

Non-coding RNAs (ncRNAs) represent an important class of transcripts that are generally missed by computational screens and highly underrepresented in EST databases[1-3]. A subclass, known as microRNAs (miRNAs), consists of a large number of small (~21-24 nucleotide) RNAs that are thought to regulate expression of specific target genes at the post-transcription level. MicroRNAs were initially discovered using genetic screens that uncovered two examples of this type of ncRNAs, lin-4 and let-7, that act as heterochronic riboregulators in *C. elegans* development[4-6]. let-7 was subsequently shown to be highly conserved in the genomes of all bilaterally symmetric animals indicating that this class of ncRNA genes might be quite expansive both in terms of function and the total number of microRNA genes[7].

MicroRNAs share certain structural and functional features with small interfering RNAs (siRNAs), another type of gene-regulatory ncRNA. Both regulate gene expression at the post-transcriptional level using sequence-specific information and are processed into short (~21-24 nucleotide) active structures from longer double-stranded RNA precursors by the enzyme Dicer[8-10]. However, there are a number of significant differences. siRNAs direct nucleitic cleavage of target mRNAs, which is consistent with their role in transposon monitoring and defense against exogenously introduced genes[11]. let-7 and lin-4, on the other hand, have been shown to downregulate translation of their endogenous target genes by binding to complementary sites in the 3'-untranslated region, leaving the mRNA stable but suppressing translation, possibly at a step after initiation[6,12-14]. At the structural level, siRNAs are processed by Dicer from perfectly base-paired dsRNA, whereas the known miRNAs are processed by Dicer from endogenous precursor stem-loop structures that contain multiple loops and mismatches. This difference in precursor structure may represent an early branch point between the miRNA and siRNA pathways in that the small RNAs processed out of imperfectly matched precursor stem-loop structures in the miRNA pathway are not likely to be retained as stable dsRNAs as in the RNAi pathway.

Recently the miRNA class of genes was greatly expanded when our group and two others, used a combination of experimental and informatics approaches to clone close to 100 novel miRNA genes from the *C. elegans*, *D. melanogaster*, and *H. sapiens* genomes[15-17]. Our directional cloning procedure was designed to select for possible ncRNAs containing the

characteristic ~22 nt length, 5'-terminal monophosphate and 3'-terminal hydroxyl generated by Dicer when processing small RNAs from precursor dsRNA, and yielded 55 novel miRNA genes. Interestingly, however, only a few of the cloned sequences that matched the *C. elegans* genome were candidate siRNA genes, despite the fact that the procedure was designed to clone either siRNA or miRNAs, indicating that miRNAs might play a more dominant role than siRNAs in regulating normal endogenous genes. Indeed, to date there is only one known example of an endogenous gene that that is regulated using siRNAs[18]. Furthermore, many of the sequences in our initial screen were represented by only one clone, indicating that the 55 miRNAs identified do not span the total set of available miRNAs in the genome. The large set of novel genes cloned previously, combined with the high incidence of single frequency sequences, led us to suspect that there might still be a large set of novel non-coding miRNA genes to be found.

Here we present the results from a large-scale cloning and sequencing project aimed at a more comprehensive sampling of the set of available miRNA sequences in the *C. elegans* genome. We conducted a greater than tenfold scale up from our initial miRNA sequencing effort and designed informatics tools to organize the large numbers of sequences and rapidly select candidate miRNA genes. In an effort to uncover novel stage-specific miRNA genes, we also expanded the number of stages from which RNA was cloned. In addition to the mixed stage population used in our initial cloning effort, we cloned RNA sequences derived from dauer, male-enriched *him-8*, and starved L1 stages.

Methods

Cloning and Sequencing of Dicer Products

Dicer RNA products between 18 and 25 nucleotides were isolated, concatamerized, and cloned into TOPO TA vectors as previously described[16]. PCR was used to screen for clones containing inserts of greater than 120 base pairs and 768 clones were submitted for sequencing. This set consisted of 384 concatamers derived from RNA of the mixed-stage population used in our initial screen, 192 concatamers from the dauer stage, 96 concatamers from the starved L1 stage, and 96 concatamers from a high incidence of male (HIM-8) population. Sequencing was followed by the use of a set of informatics tools developed to rapidly extract likely candidate miRNA genes from the large set of total sequences. Small RNA sequences were computationally extracted from the surrounding concatemer linker sequences, yielding a total of 5,319 individual sequences. Duplicate sequences were removed from this set, and the remaining sequences were compared to the set of sequences from our-previous small-scale cloning effort, yielding a non-redundant set of novel small RNA sequences.

The novel sequences were searched against the complete *C. elegans* and *E. coli* genomes using the BLAST program[19]. Sequences that matched the *C. elegans* genome and did not match a fragment of a known *C. elegans* tRNA or rRNA, were archived to a separate file consisting of these new sequences and those identified in the previous cloning effort. This archive was then ordered into groups of sequences with 5' and 3' length heterogeneity and a representative sequence with the maximum cloning frequency and sequence length was chosen for each group. This step effectively filters the number of candidate sequences that must be manually examined for desired secondary structure elements. For each of these representative sequences, two predicted secondary structures were generated using the Zuker folding algorithm. In the first structure, the candidate miRNA sequence is placed towards the 5' end of the foldback by folding it together with 15 bases and 60 bases of the surrounding 5' and 3' genomic sequence respectively. In the second structure, the candidate miRNA sequence is placed towards the 3' end of the foldback by folding it together with 60 bases and 15 bases of the surrounding 5' and 3' genomic sequence respectively. The set of predicted secondary structures was then manually inspected for the stem-loop foldback structure that is characteristic of miRNA genes[20].

Results

We used a combination of directional cloning of Dicer products and informatics tools to discover miRNA sequences from input *C. elegans* RNA. 52% of the initial 5.319 small RNA sequences matched the *C. elegans* genome and were not fragments of tRNAs or rRNAs. Of these *C. elegans* matches, 87% were determined to be candidate microRNA genes based on their predicted secondary structures. In total, 22 candidate miRNA sequences that had not been previously identified were discovered, many of which were cloned multiple times, and with varying degrees of length heterogeneity. The initial set of 5,319 sequences also contained multiple copies of nearly all the miRNAs found or predicted during our initial sequencing effort[16]. A summary of this comprehensive set is shown in Table 1.

Expression Analysis

Northern analysis (Fig. 1) confirmed expression for 18 of the 22 candidate sequences, although one of these (L1_D10-4) is unlikely to be an miRNA, as discussed below. An additional candidate sequence (D1_F11-5) showed no Northern signal, however there is an orthologue in *C. briggsae* with ninety percent sequence identity, and so D1_F11-5 is likely a new miRNA as well. Summary data for these 18 novel miRNAs including D1_F11-5 is shown in Table 2. The majority of the 17 detected miRNA sequences are constitutively expressed throughout larval development and only three sequences, N5_F02-7, L1_G01-4, and H1_G07-2, show clear developmental stage-specific expression. N5_F02-7 is a new paralog of let-7 and has an identical pattern of stage-specific L3, L4, and adult expression as the other three sequences in the let-7 family (Fig. 2). L1_G01-4 has a particularly interesting pattern with noticeably elevated levels at the L1 stage followed by decreased levels for the remainder of development (Fig. 3A). This is in contrast to the usual pattern of expression in which an miRNA is expressed at near-constant levels once turned on during larval development. D2_H10-2, which was cloned out of dauer RNA also has markedly elevated levels of expression in the dauer stage, suggesting a role in regulating dauer formation (Fig. 3B). There is also a markedly low level of processing of D2_H10-2 from the ~70 nucleotide precursor in the L3 and L4 stages, as indicated by the high ratio of precursor to mature miRNA at these two stages (Fig. 3B). Length heterogeneity in the stably expressed miRNA is also observable for a handful of sequences such as D1_F02-2,

suggesting that some miRNA sequences are prone to undergo heterogeneous processing and that the heterogeneous sequences are stable.

miRNA Sequences

A number of previously noticed sequence trends were reexamined by combining the eighteen new miRNAs with the previous set to increase the statistical relevance of these trends (Table 1). miRNA sequence lengths continue to lie in a tight distribution centered between 21 and 24 nucleotides. Biases in the sequence composition were examined by combining the 17 confirmed novel miRNAs with the previous data set. As before, there is a nearly consensus bias towards a U in the first position of the sequence. In addition, there is a bias towards purine residues at position two. However, the previously observed bias against U at positions two to four is no longer evident, and there does not seem to be significant information content beyond the first two positions. The observed sequence biases and the tight length distribution are not likely to be an artifact of the cloning as the background *E. coli* sequences have both broad length and sequence-composition distributions[16].

Twice as many of the new miRNAs are located on the 3' side of the foldback compared to the 5' side, consistent with the trend observed in the earlier set. Interestingly, when the complete set of miRNA precursor foldbacks are sorted by their genomic location it is evident that eight of the precursor sequences contain miRNA sequences processed from both sides of the foldback (Fig. 4). In the initial screen, one pair of sequences, mir-56/mir-56*, was cloned from both sides of the precursor foldback. With this larger set of data, there is now an observable disparity in the cloning frequency when comparing two miRNAs derived from both sides of one foldback, suggesting that only one of the two miRNAs is stably expressed and functional. This is consistent with the fact that in most cases only one of the two miRNAs is observable on a Northern. L1_D10-4 is the one exception where both it and the more frequently cloned mir-44/45 are both observable on Northerns (Fig. 3C). In this case the precursor foldback appears to be present at higher levels than the L1_D10-4 small RNA product. Interestingly, in most of these instances of dual miRNA foldbacks, the miRNAs are positioned within the foldback at mutually complimentary positions except at their 3' ends where there is a two nucleotide overhang (Fig. 4). This 3' two base pair overhang is reminiscent of what is observed when small interfering RNAs (siRNAs) are processed from longer precursor dsRNAs by Dicer in the RNA

interference (RNAi) pathway, suggesting a similar mechanism for Dicer processing of siRNAs and miRNAs[21].

Clustering of the miRNA data set based on sequence similarity reveals that N5_F02-7 is a new let-7 paralog (Fig. 2A). N5_F02-7 is the third paralog of let-7 cloned thus far, and all four genes in this set share identical expression profiles (Fig. 2B). This set of paralogs contains a consensus TGAGGTAG sequence at the 5' end, consistent with an overall bias towards 5' conservation within groups of related miRNA sequences within our data set. Two members of this let-7 family, N5_F02-7 and mir-48, are located within 2 kb of each other on the same side of the genome.

Paralogs and Orthologues

We searched the closely related *C. briggsae* genome as well as the more distantly related *D. melanogaster* and *H. sapiens* genomes for orthologues of the miRNA sequences. 17 of the 18 miRNAs have orthologues with good foldbacks in *C. briggsae* and all of these orthologous miRNAs, except for N1_B06-5, have ninety percent or greater identity to their *C. elegans* homolog. N3_D05-3 had no identifiable orthologue, and may be located in the remaining 10% of the *C. briggsae* genome yet to be sequenced. Alternatively, N3_D05-3 may be a sequence that has only a few targets, which would allow for more rapid sequence drift. An additional three miRNAs N1_H09-5, N1_B06-5 and D1_F02-2 have orthologues of greater than eighty percent homology in *D. melanogaster.* N1_H09-5 has identical orthologous sequences in the *C. elegans*, *C. briggsae*, *D. melanogaster*, and *H. sapiens* genomes with well-structured precursor foldbacks (Fig. 5A). The N1_H09-5 sequence is constitutively expressed and no paralogs of this sequence have been cloned. This is the second example, after *let-7*, of perfect miRNA sequence conservation across these divergent genomes. For all the orthologues, conservation tends to be highest across the miRNA sequences and falls off rapidly for the surrounding precursor sequence as, for example, in the group of N1_H09-5 orthologues (Fig. 5B). This is consistent with selective pressure against sequence drift within the functional miRNA sequence itself, which must recognize specific target sequences. Drift within the surrounding foldback sequence is likely more readily tolerated so long as it preserves the overall secondary structure.

Discussion

We present a large-scale cloning and sequencing project aimed at an in-depth sampling of the set of available microRNA (miRNA) sequences in the *C. elegans* genome. MicroRNAs are a class of small (21-24 nucleotide) gene-regulatory noncoding RNAs that are found in a wide range of genomes and are processed by the enzyme Dicer from characteristic precursor stem-loop structures.

Our data set now totals 75 confirmed miRNAs including the 18 novel miRNAs reported here. One of these novel miRNAs, N1_H09-5, is the second example after let-7 of a perfectly conserved miRNA sequence in *C. elegans*, *C. briggsae*, *D. melanogaster* and *H. sapiens*. A number of trends in the data are quite compelling. There is a noticeable asymmetry in the frequencies with which miRNAs derive from either the 3' side compared to the 5' side of the precursor stem-loop structure. Greater than two thirds of the sequences we cloned are derived from the 3' side of the foldback. In addition, we now have eight examples of precursor foldbacks where small RNAs have been cloned from both sides of the foldback. This can be due to symmetric processing in which Dicer initially processes out a small dsRNA containing both the miRNA and the RNA on the opposite side of the foldback. Alternately, Dicer might asymmetrically process only the miRNA from the precursor using sequence and secondary structure information to locate the correct small RNA. Symmetric processing is more consistent with the fact that Dicer seems to prefer a U in the first position, and the majority (five out of eight) of the small RNA sequences derived from the opposite side of the foldback do not begin with a U. These sequences may therefore be processed by Dicer as part of a larger symmetric unit containing a miRNA that begins with a U. Symmetric processing is also mechanistically more consistent with the initial steps of the RNA interference (RNAi) pathway during which long dsRNA is symmetrically processed into short 21-23 nucleotide small interfering RNAs (siRNA). One hallmark of this process is the presence of two nucleotide 3' overhangs in the processed siRNAs[21]. Indeed in each of the eight pairs of RNAs processed from both sides of the foldback, the RNAs are positioned within the stem-loop such that their ends contain 3' overhangs. However, we cannot rule out the possibility of asymmetric precursor processing. Under this model Dicer would rely on information encoded within the precursor to locate the correct miRNA. Our low frequency cloning of the opposite side may in turn reflect occasional errors in this process.

Sequence conservation is strongest at the 5' end of the miRNAs, and there seem to be families of miRNAs defined by a consensus 5' end sequence. One of these families contains four let-7-like sequences. It has been suggested that miRNAs and other regulatory noncoding RNAs find their targets using an "area code" model in which a general set of target genes is specified using sets of consensus 5' end sequences, followed by more specific gene recognition using information at the 3' end of the sequence[22]. The set of 5' consensus sequences within our data set might recognize such area codes in the *C. elegans* genome.

The majority of our miRNA sequences are constitutively expressed during development, suggesting that miRNAs play broad regulatory roles in addition to regulation of heterochronic genes during development. One developmentally-regulated miRNA, L1_G01-4, is the second known exception, after miR-49, to the rule that miRNAs are continuously expressed once turned on during larval development. L1_G01-4 shows strong expression at the L1 stage of development with markedly lower levels observable at later stages of development. This may occur transcriptionally through repression of the L1_G01-4 gene or post-transcriptionally through targeted degradation of this miRNA. No accumulation of the L1_G01-4 precursor is observed on Northerns (Fig. 3A) making it unlikely that this miRNA is downregulated at the level of Dicer processing. L1_G01-4 also shows elevated levels of expression in the dauer and starved L1 stages, and so this miRNA may be involved in regulating a starvation pathway in response to environmental stimuli. Another miRNA, D2_H10-2, also shows markedly elevated levels of expression in the dauer stage suggestive of a regulatory role in the dauer response. It will be interesting to work out the details of how the levels of miRNAs such as these are specifically up or downregulated in response to various cues.

Our approximately tenfold scale up from our initial sequencing effort yielded one third the number of novel miRNAs. Our success rate for uncovering novel miRNA genes has therefore fallen by a factor of thirty, suggesting that we may be close to sequencing the majority of the miRNA genes in the *C. elegans* genome. This is consistent with the results of recent efforts aimed at computationally predicting the locations and sequences of miRNA genes in the *C. elegans* genome which places the total number of candidate miRNA genes at less than 120 (L. Lim, N. Lau, D. Bartel, C. Burge. In preparation). However, eight of our miRNA genes have still been cloned only once, and there may therefore be additional rare miRNA genes waiting to

be discovered. Nevertheless, the set of confirmed miRNA genes is now quite large suggesting an extensive role for these small noncoding RNAs in gene regulation.

References

1. Eddy, S. R. Noncoding RNA genes. *Curr Opin Genet Dev* **9**, 695-699 (1999).
2. Burge, C. B. & Karlin, S. Finding the genes in genomic DNA. *Curr Opin Struct Biol* **8**, 346-354 (1998).
3. Marra, M. A., Hillier, L. & Waterston, R. H. Expressed sequence tags--ESTablishing bridges between genomes. *Trends Genet* **14**, 4-7 (1998).
4. Chalfie, M., Horvitz, H. R. & Sulston, J. E. Mutations that lead to reiterations in the cell lineages of C. elegans. *Cell* **24**, 59-69 (1981).
5. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).
6. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901-906 (2000).
7. Pasquinelli, A. E. *et al.* Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408**, 86-89 (2000).
8. Grishok, A. *et al.* Genes and mechanisms related to RNA interference regulate expression of the small temporal RNAs that control C. elegans developmental timing. *Cell* **106**, 23-34 (2001).
9. Knight, S. W. & Bass, B. L. A role for the RNase III enzyme DCR-1 in RNA interference and germ line development in Caenorhabditis elegans. *Science* **293**, 2269-2271 (2001).
10. Zamore, P. D., Tuschl, T., Sharp, P. A. & Bartel, D. P. RNAi: double-stranded RNA directs the ATP-dependent cleavage of mRNA at 21 to 23 nucleotide intervals. *Cell* **101**, 25-33 (2000).
11. Bosher, J. M. & Labouesse, M. RNA interference: genetic wand and genetic watchdog. *Nat Cell Biol* **2**, E31-36 (2000).
12. Wightman, B., Burglin, T. R., Gatto, J., Arasu, P. & Ruvkun, G. Negative regulatory sequences in the lin-14 3'-untranslated region are necessary to generate a temporal switch during Caenorhabditis elegans development. *Genes Dev* **5**, 1813-1824 (1991).
13. Ha, I., Wightman, B. & Ruvkun, G. A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes Dev* **10**, 3041-3050 (1996).
14. Olsen, P. H. & Ambros, V. 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 (1999).
15. Lee, R. C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862-864 (2001).
16. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).
17. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-858 (2001).
18. Aravin, A. A. *et al.* Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Curr Biol* **11**, 1017-1027 (2001).
19. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).

20. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911-940 (1999).
21. Elbashir, S. M., Lendeckel, W. & Tuschl, T. RNA interference is mediated by 21- and 22-nucleotide RNAs. *Genes Dev* **15**, 188-200 (2001).
22. Lai, E. C. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* (2002).

Table 1. 75 miRNAs cloned from *C. elegans*. Of these sequences, 57 were previously found, and nearly all were cloned again in the current effort. The 18 newly discovered miRNAs are D1_A01-1 to N5_F02-7 and are placed at the top of the table. Total cloning frequencies, and stage-specific cloning frequencies for the mixed stage N2 (N), dauer (D), *him-8* (H) and starved L1 (L) stages are as indicated.

miRNA gene	Id	miRNA sequence	Number of clones	N	D	H	L
-	D1_A01-1	AAUGGCACUG CAUGAAUUCA CGG	16	0	8	1	7
-	D1_E02-3	UCUUUGGUUG UACAAAGUGG UAUG	5	0	1	0	4
-	D1_F02-2	UUGAGCAAUG CGCAUGUGCG G	9	1	5	0	3
-	D1_F11-24	UAAAUGCAUC UUAACUGCGG UGA	10	4	4	1	1
-	D1_F11-5	UCACAGGACU UUUGAGCGUU GC	1	0	1	0	0
-	D2_B02-2	UAAGCUCGUG AUCAACAGGC AGAA	3	1	2	0	0
-	D2_F03-4	UUUGUACUCC GAUGCCAUUC AGA	2	0	2	0	0
-	D2_H10-2	UACACGUGCA CGGAUAACGC UCA	1	0	1	0	0
-	H1_G07-2	UACUGGCCCC CAAAUCUUCG CU	3	2	0	1	0
-	H1_H05-5	GUAUUAGUUG UGCGACCAGG AGA	1	0	0	1	0
-	L1_D07-2	UUGCGUAGGC CUUUGCUUCG A	1	0	0	0	1
-	L1_G01-4	UUAUUGCUCG AGAAUACCCU U	1	0	0	0	1
-	N1_B06-5	UAUUGCACUC UCCCCGGCCU GA	22	4	12	6	0
-	N1_H09-5	UAAGGCACGC GGUGAAUGCC A	26	9	11	2	4
-	N2_B02-14	UAAUACUGUC AGGUAAUGAC GCU	6	2	3	1	0
-	N2_H06-10	UUUGUACUAC ACAUAGGUAC UGG	5	4	0	1	0
-	N3_D05-3	CGGUACGAUC GCGGCGGGAU AUC	1	1	0	0	0
-	N5_F02-7	UGAGGUAGGU GCGAGAAAUG A	8	7	0	1	0
let-7	N0_181-1	UGAGGUAGUA GGUUGUAUAG UU	15	15	0	0	0
lin-4	N0_112-10	UCCCUGAGAC CUCAAGUGUG A	88	53	18	15	2
mir-1	N0_42-5	UGGAAUGUAA AGAAGUAUGU A	70	45	11	7	7
mir-2	N0_11-10	UAUCACAGCC AGCUUUGAUG UGC	173	120	32	7	14
mir-34	N0_102-7	AGGCAGUGUG GUUAGCUGGU UG	39	13	19	5	2
mir-35	N0_93-4	UCACCGGGUG GAAACUAGCA GU	22	21	0	0	1
mir-36	N0_104-8	UCACCGGGUG AAAAUUCGCA UG	28	24	0	4	0
mir-37	N0_112-2	UCACCGGGUG AACACUUGCA GU	8	8	0	0	0
mir-38	N0_99-7	UCACCGGGAG AAAAACUGGA GU	10	9	0	0	1
mir-39	N1_H05-4	UCACCGGGUG UAAAUCAGCU UG	14	11	0	3	0
mir-40	N0_77-6	UCACCGGGUG UACAUCAGCU AA	18	14	0	1	3
mir-41	N0_103-1	UCACCGGGUG AAAAAUCACC UA	3	3	0	0	0
mir-42	N5_F12-9	UCACCGGGUU AACAUCUACA GA	17	11	3	1	2
mir-43	N0_155-8	UAUCACAGUU UACUUGCUGU CGC	12	8	0	0	4
mir-44/45	N0_134-3	UGACUAGAGA CACAUUCAGC U	28	21	3	3	1
mir-46	N0_151-2	UGUCAUGGAG UCGCUCUCUU CA	28	11	8	3	6
mir-47	N0_103-10	UGUCAUGGAG GCGCUCUCUU CA	29	19	5	3	2
mir-48	N0_77-8	UGAGGUAGGC UCAGUAGAUG CGA	50	45	0	5	0
mir-49	N0_104-2	AAGCACCACG AGAAGCUGCA GA	2	1	0	0	1

mir-50	D1_E04-5	UGAUAUGUCU	GGUAUUCUUG	GG	26	10	12	0	4
mir-51	N0_2-6	UACCCGUAGC	UCCUAUCCAU	GUU	25	16	5	2	2
mir-52	N0_2-3	CACCCGUACA	UAUGUUUCCG	UGCU	311	233	52	15	11
mir-53	N0_31b-5	CACCCGUACA	UUUGUUUCCG	UGCU	26	18	4	3	1
mir-54	N0_77-7	UACCCGUAAU	CUUCAUAAUC	CGAG	88	51	24	8	5
mir-55	N0_55-7	UACCCGUAUA	AGUUUCUGCU	GAG	88	45	19	11	13
mir-56	N0_2-17	UACCCGUAAU	GUUUCCGCUG	AG	68	41	13	6	8
mir-57	N0_2-16	UACCCUGUAG	AUCGAGCUGU	GUGU	45	31	5	3	6
mir-58	N0_11-2	UGAGAUCGUU	CAGUACGGCA	AU	206	145	33	13	15
mir-59	N0_117-2	UCGAAUCGUU	UAUCAGGAUG	AUG	1	1	0	0	0
mir-60	N0_157-2	UAUUAUGCAC	AUUUUCUAGU	UCA	28	19	3	6	0
mir-61	N0_110-10	UGACUAGAAC	CGUUACUCAU	C	15	10	2	2	1
mir-62	N0_63-7	UGAUAUGUAA	UCUAGCUUAC	AG	10	5	2	0	3
mir-63	N0_77-3	UAUGACACUG	AAGCGAGUUG	GAAA	8	7	0	1	0
mir-64	N0_130-7	UAUGACACUG	AAGCGUUACC	GAA	22	12	4	1	5
mir-65	N0_187-4	UAUGACACUG	AAGCGUAACC	GAA	28	22	3	1	2
mir-66	N0_63-3	CAUGACACUG	AUUAGGGAUG	UGA	94	71	16	4	3
mir-67	N0_151-10	UCACAACCUC	CUAGAAAGAG	UAGA	3	3	0	0	0
mir-68	N0_110-6	UCGAAGACUC	AAAAGUGUAG	A	1	1	0	0	0
mir-69	predicted	UCGAAAAUUA	AAAAAGUGUA	GA	0	0	0	0	0
mir-70	N0_112-11	UAAUACGUCG	UUGGUGUUUC	CAU	19	11	3	3	2
mir-71	D2_A05-11	UGAAAGACAU	GGGUAGUGA		122	54	43	18	7
mir-72	N0_113-3	AGGCAAGAUG	UUGGCAUAGC	UGA	64	45	11	4	4
mir-73	N0_113-6	UGGCAAGAUG	UAGGCAGUUC	AGU	19	12	6	0	1
mir-74	N0_81-2	UGGCAAGAAA	UGGCAGUCUA	CA	51	34	7	4	6
mir-75	N0_124-9	UUAAAGCUAC	CAACCGGCUU	CA	18	15	1	1	1
mir-76	N0_123-7	UUCGUUGUUG	AUGAAGCCUU	GA	8	1	1	1	5
mir-77	N5_D09-2	UUCAUCAGGC	CAUAGCUGUC	CA	18	16	2	0	0
mir-78	N0_14-6	UGGAGGCCUG	GUUGUUUGUG	C	5	4	1	0	0
mir-79	N0_155-6	AUAAAGCUAG	GUUACCAAAG	CU	19	14	2	2	1
mir-80	N0_90-5	UGAGAUCAUU	AGUUGAAAGC	CGA	142	104	21	9	8
mir-81	N0_81-6	UGAGAUCAUC	GUGAAAGCUA	GU	51	31	13	5	2
mir-82	N0_4-6	UGAGAUCAUC	GUGAAAGCCA	GU	52	33	7	9	3
mir-83	N5_F06-4	UAGCACCAUA	UAAAUUCAGU	AA	32	17	9	6	0
mir-84	N0_178-1	UGAGGUAGUA	UGUAAUAUUG	UAGA	14	12	1	1	0
mir-85	N0_66-6	UACAAAGUAU	UUGAAAAGUC	GUGC	20	10	0	10	0
mir-86	N0_7-3	UAAGUGAAUG	CUUUGCCACA	GUC	112	45	39	7	21
mir-90	N2_D05-3	UGAUAUGUUG	UUUGAAUGCC	CCU	41	6	19	6	10

Table 2. 18 novel miRNAs cloned from *C. elegans*. Many sequences have length heterogeneity at the 3' terminus. The observed length range for each cloned sequence is shown. Orthologues were found for most of miRNAs using the available *C. briggsae* sequencing traces, and the orthologue identity to the *C. elegans* miRNA is noted as +++ (100%), ++ (90%), and + (>75%).

miRNA gene	miRNA sequence	Length	*C. briggsae* homology	Fold-back arm		Chromosome & distance to nearest gene
D1_E02-3	UCUUUGGUUG UACAAAGUGG UAUG	23-25	+++	5'	I	1.6 kb from end of T04D1.2, antisense
N1_B06-5	UAUUGCACUC UCCCCGGCCU GA	22-22	+	3'	I	0.7 kb from start of T09B4.7
N2_B02-14	UAAUACUGUC AGGUAAUGAC GCU	21-24	+++	3'	II	0.3 kb from end of C52E12.1, antisense
L1_G01-4	UUAUUGCUCG AGAAUACCCU U	21-21	+++	3'	II	1.6 kb from end of Y54G11B.1, antisense
D2_B02-2	UAAGCUCGUG AUCAACAGGC AGAA	23-24	++	3'	III	10.4 kb from start of C07H6.7
D2_F03-4	UUUGUACUCC GAUGCCAUUC AGA	23-23	++	3'	III	6.7 kb from end of F44E2.2, antisense
L1_D07-2	UUGCGUAGGC CUUUGCUUCG A	21-21	++	5'	IV	0.9 kb from end of F36A4.14, antisense
N3_D05-3	CGGUACGAUC GCGGCGGGAU AUC	23-23	-	3'	IV	1.0 kb from start of R08C7.1
D1_A01-1	AAUGGCACUG CAUGAAUUCA CGG	23-24	+++	5'	IV	0.2 kb from end of T12E12.5, antisense
D1_F11-24	UAAAUGCAUC UUAACUGCGG UGA	23-23	+++	3'	IV	1.2 kb from end of F13H10.5, antisense
N1_H09-5	UAAGGCACGC GGUGAAUGCC A	21-21	+++	3'	IV	1.2 kb from end of C29E6.6
N5_F02-7	UGAGGUAGGU GCGAGAAAUG A	21-21	++	5'	V	7.1 kb from start of F56A12.1, antisense
D2_H10-2	UACACGUGCA CGGAUAACGC UCA	23-23	++	3'	X	in intron of AH9.3
D1_F11-5	UCACAGGACU UUUGAGCGUU GC	22-22	++	3'	X	2.7 kb from start of Y41G9A.6
H1_H05-5	GUAUUAGUUG UGCGACCAGG AGA	23-23	++	3'	X	0.4 kb from end of F13D11.3, antisense
H1_G07-2	UACUGGCCCC CAAAUCUUCG CU	22-22	++	3'	X	1.7 kb from start of C39D10.3
N2_H06-10	UUUGUACUAC ACAUAGGUAC UGG	22-23	++	5'	X	6.1 kb from start of C34E11.1
D1_F02-2	UUGAGCAAUG CGCAUGUGCG G	21-23	+++	3'	X	in intron of W03G11.4

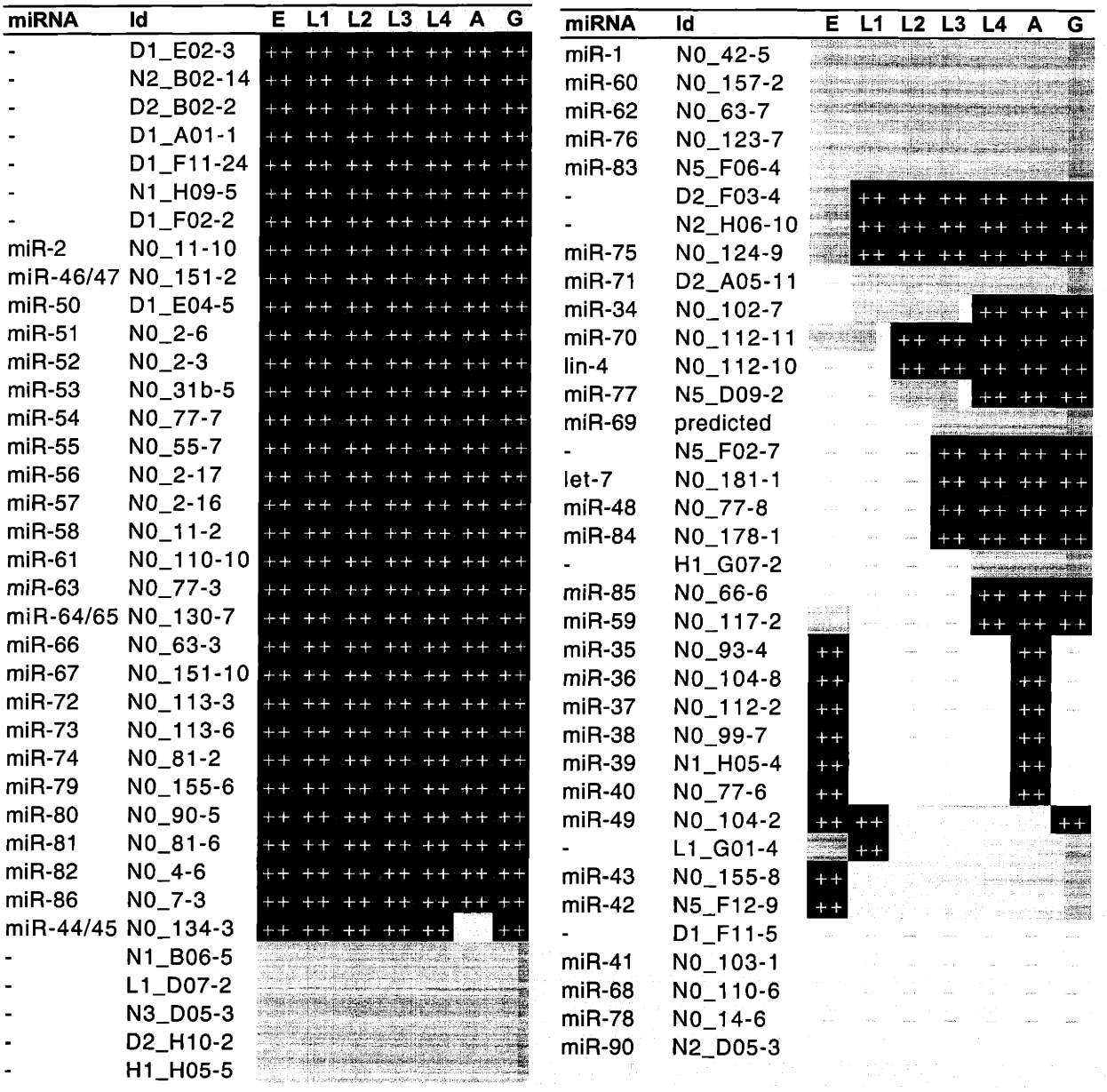


Fig. 1. Developmental expression of the miRNAs cloned from *C. elegans*. Expression was probed in the embryo (E), the L1-L4 larval stages, the adult stage (A), and *glp-4* mutant adults (G). High expression (++), low expression (+), and no detectable expression (-) are as indicated.

A.

C.e. let-7 RNA	UGAGGUAGuagguuguauaGuu-
C.b. let-7 RNA	UGAGGUAGuagguuguauaGuu-
C.e. miR-84	UGAGGUAGuauguaauauuGua-
C.b. miR-84	UGAGGUAGuuugcaaugcuGuc-
C.e. miR-48	UGAGGUAGgcucaguagauGcga
C.b. miR-48	UGAGGUAGgcucaguagauGcga
C.e. N5_F02-7	UGAGGUAGgugcgagaaauGa--
C.b. N5_F02-7	UGAGGUAGgugugagaaauGa--
	******** *

B.

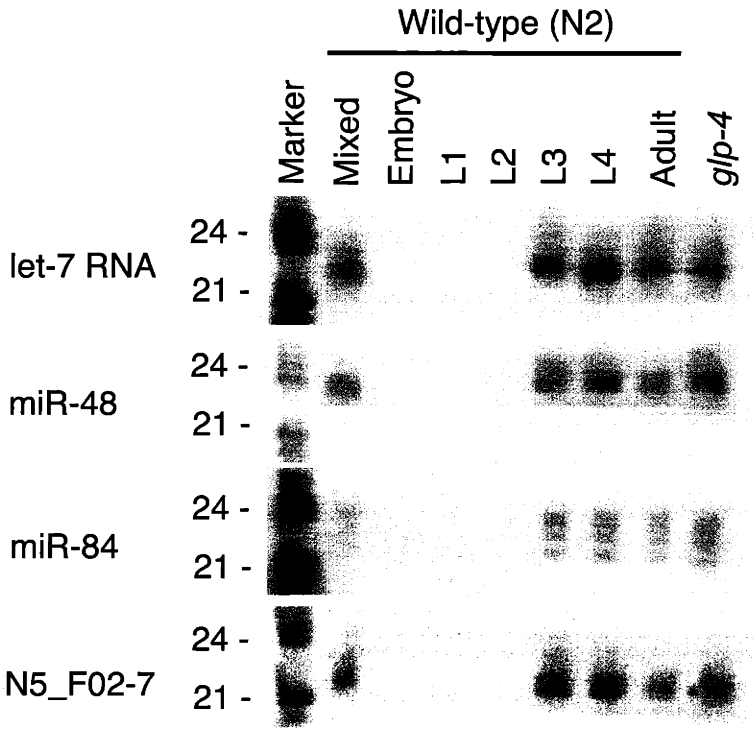


Fig. 2. The *let-7* family of miRNAs in *Caenorhabditis.*
(A) Sequence alignment of the four *C. elegans* miRNAs and their *C. briggsae* orthologues with conserved residues shown in upper case.
(B) Coordinate developmental expression of the *let-7* family *in C. elegans.* The developmental stages are as listed in Fig. 1.

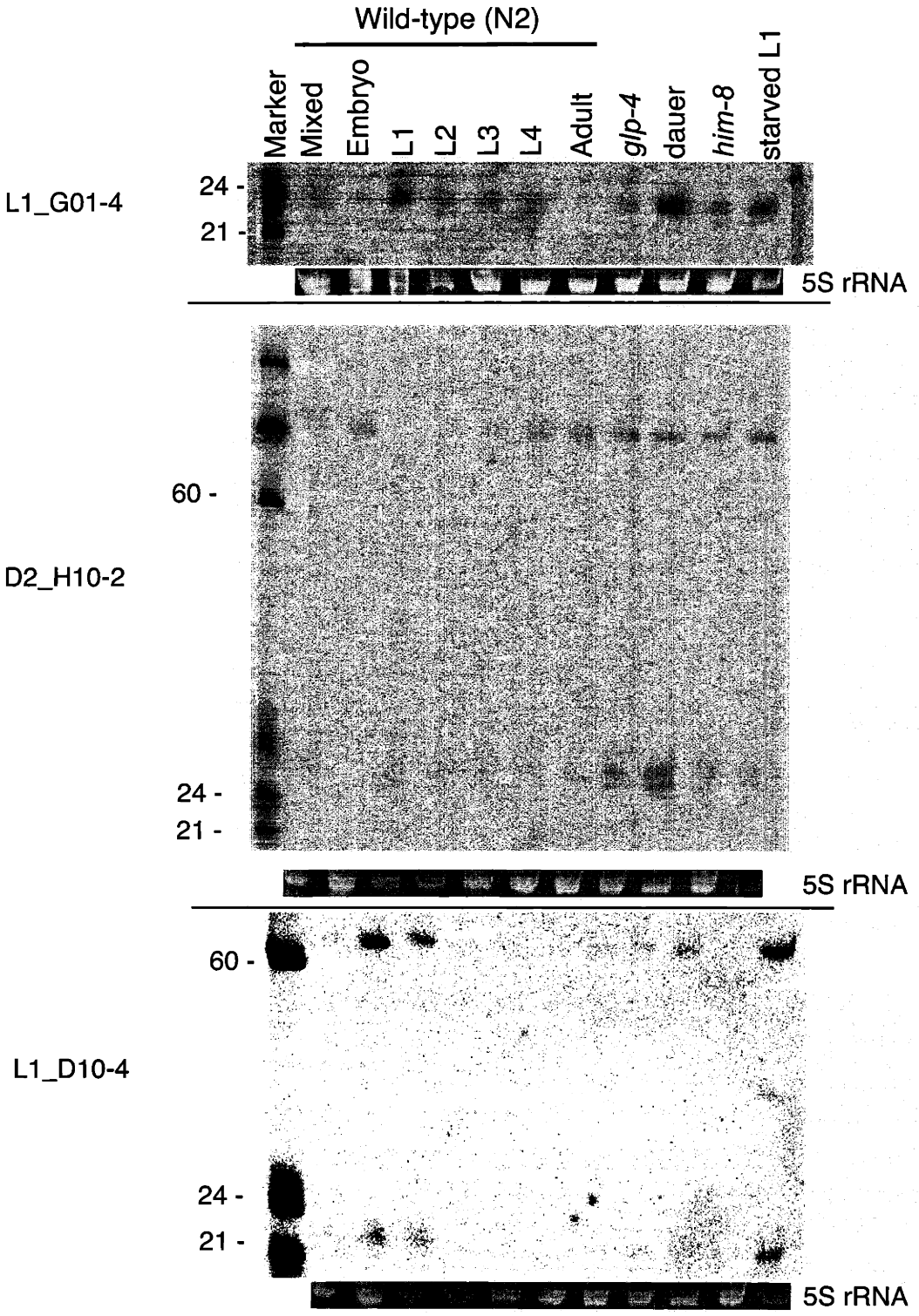


Fig. 3. Expression analysis of two miRNAs with unusual expression during development, and one small RNA (L1_D10-4) cloned from the opposite side of the miR-45 precursor.

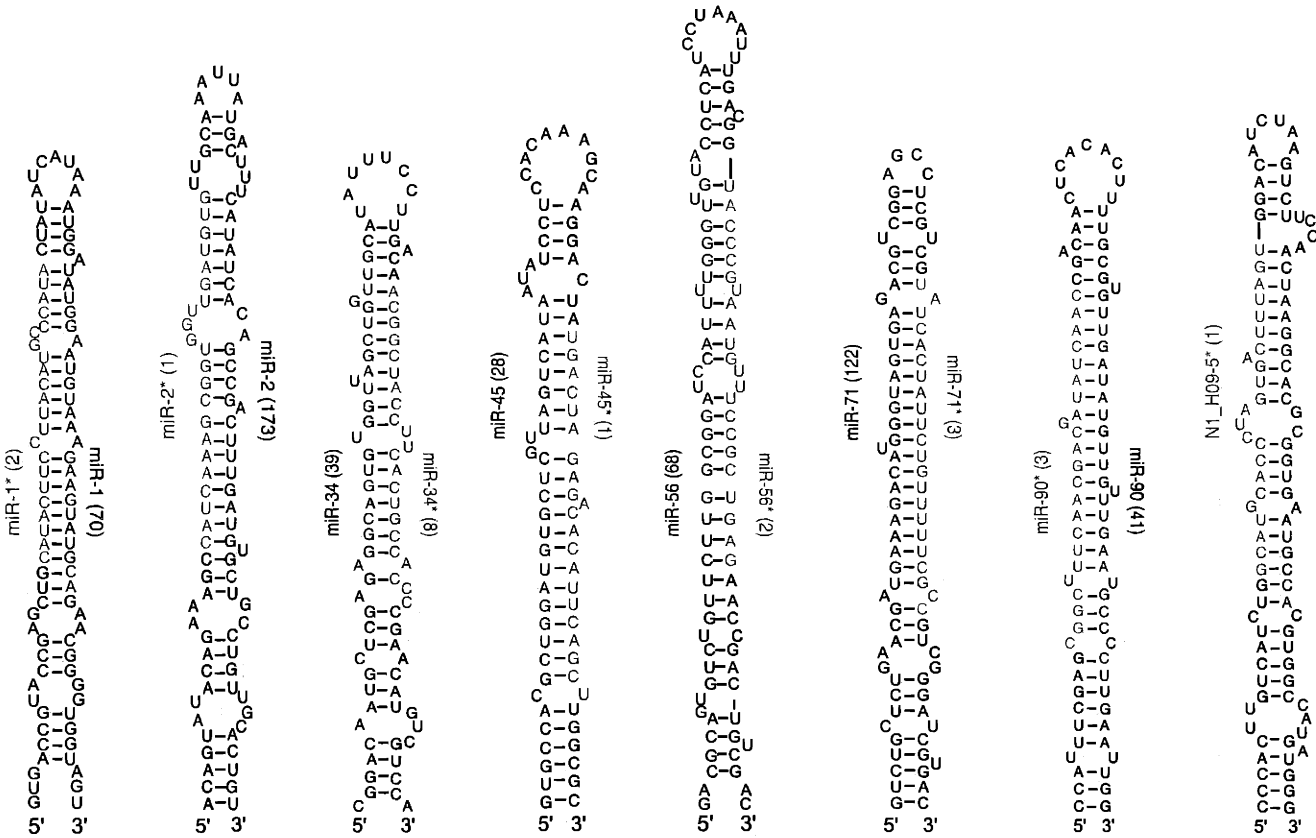
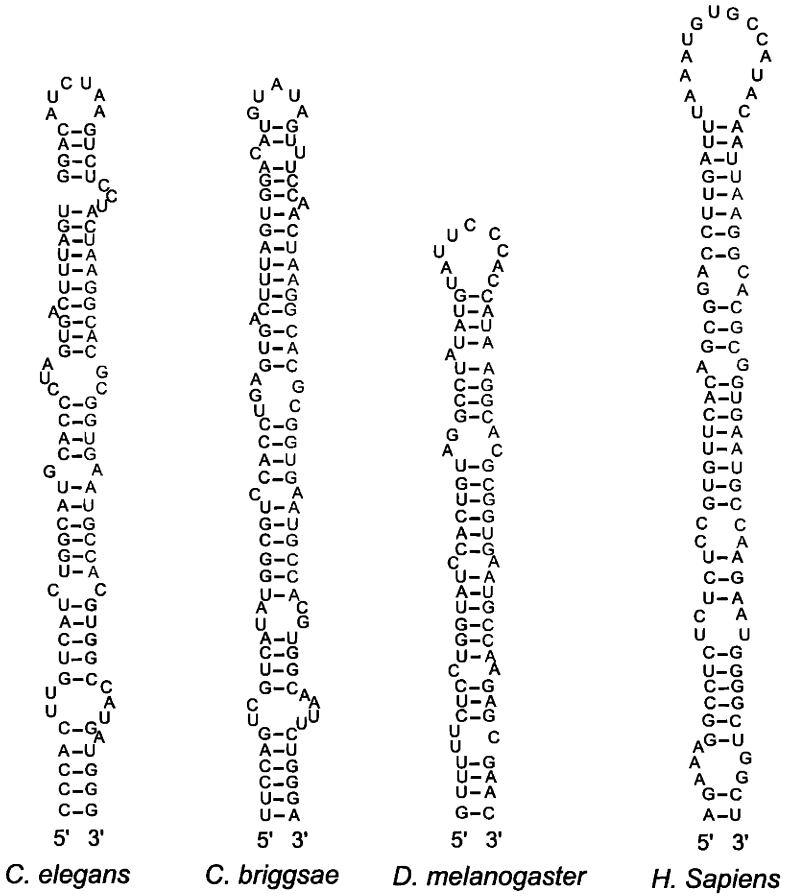


Fig. 4. RNAs cloned from both sides of precursor foldbacks. Predicted secondary structures for the eight foldbacks containing two cloned RNAs are shown. The more frequently cloned sequence from each foldback is shown in red, and the less frequently cloned sequence is shown in blue. Cloning frequencies for each of the RNAs are shown in parentheses.

A.



B.

```
drosophila      ---------------UUCGUCAUUUGGUACGUUUUUCUCCUG-GUAUCCACUGUAGGCCUAUAUGUAUUUCCAC-CA--UAAGGCACGCGGUGAAUGCCAAGAGCGAACGCAGUU-
human           ----------------UCAAGAUUAGAGGCUCUGCUCUCC---GUGUUCACAGCGGACCUUGAUUUAAUGUCAUACAAUUAAGGCACGCGGUGAAUGCCAAGAGCGGAGCCUACG-
elegans         --------------GUCCCACUUGUCAUCUGGCAUGCACCCUAGUGACUUUAGUGGACAUCUAAGUCUUCCAAC-----UAAGGCACGCGGUGAAUGCCACGUGGCCAUGAUGGG-
briggsae        UCAUUGAUACAUUUUUUCCAGUCGUCAUAUGGCGUCCACCUGAGUGACUUUAGUGGACAUGUAUAGUUUCCAAC-----UAAGGCACGCGGUGAAUGCCACGUGGCAAUUCUGGGA
                          *                  * **   **       *  * * *  *     *   *      ********************* * *   *
```

Fig. 5. N1_H09-5, a miRNA absolutely conserved in *C. elegans, C. briggsae, D. melanogaster* and *H. sapiens*.

(A) Predicted secondary structures for the precursor foldbacks, with the conserved miRNA sequence shown in red.

(B) Sequence alignments of the foldback sequences with the conserved miRNA shown in blue.

An Experimental and Bioinformatics Approach to MicroRNA Gene Cloning and Sequencing

The work presented in this chapter was a collaboration between myself, Nelson Lau and Lee Lim. On the experimental side, Nelson worked out the original experimental directional cloning protocol. On the bioinformatics side, Lee wrote the interface from our suite of programs to the BLAST software, and wrote a script to locate and label candidate miRNA sequences in the downstream postscript files.

Overview

Here I present an experimental and informatics approach to identifying large sets of novel miRNA genes in a given genome. This involves a directional cloning methodology followed by downstream informatics analysis of sequenced vectors to rapidly locate miRNA genes in large sets of input sequencing data.

Experimental Methodology

Experimentally, the goal is to clone and sequence the ncRNAs in the genome that contain the characteristic 5'-terminal phosphate and 3'-terminal hydroxyl groups generated by Dicer in processing miRNAs and siRNAs. Detailed experimental protocols are found elsewhere[1] and here I would like to describe the rationale behind our experimental cloning and sequencing methodology (Figure 1).

Two characteristics of miRNAs make them amenable to selective cloning from a heterogeneous population of starting RNAs. The small size of miRNAs allows them to be separated from longer RNAs using a size-selective electrophoresis step with total genomic RNA as the input, followed by excision from the gel of a band containing all RNAs between 18-26 nucleotides. This step yields a heterogeneous population of small RNAs including miRNAs and a wide range of RNA fragments derived from mRNAs, tRNAs, rRNAs, introns, etc. The 5'-terminal phosphate and 3'-terminal hydroxyl groups that are characteristic of Dicer processing serve as useful features for selective miRNA cloning from this heterogeneous population of small RNA sequences. To selectively clone products of Dicer processing, we use two different adapter oligonucleotides that are specific for 5'-terminal phosphate and 3'-terminal hydroxyl groups, and that contain primer-binding sites for selective PCR amplification of these products. The 3'-terminal hydroxyl located on the small RNAs, is first used to ligate on a 5'-adenylated adaptor oligonucleotide containing a protected 3' end using T4 RNA ligase. This reaction is done in the absence of ATP to avoid T4 RNA ligase circularization of the input miRNAs. The 5'-terminal phosphate is then used to ligate on a second adaptor oligonucleotide in the presence of ATP using T4 RNA ligase. Circularization is not a problem during the 5' adapter ligation due to the presence of a blocking group on the 3' end of the first adaptor sequence. Both the 5' and 3' adaptor oligos contain non-palindromic BanI restriction sites that are used in a later step.

Reverse transcription of the adapter-ligated RNA population followed by PCR amplification using primers specific to the adaptor sequences yields a cDNA library that is enriched in small RNAs that contain a 5'-terminal phosphate and a 3'-terminal hydroxyl. In order to efficiently sequence large numbers of these small cDNAs, multiple short sequences are concatamerized by cutting the cDNAs at the Ban I sites located in the adapter sequences, and ligating together multiple cDNAs with T4 DNA ligase. Because the BanI sites are non-palindromic, all of the small RNA sequences necessarily orient in the same direction within the concatemer, which is then cloned and sequenced.

Computational Methodology

The experimental cloning, concatemerization and sequencing generates a large set of sequence traces, each of which contains multiple RNA sequences. This data set must then be processed to locate candidate miRNAs. A suite of PERL scripts was developed for this purpose, and allows the user to rapidly go from input sequence traces to an ordered list of the best candidate miRNA sequences. All aspects of this suite of programs were designed to be fully modular so that new sequences and updated genomic data could be added at any time. The overall bioinformatics scheme is diagrammed in Figure 2 and described in more detail below.

Sequence Extraction

The programs take, as input, folders containing multiple sequenced cloning vectors each containing a sequenced concatemer of small cDNAs. For each sequenced cloning vector, the first step is to extract the concatemer of cDNA sequences from the surrounding vector sequence. Individual cDNA sequences within the concatemer are then extracted from the constant linker sequences that adjoin each of them in the concatemer. All of these steps are carried out by a program called linkerstrip.csh that first defines the edges of the concatemer using the known sequences of the 5' and 3' adapter oligos and removes the surrounding sequence from the cloning vector. Next, the sense or antisense orientation of the constant known linker sequences formed from ligating the 5' and 3' adaptors at the non-palindromic Ban I sites is used to define the orientation of the cDNAs within the concatemer. cDNA sequences are then extracted by splitting a string representation of the concatemer sequence along the constant linker sequences. The orientation information from the constant linker sequences is then used to correct the

orientation of the extracted small cDNAs, where a reverse complement function is applied to the cDNA sequences if the constant linker sequences were determined to be the antisense versions. It is crucial that the cDNA sequences be accurately extracted from the concatemer as they are essentially defined at this step for all subsequent steps of the informatics analysis. Care must be taken, therefore, to accurately define the sequence borders with the adjoining linker sequences. In order to do this, we empirically determined that a fifteen percent degree of mismatch to the known linker sequences should be allowed to balance the need for tolerating sequencing errors with the need for accurately defining the ends of the small RNA sequences.

Individual extracted sequences are then sequentially numbered based on their position in the concatemer. A size filter is also applied at this step to remove sequences that are smaller than five nucleotides as these would interfere with downstream steps. For sequencing, we utilize a 96-concatemer/folder format, where each concatemer comes from a unique position in a 96-well plate and is sequentially assigned a unique identifier of the form XY where X is a letter ranging from A-H and Y is a number ranging from 1-12. Each RNA extracted from a given concatemer is then uniquely identified within the 96-concatemer folder by combining the concatemer ID with the sequence location ID. For example, sequence C12-6 is the sixth sequence in concatemer C12. Each 96-concatemer folder is in turn assigned a two character code where the first character contains information about the source of the RNA, and the second character is a number used to sequentially number folders of concatamers from the particular RNA source. This id is then combined with the sequence concatemer id, yielding an RNA id that uniquely specifies every cloned RNA in the growing database. For example, L3_C12-6 is the sixth sequence in concatemer C12 obtained from the third folder of sequences cloned from the L stage. These unique identifiers are used to keep track of the small RNA sequences in all subsequent steps.

Database Updating

A program called getnewseqs.pl is then used to order the sequences extracted from the 96-concatemer folder and to update databases of all the sequenced RNAs. Duplicate sequences are first removed from each folder and the internally unique sequences for the input folder written to a separate file for that folder. This file of internally unique sequences is compared to an archive of all the unique sequences cloned up to that point and any sequences not already

represented in the uniques archive are written to a separate file of new sequences for that folder. This file of new sequences is in turn copied to the archive of unique sequences and formatted in FASTA format. Statistics for each folder such as the total number of sequences, percent of the sequences in the folder that are internally unique, and the percent of sequences that are genuinely new are then calculated as a means of keeping track of the progress in sequencing the available miRNAs in a genome.

BLASTing and Sequence Binning

For each input folder of sequenced concatamers, there is now a FASTA file of new sequences derived from that folder. This initial set of new sequences includes fragments of mRNAs, tRNAs, rRNAs, other unknown sequences, and sequences from background RNA from other genomes such as the *E. coli* genome in this instance. The task, therefore, is to bin these sequences based on their identity to sequences within databases of tRNAs, rRNAs, the genome of interest, other genomes, etc. To do this, sequences are first BLASTed using WU-BLAST[2] against the genome of interest, any potential background genomes, and separate files containing known rRNA and tRNA sequences. BLAST parameters were set at E=15 Q=10 R=5 W=6 (E= expectation value, Q= gap opening penalty, R= gap extension penalty, W= word size). The BLAST output is parsed and candidate sequences binned into separate files depending on which of the genome databases they match best. A cutoff of 90 percent identity between the sequence and the genome was set in order to filter the list of matches and in the process tolerate one to two mismatches due to errors in the sequencing of the original concatemer or, less likely, in the genome of interest. The bins are then compared, and sequences found in the bin corresponding to the genome of interest and not also found in the other bins containing sequences that match tRNA, rRNA, background genomes, etc. are written to a separate folder-specific file containing sequence ids with their genomic locations, lengths, sequences and percent match to the genome as determined in the BLAST step. A file called summary.BLAST that summarizes the BLAST results for all of the sequences is also generated.

Heterogeneity Grouping and Filtering

At this step, each input folder of concatamers contains a file of candidate miRNA sequences that specifically hit the genome and are not fragments of tRNAs and rRNAs, or

background sequences from another genome. In order to further parse down the set of candidate miRNAs, we make use of a feature of Dicer processing of the miRNAs that was noticed as we began building our data set. In particular, for many sequences there is noticeable length heterogeneity at the 3' end of the miRNAs and less often at the 5' end as well. A program called hgroup.pl first archives the candidate sequences together with their genomic locations to a file containing all the previously determined candidate miRNAs. This archive is then structured into heterogeneity groupings using the genomic locations of the miRNAs. Sequences are placed in a heterogeneity group when they share the same 5' position but have different 3' positions (3' heterogeneity), share the same 3' position but have different 5' positions (5' heterogeneity), or when one sequence is fully overlapped by a longer sequence (5' and 3' heterogeneity). These groups often contain multiple sequences, and are in turn ordered to determine a parent sequence that is most representative of the group. Ordering of each heterogeneity group is done by first sorting the group based on the percent match of each individual sequence to the genome, then on cloning frequency (higher cloning frequency is preferred), and finally on length (higher sequence lengths are preferred). This ensures that the representative parent sequence is not only the best match to the genome, but also the most frequently cloned of the sequences in the group. In cases where there is more than one sequence in the group that match the highest cloning frequency, the longest sequence out of this subset is selected. An updated list of all the archived parent sequences is then generated, and the list of new candidate miRNA sequences filtered against this archive. Candidate sequences that are not also parent sequences are then removed, thereby further minimizing the set of novel candidate miRNA sequences.

Secondary Structure Predictions

Two potential precursor foldback structures are then generated for each candidate miRNA, with the miRNA placed at either the 5' end or the 3' end of the foldback. Using a program called fold_uRNAs.pl, the genomic coordinates of the filtered set of candidate miRNA sequences are first used to extract additional surrounding sequence from the genome for secondary structure predictions. The correct strand of the genome to extract sequence information from, is determined using the candidate miRNA BLAST coordinates. The sense strand of the genome file is used if the miRNA ending position is greater than the miRNA beginning position, whereas the antisense strand is used if the ending position is less than the

beginning position. To generate the precursor foldbacks with the miRNA at either the 5' or 3' ends, we use a window on the appropriate strand of the genome encompassing an additional 15 bases upstream of the beginning position of the miRNA and 60 bases downstream of the ending position of the miRNA for the 5' precursor, and another window containing 15 bases downstream and 60 bases upstream for the 3' precursor. Two additional structures are also generated using windows of 1000 bases either upstream or downstream of the candidate miRNA to determine if the candidate miRNA belongs to a cluster containing multiple genomically-proximal miRNAs.

All secondary structure calculations are done using the Zuker folding algorithm within the Vienna RNA software package[3], and postscript files containing the candidate miRNA sequence highlighted in red written to a separate folder. The set of predicted secondary structures is in turn examined to select good precursor foldbacks containing new miRNA sequences. This step involves some judgment on the part of the user. We generally look for secondary structures that are stem-loops with a well-paired stem and a loop that is not too large (<10 nucleotides). A certain degree of mismatches and bulges is tolerated throughout, however the miRNA should be located in a well-paired region and a few bases beyond the miRNA 5' and 3' termini should generally be completely base-paired. Additionally, the miRNA should be located within approximately ten bases of the loop. MicroRNAs from the candidate list that fit this description are considered to be potentially novel miRNA genes.

Gene Annotation and Spreadsheet Generation

A number of additional datum are then determined for each potential miRNA gene. The archive of all the sequences ever cloned is first grouped using a variant of the hgroup.pl program described above. For each miRNA gene, the frequency with which that exact sequence was cloned is calculated, together with a total frequency that includes length-heterogeneous examples of the miRNA. A length range spanning all of the sequences within the length heterogeneity grouping is also calculated. If sequences were cloned from different stages or lineages, stage/lineage-specific cloning frequencies for the miRNA gene are then calculated using the stage/lineage identifier for each sequence within the heterogeneity grouping.

The miRNA genes are then annotated by feeding the BLAST coordinates to a program that searches and parses Genbank annotation files for the source genome. Information that is

retained includes whether the miRNA is located within an annotated exon or intron, the distance to the nearest gene, whether the miRNA matches a portion of an expressed sequence tag (EST), and what side of the genome the sequence is located on. The other sequences that are not predicted to form foldback structures are also annotated at this stage. This information is then used by a downstream script that predicts whether the cloned sequence falls into one of the classes of miRNA, siRNA, mRNA fragment, or unknown, using the following criteria.:
miRNA: Good predicted foldback. Not located in an exon. Does not match the antisense sequence of a known EST.
siRNA: Located in an exon or an intron. Matches the antisense sequence of a known EST.
mRNA fragment: Located within an exon or within 200 bases upstream or 500 bases downstream of an annotated exon or in an intron with no good foldback structure. Does not match the antisense sequence of a known EST.
Unknown: Sequence is neither an miRNA, siRNA, nor an mRNA fragment.

Homology Modeling

Potential paralogs (homologues within the same genome) and orthologues (homologues across different genomes) are found using the sequences of the cloned miRNAs genes. Paralogs are found by generating phylogenetic trees of the set of miRNA sequences, and identifying sequences that cluster together on branches of the tree. Orthologues of the miRNAs are found by BLASTing other genomes using either just the miRNA sequence or the entire precursor foldback sequence. For closely related genomes such as the *C. briggsae* genome in this instance, we found that it is possible to locate orthologous miRNAs by BLASTing the entire foldback sequence as there is still sufficient conservation in the region surrounding the miRNA sequence. When BLASTing the entire foldback, we used default nucleotide BLAST parameters. We then used a filter containing a seventy percent identity cutoff for the match between the foldbacks and its orthologue, and a 50-nucleotide cutoff on the size of the minimum matching region, in order to order the list of candidate orthologous foldbacks. The position of the candidate orthologous miRNA within a candidate orthologous foldback is then determined by aligning the original miRNA sequence and the entire candidate orthologous foldback and extracting the sequence with the maximum alignment score from the orthologous foldback. This sequence is considered to be an orthologous miRNA sequence if there is a good foldback with a well-positioned match for the

original miRNA within the foldback. For BLASTing more distantly related genomes such as *D. melanogaster* and *H. sapiens*, we found it necessary to BLAST using just the miRNA sequence with WU-BLAST parameters set at E=20 Q=10 R=5 W=9 (E= expectation value, Q= gap opening penalty, R= gap extension penalty, W= word size). This strategy works due to the higher conservation of the miRNA sequence relative to the entire foldback sequence, which reflects the higher selective pressure against sequence drift in the functional miRNA sequence relative to the precursor foldback sequence. A seventy percent identity cutoff was set to filter the BLAST output. The genome coordinates from the BLAST output were then used to extract surrounding sequence from the genome followed by secondary structure prediction for the candidate miRNA orthologue in the context of the surrounding sequence. Good foldbacks with well-positioned orthologous miRNAs are taken to be the likely orthologous miRNAs.

References

1. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).
2. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-410 (1990).
3. Mathews, D. H., Sabina, J., Zuker, M. & Turner, D. H. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J Mol Biol* **288**, 911-940 (1999).

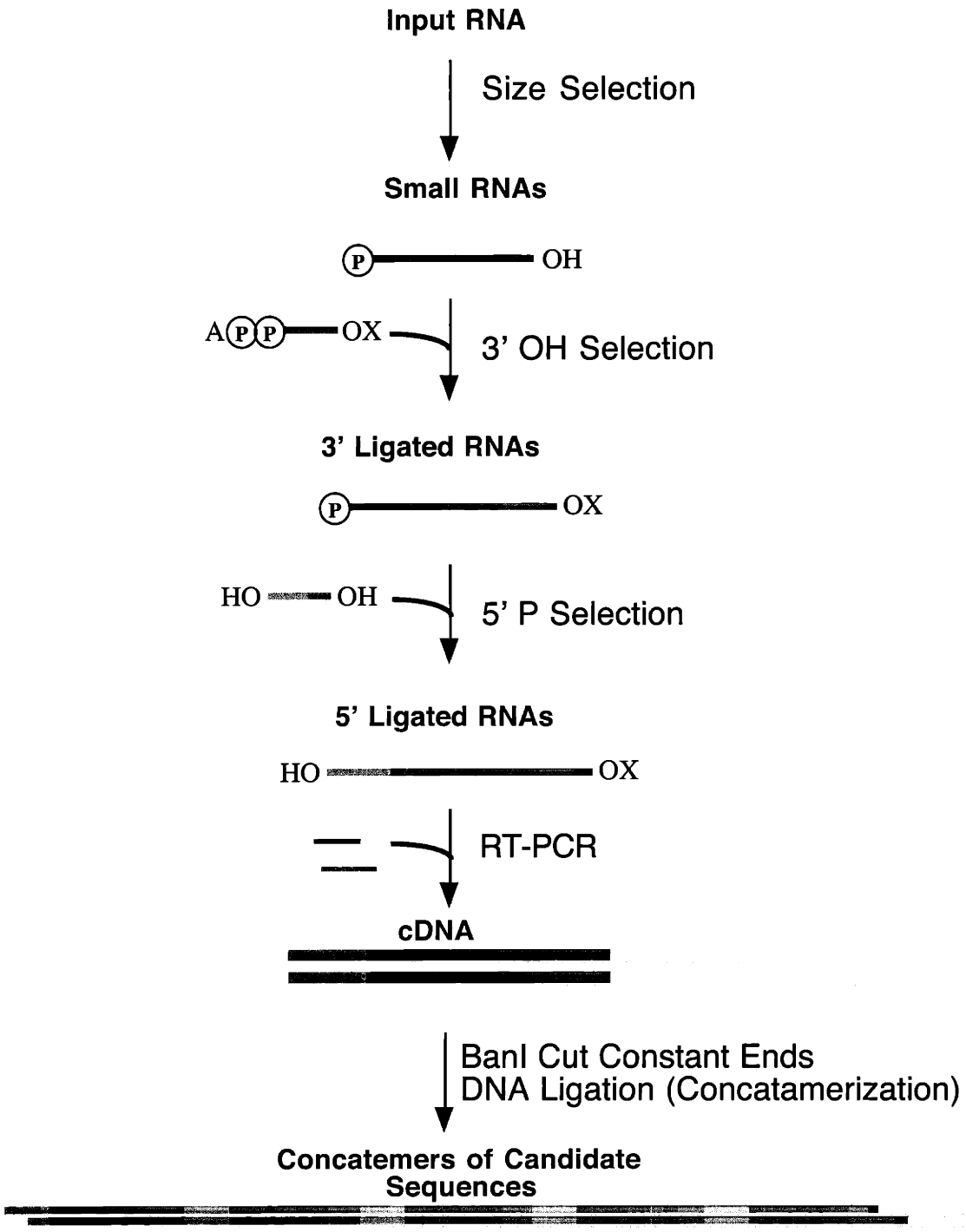


Figure 1. Cloning and sequencing large numbers of small RNA Dicer products. Total RNA is size selected and ligated to two different adapter oligos containing non-palindromic Ban I sites. These oligos are specific for Dicer products containing a 3'-terminal hydroxyl group and a 5'-terminal monophosphate group. Dicer products ligated to the two adapter oligos are then selectively amplified in an RT-PCR step. The resultant cDNAs are concatamerized and sequenced.

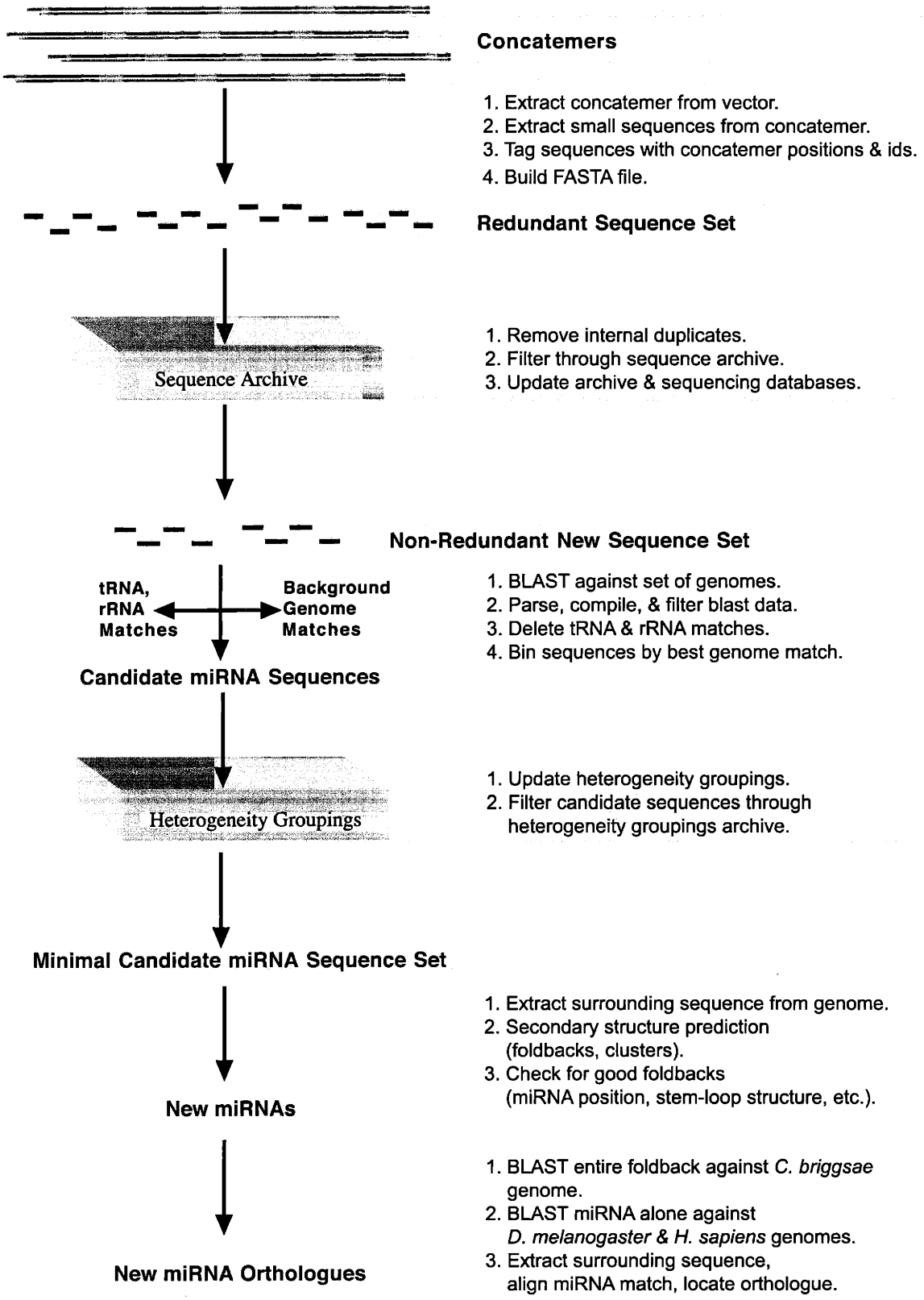


Figure 2. Schematic of the bioinformatics methodology.

Future Directions

Using an integrated experimental and informatics approach, we have now cloned a total of 75 miRNA genes in *C. elegans*. Together with 4 additional *C. elegans* miRNAs[1], 14 miRNAs cloned from *D. melanogaster*[2], 19 miRNAs cloned from *H. sapiens*[2] and 31 miRNAs cloned from mouse[3], the total number of confirmed miRNA genes across these four genomes is now over one hundred. Additional large-scale miRNA cloning projects underway in the mouse have already found over one hundred candidate miRNAs, the majority of which have orthologues in the human genome. MicroRNAs are therefore an extensive class of noncoding RNAs, with representative examples found across a number of eukaryotic genomes. Yet, only two of these genes, *lin-4* and *let-7,* were identified using classical forward genetic approaches and characterized in any detail. The existence of such a large data set begs the more functional questions related to miRNAs. Fortunately, large data sets also contain hidden trends and data points. Here I present some interesting questions related to the function, mechanisms, and regulation of miRNA genes. I also outline possible experimental and computational approaches to these questions with an emphasis on how the growing set of miRNA genes might be utilized in addressing these questions.

MicroRNA Targets

An understanding of the code for miRNA gene-specific recognition is particularly important as an initial step towards integrating miRNAs into classical models of gene regulatory networks where the emphasis has traditionally been on transcription factors as the primary gene-regulatory molecules. Thus, one would like to understand which miRNAs regulate which genes. This regulation need not be one to one. Rather there may be single miRNAs that regulate sets of genes. Conversely there may be sets of related miRNAs that co-regulate a single gene or a set of functionally related genes. In all cases, however, there must be information that is used to "wire" a gene regulatory network such that specific miRNAs are linked to the regulation of specific genes. More specifically, one would like to get at what information is contained within a regulated gene that specifically directs recognition and regulation by a given miRNA sequence.

Gene recognition for both *lin-4* and *let-7* is specified by complementary sequence elements in the 3' UTR of target genes[4-7]. Determination of likely target sites was possible in

these two instances because the miRNAs and their target genes were identified based on observed phenotypes and could therefore be placed within the framework of a known gene-regulatory pathway. Given a set of likely target genes, potential target sites were identified based on their complementarity to the known miRNA sequences. and these target sites were experimentally tested for functionality. Identifying target genes and target sequences for the miRNAs found using the current strategy is more challenging due to the lack of known phenotypes for these miRNAs.

Ideally, to identify target genes for sets of miRNAs one would simply search for complementary sites within well-defined 3' UTRs. However the problem is complicated by the fact that, as observed for *lin-4* and *let-7,* target sequences are generally not the absolute antisense of the miRNA sequence. Thus, all of the target sites identified and confirmed for *lin-4* and *let-7* contain bulges and mismatches, and in one instance, a bulged C seems to be essential for specifying target gene regulation[8-10]. Indeed because the RNAi pathway makes use of perfectly complementary sequences of 22 nucleotides to direct target mRNA degradation, miRNA recognition of simple antisense sequences might be rare in order to avoid crosstalk between the factors involved in these two pathways. Perhaps, miRNAs make use of a more complex code that specifies target recognition.

Here I would like to outline how a mix of computational prediction and experimental verification might be used to bootstrap a solution to the problem of target gene and target site prediction. Given that target recognition is likely to be complex as discussed above, target prediction is best done using a statistical model that captures functionally significant features of miRNA/target recognition and scores potential target sequences as likely targets for an input miRNA sequence. In order to begin building such a statistical model, however, many more examples of experimentally verified target sequences are needed. As a first step towards bootstrapping this problem, I wrote a PERL script called targets.pl that takes an input miRNA and UTR sequence pair and generates a set of candidate target sequences. Six parameters that characterize the interaction between an miRNA and a predicted target were incorporated into targets.pl, and I initially defined non-statistical cutoff values for each of these parameters based on the predicted structures of the known *let-7* and *lin-4* miRNA/target site interactions. The parameters are as follows:

1. Percent miRNA/target bases paired: minimum=50%.
2. Percent of G/C + G/U base pairs that are G/C: minimum= 30%.
3. Number of bases looped out on one side of the miRNA/target structure: maximum= 5 bases.
4. Total number of loops: maximum= 2 loops.
5. Number of unpaired bases at the 5' end: maximum= 3 bases.
6. Number of unpaired bases at the 3' end: maximum= 3 bases.

The purpose of targets.pl is to generate small sets of likely target sequences for a given UTR that can be experimentally verified as a first step towards building a more statistical model of miRNA/target recognition. Multiple targets will need to be verified, and it is therefore useful to start with a set of candidate miRNAs and a set of candidate gene UTRs that are likely targeted by these miRNAs. The heterochronic pathway is a particularly useful model system in this regard. Using the Northern data presented here and elsewhere for miRNAs in the heterochronic pathway useful pairs of input miRNA and UTR sequences can be derived by clustering miRNAs into sets containing members with similar expression profiles. These clusters can then be correlated with known genes that show altered expression profiles in the presence of the miRNAs within the cluster, thereby yielding a set of miRNAs and coding gene pairs that are likely to be connected in some regulatory network. Use of this input set should help narrow the initial search space of potential target genes and maximize the chances that predicted targets are confirmed experimentally. In addition, anti-correlated miRNA/gene pairs can later be used as a negative control to check the developing statistical model.

Targets.pl first scans a UTR for imperfect antisense versions of the miRNA sequence of interest, allowing for a user-defined mix of mismatches, deletions, and insertions between the miRNA sequence and candidate target sequences. For each candidate target sequence, a composite miRNA/target sequence is generated by linking the miRNA sequence to the target sequence via a set of four N bases. Secondary structures for these composite sequences are then calculated using the Zuker RNA folding algorithm, where each of these structures is intended to model the potential recognition of the target candidate by the input miRNA. For each of these secondary structures, values for the parameter set listed above are calculated, and structures that meet the cutoff values are promoted to a list of likely candidate target sites.

Two additional features were built into targets.pl for ordering sets of candidate target sequences. An average distance measure is calculated for the set of candidate targets that captures any clustering of predicted targets in the UTR. This is not used as a cutoff parameter, and is instead based on the observed clustering of *let-7* and *lin-4* sites within the *lin-14* and *lin-41* UTRs[4]. Candidate target sequences that are near to each other are then grouped and aligned to determine if there is significant conservation amongst them. Based on *lin-4* and *let-7*, sequence conservation within closely spaced candidate target sequences may suggest functional relevance.

A second feature makes use of target alignments from two orthologous genomes. Because there is selective pressure against sequence drift in functionally relevant sequences, targets can be ranked based on conservation across orthologous genomes. Thus, two well-conserved target candidates are generally more likely to be functionally relevant and can be given higher scores. This is particularly useful in cases where an orthologous miRNA to the one being tested for targets has been predicted or cloned in a closely related genome. In particular, covariation in the miRNA/target structure between closely related genomes, where this covariation is consistent with preservation of base-pairing between the miRNA and its target sequence, may be a good metric for ranking potential targets. For example, genome A might have the miRNA sequence 5' G*G*GAAACCC 3' and a candidate target sequence 5' GGGUUUC*C*C 3'. Genome B might then contain an orthologous miRNA 5' G*C*GAAACCC 3' and a candidate target sequence 5' GGGUUUC*G*C 3'. Here, miRNA B varies from miRNA A in the second position and contains a C instead of a G. In turn, the candidate target site in genome B contains a G that may have arisen as a compensatory mutation to preserve a G/C base pair as part of a functionally useful secondary structure (i.e. a structure involving miRNA binding to this target site). Compensatory mutations such as these can be scored over all the co-varying positions and in turn used to rank candidate target sequences. Thus, given two orthologous pairs of miRNAs and sets of potential targets from two orthologous UTRs, the program finds pairs of optimally aligned target sequences across the two sets of orthologous candidate target sequences. The orthologous miRNAs are then aligned and positions that co-vary in the alignment correlated with the predicted miRNA/target interaction.

There is an additional feature apparent in our data set that might be built into a predictive model such as this. In particular, dendritic treeing and sequence alignments of our total miRNA

sequence set reveals that miRNA sequences tend to be more conserved at their 5' ends, with conservation falling off after approximately eight bases in most cases (Figure 1). Additionally, there are a number of distinct sequence families that contain paralagous sequences with 5' conservation. This suggests that there may be 5' motifs that are used by miRNAs in distinct pathways to recognize sets of genes involved in that pathway. For example, we now have a set of four paralogs of the *let-7* sequence (E. Weinstein, N. Lau, D. Bartel. In preparation). All four sequences in this set contain a consensus UGAGGUAG sequence at their 5' end, suggesting that they may recognize a common set of genes containing the antisense of this motif in their 3' UTR sequences. Further discrimination of specific genes and target sequences by miRNAs such as these *let-7* paralogs may then come from information specified in the 3' end of the miRNA sequence. In predictive or statistical models, it may therefore be useful to weight base-pair scoring to the target sequence more at the 5' end of the miRNA than at the 3' end and to search for target sequences that specifically contain the antisense of conserved 5' motifs for cases such as this *let-7* miRNA group. This type of model is consistent with the recently proposed "area code" model in which a general set of target genes is specified using sets of consensus 5' end sequences, followed by more specific gene recognition using information at the 3' end of the sequence[11].

As a test case, the 3' UTR of the *hbl-1* gene was scanned with targets.pl using the sequence of the *let-7* miRNA that is thought to be involved in regulating *hbl-1*. This gene was chosen because its 3' UTR sequence is available for three closely related genomes, namely *C. elegans, C. briggsae*, and *C. remanei*. Nine targets with orthologous versions in at least two out of the three genomes were found. One of these targets was absolutely conserved in all three of the orthologous UTRs, and another was absolutely conserved between the *C. elegans* and the *C. remanei* UTRs. Interestingly, conserved sets of targets tend to occur at similar positions relative to the start of their respective UTRs and a cluster of five targets spanning approximately 400 bases is predicted in the *C. elegans* UTR.

Target prediction using targets.pl is the first step towards building a statistical model of miRNA target recognition. Sets of predicted targets must then be experimentally confirmed in order to get accurate sequence and structure statistics for such a model. Gene fusions of UTR sequences to reporter genes such as *lacZ* are a useful tool for confirming the functionality of predicted target sequences. This technique was previously used to confirm predicted *let-7* sites

in the 3' UTR of *lin-41* by showing that the *lacZ/UTR* fusion has an identical temporally-regulated expression profile to *lin-41* and that this profile is dependent on the predicted *let-7* sites, as indicated by loss of this profile upon deletion of the *let-7* target sites[4]. Optimal target sites predicted by targets.pl can be confirmed in a similar manner for genes with defined expression profiles. One would then want to get at which features of the target are most important for specifying miRNA recognition. Thus, instead of deleting target sites from the 3' UTR, confirmed sites could be systematically altered and the resultant calculated miRNA/target secondary structure correlated with changes in the expression profile for the reporter gene. MicroRNAs such as those in the *let-7* paralog set and miRNAs with known orthologues are initially good candidates for experimental target verification.

In summary, the rapidly expanding database of known miRNA sequences can be used to build a statistical model of miRNA gene target site recognition sequences by first correlating sets of miRNAs with sets of genes they are likely to regulate. The heterochronic pathway is a useful model network in this regard because the expression profiles and phenotypes for a number of protein-coding genes in the pathway are known, and these profiles can be correlated with Northern expression profiles for the miRNAs in the database. Given a set of protein-coding genes and a set of miRNA genes whose expression is well correlated with the protein-coding genes, predictive programs such as targets.pl can be used to generate sets of likely candidate target sequences. Targets can be then be confirmed experimentally using reporter gene fusions to UTRs with and without the predicted target sites. Predicted secondary structures for miRNA sequences bound to confirmed target sequences can then be fed back in to an evolving statistical model containing parameters such as those used in targets.pl. This evolving model can in turn be used to better predict candidate target sequences for experimental verification thereby steadily improving a statistical model for target sequence prediction.

Regulation of miRNA Genes

MicroRNA genes, like protein-coding genes, are likely subject to regulation that is in turn correlated to their role in specific gene-regulatory networks. There are a number of levels at which regulation of miRNA genes might take place and the large data set of confirmed miRNA genes is likely to be very useful in fully delineating each of these regulatory levels. Here I would like to outline approaches to understanding the mechanisms and relative importance of regulation

at the levels of transcription, precursor processing from a primary transcript, miRNA processing from a precursor foldback structure, and the regulation of clusters of genomically-proximal miRNA sequences.

Transcription

Cis-regulatory elements in sites upstream of a gene influence regulation of that gene at the transcriptional level through recruitment of specific transcription factors and RNA polymerases. Identifying these gene-regulatory transcriptional elements (TEs) is a very difficult problem due to the high number and variability of TE sequences. For instance the TRANSFAC database, which catalogs known TEs, contains over 4,000 sequence motifs[12]. These motifs are often highly variable and even well-conserved motifs such as the TATA sequence are not always present when expected (i.e. in pol II promoters)[13].

Many of the miRNAs in our data set are located at large distances from the nearest gene, and are therefore likely transcribed as independent units. Some of these miRNAs are also located in clusters that may share a set of common miRNA TEs (Figure 2). However, identifying miRNA TEs is further complicated by the fact that the polymerase responsible for transcribing microRNA genes is as yet unknown. Additionally, there may be sets of motifs that are unique to specific sets of microRNA genes. For example the four *let-7* paralogs are coexpressed beginning at the L3 stage, and this coexpression may be regulated by a common TE.

Here I would like to focus on the use of a combination of inter- and intra-genome sequence comparisons as a method for identifying putative miRNA TEs based on conservation of sequence (i.e. function)[14]. Comparative genomics approaches such as this were successfully used in a recent paper that aligned human and mouse sequences to identify binding sites for muscle-specific transcription factors[15]. The large set of sequenced miRNA genes together with the set of orthologous miRNA genes identified in a number of genomes naturally lends itself to applying TE-finding methods such as this based on sequence conservation.

A necessary first step in defining miRNA TEs using sequence comparisons is to locate the regions likely to contain TEs for each miRNA gene in the data set. This might be accomplished using a multi-step process involving localization of the 5' transcription start site followed by inter-genome alignments of closely related species to define more specific regions

that are likely to contain regulatory elements. Intra-genomic alignments across the set of known miRNAs can then be used to define specific TE sequence motifs.

Rapid amplification of cDNA ends (RACE) applied to the 5' end of the miRNA genes (5' RACE) is a useful technique for accurately defining the 5' end of the full length miRNA transcript[16]. Briefly, DNA-free RNA transcripts are reverse transcribed and a specific constant sequence ligated to the 3' end of the cDNA. These are then PCR-amplified using a gene specific primer and a primer to the 3'-ligated constant sequence and subcloned and sequenced. 5' RACE might thereby be used to rapidly define the 3' edge of the transcription-regulatory upstream regions for all the miRNA genes in the data set.

The second step is to limit the search space to specific regions of the upstream sequence beyond the transcription start site. This is essential when searching for TEs in the noncoding upstream regions of higher eukaryotes genes as the relevant upstream region can be both very long and of highly variable size. For example, gene density in *D. melanogaster* is on the order of one gene per 9 kb. Yet the average size of a transcribed RNA is only 3058 bases[17]. There is therefore generally a large amount of sequence upstream of any given gene in which transcriptional regulatory elements might reside. A number of recent papers have shown that inter-genomic sequence comparisons can be used to locate potential regulatory elements based on sequence conservation[18,19]. A useful method for narrowing the search space in these comparisons, is to first find blocks of conserved ungapped sequence in which specific regulatory elements are most likely to reside. Recently, a Bayesian block alignment (BBA) algorithm[20] was adapted to find such conserved ungapped blocks by using nucleotide comparisons of the upstream regions of a set of 28 orthologous gene pairs that are specifically upregulated in skeletal muscle[15]. The BBA works by calculating every possible alignment for an input sequence pair. The probability that a base at position i in the first sequence of the pair is conserved with a base at position j in the second sequence of the pair is then plotted for each possible i and j. By summing the probabilities over j for each position i in the first sequence, ungapped blocks of high sequence conservation in the first sequence are identified. Application of the BBA to the set of 28 orthologous gene pairs that are specific to skeletal muscle allowed the authors to narrow their search for transcription factor binding sites to regions comprising just 19% of the upstream sequences where these regions had greater than 50% probability of being

located in a conserved sequence block[15]. A similar application of the BBA to the large set of orthologous *C. elegans/C. briggsae* and mouse/human gene pairs should prove useful here once the transcription start sites have been identified.

The next step is to locate specific motifs within the sets of larger blocks of statistically conserved sequences using intra-genomic sequence comparisons. It may also be useful to cluster miRNA genes prior to this step based on coexpression data as discussed below. The challenge is to locate motifs in the absence of *a priori* knowledge as to their size, location and even degree of conservation amongst the miRNA genes in a single genome. Statistical alignment methods such as the Gibbs sampling algorithm[21,22] should prove useful here. Very generally, the Gibbs sampling algorithm is an iterative sampling and optimization procedure whereby sequence motifs are identified based on an evolving weight matrix describing the alignment over all the input sequences. The algorithm is iterated to maximize the probability that the identified motifs are located within the regions covered by the weight matrix and not located in the remaining background regions. Because the algorithm is initially seeded with a random alignment, and subsequent motifs randomly selected at each iteration from a distribution derived from the evolving weight matrix and background frequencies, no *a priori* specifications as to their position and sequence composition need be made. Furthermore, the algorithm tends to converge to a solution in N-linear time where N is the number of input sequences, which should allow for alignments over the large data set of miRNA genes.

As mentioned above, clustering of miRNA genes based on their expression profiles will also be useful as a means of classifying TEs identified based on sequence comparisons. Thus, coexpressed miRNAs may share a specific set of TEs whereas other TEs might be located more broadly amongst all miRNA promoters. The assumption here is that coexpression reflects coregulation and that coregulation is caused by an underlying set of common TE motifs. Coexpression clustering is an increasingly common technique in analyzing expression profiles derived from microarray data to identify cis-regulatory elements[23,24]. For miRNA genes, coexpression of the short miRNA sequence is not necessarily the best indicator that these sequences share common TEs, as transcribed miRNAs are likely subject to further regulation at the level of processing of the miRNA from precursor foldbacks and possibly also the processing of precursor foldbacks from longer primary transcripts (see below). Microarray experiments using probes to defined full-length transcripts identified using 5' RACE may therefore provide

the most useful data for clustering sets of coregulated miRNA genes. It is likely that clustering prior to sequence alignments could prove very useful for reducing the noise in these comparisons by restricting the input sequences to a set of genes that are more likely to share common regulatory motifs. Clustering miRNAs based on their known target genes might also allow for identification of motifs that regulate miRNAs involved in common gene-regulatory networks.

MicroRNA Processing

MicroRNAs are processed from precursor foldbacks. Additionally, in certain cases, such as the miR-35 cluster (Figure 2), we have been able to observe intact precursor foldbacks that were presumably processed from a longer primary transcript[25]. Additional regulation of miRNA expression may therefore take place post-transcriptionally at two levels of processing, namely processing of the precursor foldback from a longer primary transcript, and processing of the miRNA from the precursor foldback. Interesting questions in this regard are what information directs both of these processing events, and how is this information encoded at the primary sequence and secondary structure levels. Furthermore, this information will need to be integrated into some model describing how complexes containing Dicer, the Argonaute family of proteins and other, as yet unidentified, members, utilize this information to temporally and spatially regulate miRNA formation. Here again the large data set of miRNA genes lends itself to sequence comparison approaches to this problem.

At the level of miRNA processing from the foldback, one immediately noticeable feature is the nearly consensus U residue at the first position of all the miRNA sequences (Figure 3). This likely reflects the known preference of Dicer for processing out small RNAs that begin with a U. Additional information might be encoded in the second position where there is a two to one bias towards purine residues. Beyond this, however, there does not appear to be any consensus sequence in either the surrounding foldback sequence or the actual miRNA sequence that could code for processing of a specific small RNA from the large precursor foldback. Secondary structure elements such as the location of the miRNA sequence relative to the loop, base-pairing within the miRNA, loop sizes, etc. are therefore likely to contain additional information. Indeed these secondary structure parameters are an essential part of a predictive program for locating miRNA sequences in the genome (L. Lim, D. Bartel, C. Burge. In preparation). In order to deconvolute the relative importance of these primary and secondary structure parameters it

would be interesting to systematically vary either the surrounding foldback context sequence and miRNA position using a known miRNA sequence or the miRNA sequence in the context of a known foldback sequence. Using an *in vitro* system similar to that developed for studying processing in RNAi[26] processing of the miRNA from the precursor foldback could then be monitored.

At the level of foldback processing from a larger primary transcript, the challenge is to locate sequence elements in the primary transcript proximal to the foldback that might direct processing. Local alignments using a combination of inter- and intra-genomic comparisons as outlined in the section on finding transcriptional regulatory elements might prove useful for finding such sequence motifs. Alternately, the foldback secondary structure might be recognized by some processing molecule independently of particular foldback-proximal sequence motifs. Use of an *in vitro* processing system might again be useful here. For instance, known foldback sequences might be placed in the context of highly degenerate surrounding sequences, and processing of the foldback from the surrounding degenerate sequence assayed, to determine if there is any contextual effect.

Why miRNAs

I would like now to briefly discuss some of the broader theoretical questions related to this extensive set of genes. In particular, it is interesting to speculate as to the origin and usefulness of this class of genes. In addition, in considering the seeming imbalance between the size of the proteomes and the phenotypic complexity of higher order eukaryotes, it is interesting to determine if the function of miRNAs and the extent of this class might account for some of this biological "dark matter".

MicroRNAs may be vestiges of a precursor RNA world[27] co-opted to act as regulators of gene expression or they may have arisen later in evolution. It is somewhat difficult to imagine a role for miRNAs in the RNA world as very small RNA sequences are likely to have very limited functionalities on their own. However, under either scenario, miRNAs must possess some niche utility or advantage that gene-regulatory protein molecules did not evolve to replace. The rapidity with which RNA is synthesized compared to proteins is one such advantage. This is particularly useful in tightly regulated gene networks, where signaling molecules such as miRNAs would have to be turned on rapidly at specific points[28]. MicroRNAs also possess high

target specificity for other genes due to their sequence information content, which is sufficient to uniquely specify any sequence in even the human genome. This is another property that is useful in their role as signaling molecules and which is more difficult to achieve with small protein molecules. Indeed some have suggested that miRNAs might mediate synaptic plasticity[28] by rapidly and specifically regulating expression of dendritically-localized mRNAs.

Because miRNAs possess high sequence specificity and can be rapidly synthesized they may be the ideal signaling molecules for "wiring up" gene-regulatory networks. This is likely to be of great interest from the standpoint of the increasingly popular class of network-oriented models for biological complexity[29-31]. Under these models, sheer proteome size need not be the sole determinant of phenotypic complexity. Instead increased phenotypic complexity can arise from recruiting a fixed set of components and evolving novel connections and increasingly complex network dynamics. Indeed much progress in understanding how complex phenotypes can emerge from evolving regulatory network architectures has come from modeling gene regulatory networks, where a fixed set of components can demonstrate complex circuit behavior such as multi-stable states, fixed point attractors, bimodal switch-like behavior, etc[29].

In addition, as proposed by Mattick et al.[32], small RNAs derived from introns may function as signaling molecules that can multitask the output of the genome by acting as a second signaling output from a gene transcription event in addition to the coding exons[33] . Indeed both *lin-4* and *let-7* come from regions that appear to be vestigial exons based on their base composition, suggesting that they may have originally arisen as multitasking signaling introns[32] . A number of the miRNAs in our data set are located in annotated introns, and it will be interesting to determine if there are signals of vestigial exons surrounding the other sequences.

Because miRNAs are such ideal wiring molecules, and perhaps multitasking molecules as well, they may be recruited to form novel higher order networks in the genomes of higher order species. Ultimately, therefore it will be very informative to develop models of specific gene-regulatory networks that incorporate miRNAs to determine if they exhibit higher-order dynamics. It may turn out that, as an extensive class of well-suited gene-regulatory molecules, miRNAs are the "dark matter" underlying much of biological complexity.

References

1. Lee, R. C. & Ambros, V. An extensive class of small RNAs in Caenorhabditis elegans. *Science* **294**, 862-864 (2001).
2. Lagos-Quintana, M., Rauhut, R., Lendeckel, W. & Tuschl, T. Identification of novel genes coding for small expressed RNAs. *Science* **294**, 853-858 (2001).
3. Lagos-Quintana, M. *et al.* Identification of tissue-specific microRNAs from mouse. *Curr Biol* **In Press.** (2002).
4. Reinhart, B. J. *et al.* The 21-nucleotide let-7 RNA regulates developmental timing in Caenorhabditis elegans. *Nature* **403**, 901-906 (2000).
5. Seggerson, K., Tang, L. & Moss, E. G. Two Genetic Circuits Repress the Caenorhabditis elegans Heterochronic Gene lin-28 after Translation Initiation. *Dev Biol* **243**, 215-225 (2002).
6. Lee, R. C., Feinbaum, R. L. & Ambros, V. The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* **75**, 843-854 (1993).
7. Moss, E. G., Lee, R. C. & Ambros, V. 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 (1997).
8. Slack, F. & Ruvkun, G. Temporal pattern formation by heterochronic genes. *Annu Rev Genet* **31**, 611-634 (1997).
9. Wightman, B., Ha, I. & Ruvkun, G. Posttranscriptional regulation of the heterochronic gene lin-14 by lin-4 mediates temporal pattern formation in C. elegans. *Cell* **75**, 855-862 (1993).
10. Ha, I., Wightman, B. & Ruvkun, G. A bulged lin-4/lin-14 RNA duplex is sufficient for Caenorhabditis elegans lin-14 temporal gradient formation. *Genes Dev* **10**, 3041-3050 (1996).
11. Lai, E. C. Micro RNAs are complementary to 3' UTR sequence motifs that mediate negative post-transcriptional regulation. *Nat Genet* (2002).
12. Wingender, E., Dietze, P., Karas, H. & Knuppel, R. TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* **24**, 238-241 (1996).
13. Pugh, B. F. Control of gene expression through regulation of the TATA-binding protein. *Gene* **255**, 1-14 (2000).
14. Hardison, R. C. Conserved noncoding sequences are reliable guides to regulatory elements. *Trends Genet* **16**, 369-372 (2000).
15. Wasserman, W. W., Palumbo, M., Thompson, W., Fickett, J. W. & Lawrence, C. E. Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* **26**, 225-228 (2000).
16. Liu, X. & Gorovsky, M. A. Mapping the 5' and 3' ends of Tetrahymena thermophila mRNAs using RNA ligase mediated amplification of cDNA ends (RLM-RACE). *Nucleic Acids Res* **21**, 4954-4960 (1993).
17. Ohler, U. & Niemann, H. Identification and analysis of eukaryotic promoters: recent computational approaches. *Trends Genet* **17**, 56-60 (2001).
18. Duret, L. & Bucher, P. Searching for regulatory elements in human noncoding sequences. *Curr Opin Struct Biol* **7**, 399-406 (1997).
19. Wasserman, W. W. & Fickett, J. W. Identification of regulatory regions which confer muscle-specific gene expression. *J Mol Biol* **278**, 167-181 (1998).
20. Zhu, J., Liu, J. S. & Lawrence, C. E. Bayesian adaptive sequence alignment algorithms. *Bioinformatics* **14**, 25-39 (1998).
21. Lawrence, C. E. *et al.* Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science* **262**, 208-214 (1993).
22. Liu, X., Brutlag, D. L. & Liu, J. S. BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. *Pac Symp Biocomput*, 127-138 (2001).

23. Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nat Genet* **22**, 281-285 (1999).
24. Spellman, P. T. *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell* **9**, 3273-3297 (1998).
25. Lau, N. C., Lim, L. P., Weinstein, E. G. & Bartel, D. P. An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* **294**, 858-862 (2001).
26. Tuschl, T., Zamore, P. D., Lehmann, R., Bartel, D. P. & Sharp, P. A. Targeted mRNA degradation by double-stranded RNA in vitro. *Genes Dev* **13**, 3191-3197 (1999).
27. *The RNA World* (ed. Gesteland, R. F., Cech, T. R. & Atkins, J.F.) (Cold Spring Harbor Laboratory Press, New York, 1999).
28. Ruvkun, G. Molecular biology. Glimpses of a tiny RNA world. *Science* **294**, 797-799 (2001).
29. Hasty, J., McMillen, D., Isaacs, F. & Collins, J. J. Computational studies of gene regulatory networks: in numero molecular biology. *Nat Rev Genet* **2**, 268-279 (2001).
30. Duboule, D. & Wilkins, A. S. The evolution of 'bricolage'. *Trends Genet* **14**, 54-59 (1998).
31. Hartwell, L. H., Hopfield, J. J., Leibler, S. & Murray, A. W. From molecular to modular cell biology. *Nature* **402**, C47-52 (1999).
32. Mattick, J. S. & Gagen, M. J. The evolution of controlled multitasked gene networks: the role of introns and other noncoding RNAs in the development of complex organisms. *Mol Biol Evol* **18**, 1611-1630 (2001).
33. Mattick, J. S. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep* **2**, 986-991 (2001).

Figure 1. Groups of paralagous sequences with conserved 5' motifs.
Dendritic treeing and sequence alignments of the entire data set reveals groups of related sequences. Conserved 5' motifs are shown in blue.

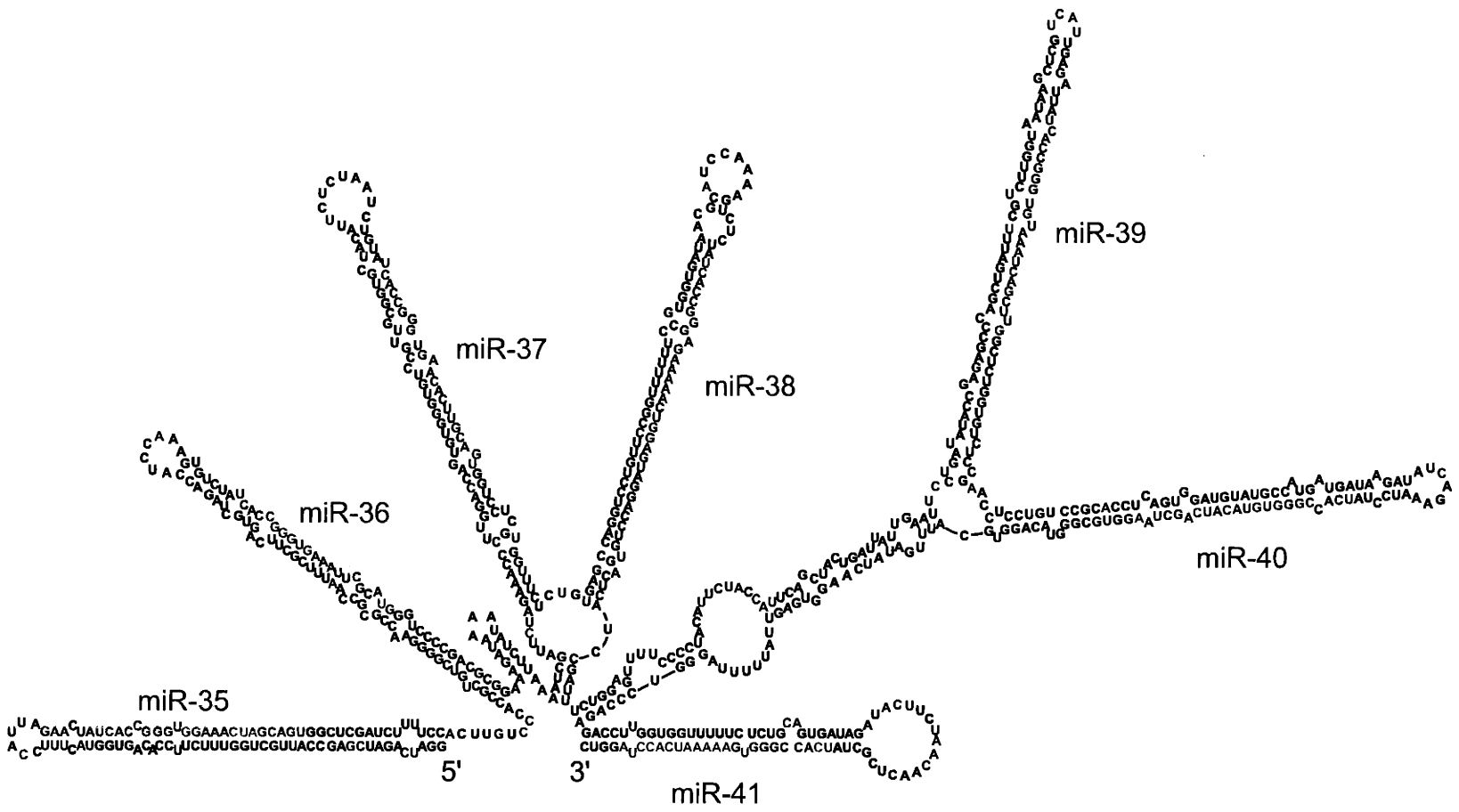


Figure 2. Cluster of microRNAs.
An example of seven microRNAs proximally located within a cluster in the genome. The microRNA sequences are shown in red.

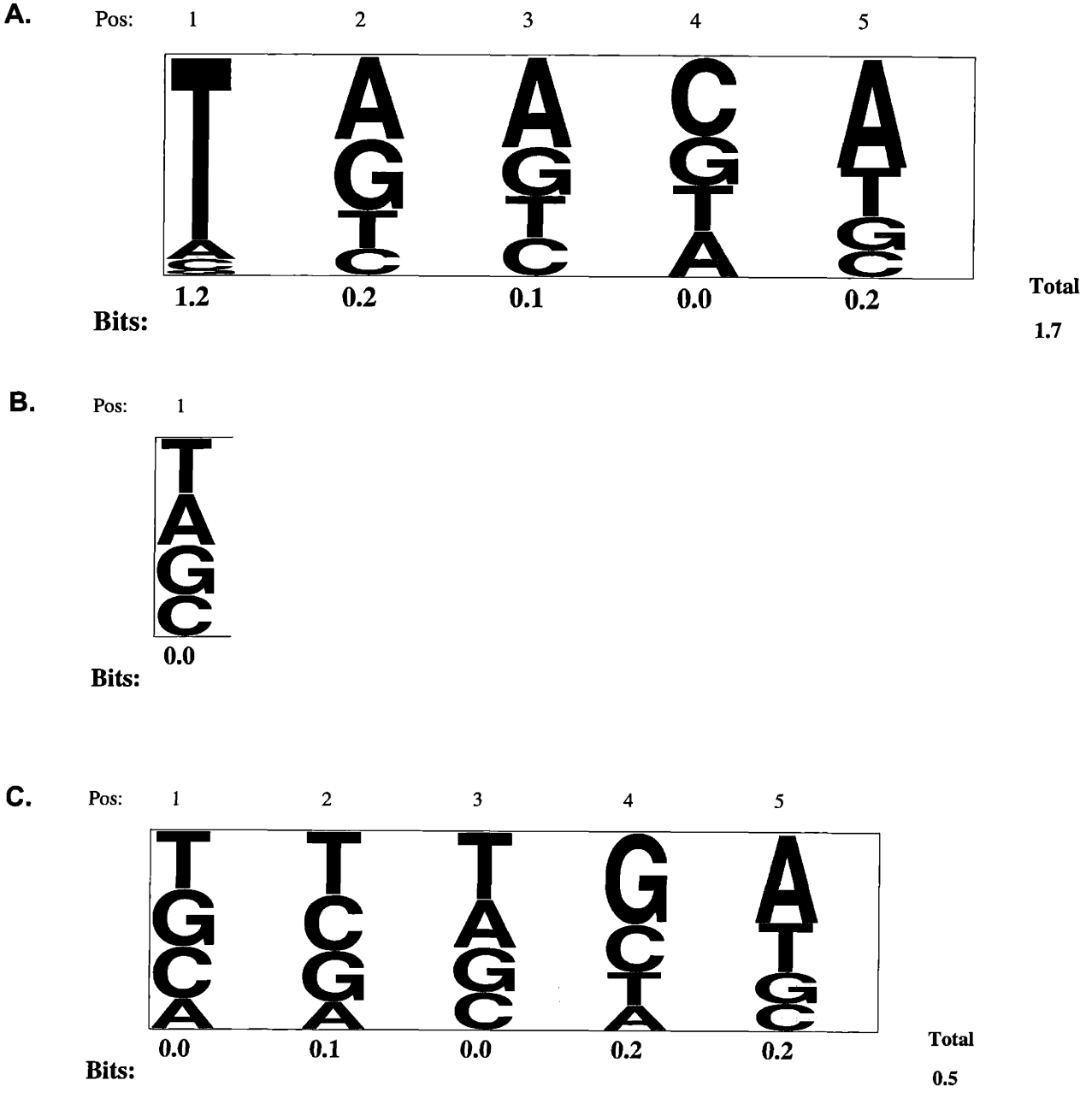


Figure 3. Information present in the microRNA sequences.
Sequence logos indicate the relative frequency of each of the four bases at each position of the microRNAs within the total data set. The information content for each position is indicated in bits.

(A) The first five positions of the microRNAs.

(B) A single column represents the middle portion of the sequences.

(C) The last five positions of the microRNAs.

Part II: A Photoactivatable Bifunctional tRNA for Directed Molecular Evolution

Introduction

Directed Molecular Evolution

Rational molecular design from physio-chemical first principles is generally a difficult problem. This is due to an incomplete understanding of the physio-chemical principles governing molecular structure and dynamics and to the difficulty of modeling well-understood principles at a high degree of precision[1]. In contrast, evolutionary processes found throughout nature represent an "irrational" design paradigm[2,3]. Under this paradigm, optimal solutions to a design problem are converged upon through random alterations and selection of functional variants rather than through top-down application of known principles. Evolution thereby enables generation of functional objects in the absence of any *a priori* understanding of the principles governing the design and function of such objects.

The design of systems that can recapitulate general evolutionary processes has garnered attention from a variety of fields. Examples include the development of genetic algorithms in computer science, use of neural networks in artificial intelligence, *in vitro* and *in vivo* systems for protein design, and the study of ribozymes and self-replicating peptides in evolutionary theory[4-8]. A common requirement for such systems is that they contain analogues of the operators that underlie evolutionary processes in nature, namely randomization and amplification, as well as a method for selecting objects bearing iterative improvements in some property of interest. This approach is referred to as directed evolution. Here I am interested in the construction of a novel directed evolution system that utilizes the cell's protein-synthesis machinery for the design of functional biomolecules.

In constructing a directed evolution system, it is useful to consider whether the system will have sufficient "evolutionary capacity" to solve a given design problem. Evolutionary capacity is used here to denote the ability of a directed evolution scheme to reach and explore areas of the state space containing every possible variant of the class of object being designed. This capacity is a combination of both the number of variants that the evolutionary scheme is capable of sampling at any given time point and the speed at which the scheme can iterate through the general evolutionary process of variation followed by selection. If this evolutionary capacity is too low, the directed evolution scheme will be unlikely to converge on a solution in a reasonable amount of time. For example, the sequence space for polypeptides is quite large. Even a small

100 amino acid peptide is located in a one hundred dimensional sequence space containing 20^{100} total objects. For any given sampling of this space, the sample must contain peptides with at least a minimal threshold level of functionality such that they can be selected from the general population. Yet, vast portions of the total sequence space are unlikely to contain variants with this threshold level of functionality where the size of these non-functional regions depends on the density of functional molecules within the space. It is essential, therefore, that a system for directed biomolecular evolution be capable of large samplings and rapid iteration of the sampling and selection steps in order to trace out a path in sequence space that leads to molecules with desired functionalities.

In general, using directed evolution one begins with a numerically large pool of diverse molecules that is subjected to some desired selective pressure. Selected molecules are randomized and amplified, and the selection scheme is reapplied. The stringency of the selection, which dictates the percent of the molecules in the initial population that pass through to the next round, is generally increased with each iteration of the selection cycle. For small sequences, the optimal molecule may already be present in the starting library and the challenge is to selectively amplify this molecule. More common, however, is the scenario in which molecules of low functionality are initially present, and only a small portion of sequence space is sampled in each round. Under such conditions it is essential to reintroduce diversity after each selection cycle so that the population of molecules begins to trace a path of increased fitness within the functional landscape instead of settling on some local optimum.

The general approach outlined above has proven instrumental for selecting functionally novel RNA molecules in studies of early molecular evolution[6,9,10]. Directed evolution works well with nucleic acids because they are easily amplified and randomized using polymerase chain reaction (PCR)-based methods[11-13]. Selected molecules can also be rapidly sequenced and characterized. However, combinatorial libraries of protein molecules cannot be easily evolved, as there is as yet no analogue of PCR for amino acid polymers. One solution is to tag the protein molecules in the library with their respective coding messenger RNAs (mRNAs) yielding evolvable fusions of information-bearing nucleic acids and functional protein molecules . The resultant library of hybrid peptide-mRNA molecules can then be used in a selection as outlined (Figure 1), where the mRNA tag enables PCR-based amplification and randomization of selected peptides followed by sequencing and characterization of evolved molecules. Previously

elaborated *in vitro* protein evolution schemes including ribosome display, mRNA-peptide fusions using puromycin, plasmid display, and phage display, as well as various *in vivo* schemes, have all been formatted to maintain this cis linkage between an information-bearing mRNA molecule and its encoded polypeptide[7,14-18]. Many of these systems are limited, however, by relatively small library sizes, constraints on the range of possible selection conditions, inherent limitations of *in vivo* steps, the need to do screens rather than selections in certain instances, and long time scales for multiple iterations of the selection cycle.

The following chapter describes a novel method for generating libraries of protein molecules linked to their encoding mRNAs that circumvents many of the above limitations. Central to this scheme is the synthesis and use of a photoactivatable bifunctional tRNA that serves as the linker between an encoded protein molecule and its respective mRNA. The system functions entirely *in vitro* and covalently links peptides to their encoding mRNAs via a stable amide bond, thereby expanding the range of possible selective conditions. In addition, encoded peptide libraries can be generated using a readily obtained prokaryotic translation system. I turn now to a brief description of the synthesis and properties of this bifunctional tRNA, as well as its use in generating evolvable combinatorial libraries.

Photoactivatable Bifunctional PHE-N- tRNA^{phe}

The system makes use of transiently stable mRNA-tRNA-polypeptide complexes normally formed during translation on the ribosome to generate libraries of covalently-linked mRNA-tRNA-polypeptide ternary complexes that can then be used to select for some property of interest. The tRNA in the complex is a synthetic bifunctional molecule (Figure 2) where the usual ester bond to its own amino acid has been replaced by a highly stable amide bond[19]. In addition a naturally occurring UV-activatable wybutine base (Y base) is situated near the anticodon portion of the tRNA. These unique features allow the bifunctional tRNA to form covalent bonds to both a coding mRNA and a polypeptide during translation of some initial random mRNA library thereby connecting a protein molecule to its encoding mRNA (Figure 3). Briefly, during the course of translation, a bifunctional tRNA enters the A site of the ribosome in response to the presence of its codon at the 3'-termini of the mRNA's in the library. The bifunctional tRNA then accepts the growing polypeptide from tRNA in the P site of the ribosome and is translocated to the P site. Due to the stable amide-linkage between the bifunctional tRNA

and its own amino acid, further transfer of the nascent polypeptide to incoming tRNA in the A site does not take place . Instead, translation halts and the bifunctional tRNA remains bound to the ribosome, linked via a non-hydrolysable amide bond to the nascent polypeptide. The resultant tRNA-peptide complex is then UV crosslinked in cis to the coding mRNA in the ribosome via a Y base located adjacent to the anticodon sequence . The resultant ternary complex represents the desired covalent cis linkage between an information-bearing, amplifiable mRNA molecule and a functional, selectable polypeptide using a bifunctional tRNA to link the two components.

The central component of our scheme is a bifunctional $tRNA^{phe}$ that covalently links a polypeptide to its encoding mRNA during translation. All tRNAs end in a consensus CCA sequence at their 3'-termini. The 3'-hydroxyl of the final adenosine residue in the 3'-CCA region is normally linked to the appropriate amino acid residue through a labile ester bond. We make use of purified truncated $tRNA^{phe}$ that lacks the usual terminal adenosine. I synthesized a modified adenosine nucleoside triphosphate[20,21] (Figure 4) where the usual 3'-OH is replaced with a 3'-NH_2 group capable of forming an amide bond to phenylalanine. This 3'-NH_2, 3'-deoxy adenosine triphosphate was then added to the truncated $tRNA^{phe}$ in a nucleotidyl transferase reaction to form 3'-NH_2- $tRNA^{phe}$. The final desired bifunctional tRNA containing an amide-linked phenylalanine, herein referred to as PHE-N-$tRNA^{phe}$ or $tRNA^{x}$, is formed in an aminoacyl-transferase reaction using phenylalanine and 3'-NH_2- $tRNA^{phe}$ as substrates as described in Chapter III.

Experimental Considerations & Pilot Experiments

The above translation scheme hints at some fundamental considerations that we addressed in developing this system. The presence of stop and phenylalanine codons at positions 5' of the poly(U) region will prematurely terminate synthesis of the nascent polypeptide through a normal termination mechanism or formation of a stable amide-linked polypeptide-$tRNA^{x}$ complex that cannot participate as a donor substrate for translation of the remainder of the full length mRNA. This is particularly vexing as the probability *P* of a given codon *C* occurring at random in a sequence of length *n* codons is given by $P(C)= 1 - (63-64)^n$ which rapidly approaches unity as *n* increases. The rather large absolute size of the resultant libraries tends to alleviate such considerations. Thus, if even 90% of the mRNA in an initial library of $>10^{12}$

molecules contains at least one of the five aforementioned undesirable codons, 10%, or $>10^{11}$ of the mRNA molecules in the library will still contain a full length open reading frame that codes for selectable molecules. In addition, for randomized regions containing less than forty codons, the mRNA sequences can be reliably biased by synthesizing sequences using base mixtures that yield some desired probability distribution of amino acids in an open reading frame[22]. This forty codon upper limit is the result of the 0.8% frequency for insertions and deletions at each position that can occur in the original DNA synthesis, which tends to alter the desired distribution through frame shifts. For longer sequences, it is possible to construct libraries of open reading frames that are translated through the full length mRNA by selecting based on some constant motif, such as a 6-His tag, placed at the end of the randomized coding sequence[23]. In this instance, only library members that contain open reading frames will translate through to this 6-His tag and be selected, and these ORFs can then be linked together to form longer open reading frames. Additionally, the probability of $tRNA^{x}$ binding to internal UUU or UUC degenerate phenylalanine codons can be controlled by varying the ratio of $tRNA^{x}$ and normal PHE -$tRNA^{phe}$. Desirable low ratios can be compensated for by increasing the length of the poly(U) region thereby increasing the probability with which low concentration $tRNA^{x}$ is recruited to the ribosome.

The stability of mRNA molecules in the initial pool is an additional consideration given that mRNA levels are controlled in prokaryotes through the use of highly robust nucleases that rapidly degrade mRNA. Nuclease degradation of the mRNA tag during translation will make it impossible to PCR amplify selected sequences. Previous work in the literature had shown that the presence of a stem-loop structure at the terminal 5'-end of an mRNA greatly increases mRNA half-lives *in vivo*[24]. Introduction of such a stem-loop failed to stabilize our mRNA i*n vitro*, as evidenced by the extremely short half-life of radiolabelled mRNA in the translation system (too short to measure). A second type of short stem-loop structure placed at the 5'-end of a DNA sequence had been previously shown to be highly stable to nucleases *in vitro*[25]. We designed a DNA oligo called adDNA1 that contains this second type of short stem-loop structure placed after a region that is complementary to the 3' end of the mRNA sequences in our library. We found that annealing of this oligo (40 uM) to the 3' end of the mRNAs (10 uM) in our library prior to addition of the translation system yielded a greatly increased half-life of approximately

twenty minutes (Figure 5). A full sixty minute half-life was achieved by increasing the amount of adDNA1 added (80 uM). This half-life is quite long in comparison to the time it takes to complete a typical *in vitro* translation reaction (approximately ten minutes).

We synthesized a large amount of 3'-NH_2, 3'-deoxy adenosine triphosphate[20,21] (Figure 4) and used it to replace the 3'-terminal adenosine of a Y-base containing tRNA. The resultant bifunctional $tRNA^x$ contains a stable amide-linked amino acid and a photoactivatable cross-linker. We then tested whether this $tRNA^x$ could be stably crosslinked to a short message containing just a start codon followed by a single phenylalanine codon (Figure 6). Complexes of 30S and 50S ribosomal subunits were allowed to assemble on short AUGUUU RNAs in the presence of radiolabelled ^{35}S methionine initiator tRNA and $tRNA^x$. The bifunctional $tRNA^x$ accepts the labeled methionine from the initiator tRNA and the resultant tRNA-peptide complex is stable to base hydrolysis. In the presence of UV irradiation at 320 nm, an additional band corresponding to the tRNA-peptide-mRNA complex is observed above the tRNA-peptide band due to $tRNA^x$ crosslinking to the AUGUUU mRNA. This confirmed that the synthesized bifunctional $tRNA^x$ contains the two necessary moieties for encoding libraries of peptides with their respective mRNAs. Subsequent experiments using longer mRNAs and complete translation systems showed that libraries of protein-$tRNA^x$-mRNA fusions can be generated (Chapter III). These libraries can then be used in selections to do *in vitro* protein evolution. Future efforts in this regard should be focused on increasing the cross-linking efficiency perhaps using different photoactivatable bases. It will also be interesting to attempt selections from complex libraries for functionally interesting molecules.

References

1. Street, A. G. & Mayo, S. L. Computational protein design. *Structure Fold Des* **7**, R105-109 (1999).
2. Darwin, C. *The Origin of Species* (Norton, New York, 1975).
3. Holland, J. H. *Adaptation in natural and artificial systems : an introductory analysis with applications to biology, control, and artificial intelligence.* (MIT Press, Cambridge, 1992).
4. Forrest, S. Genetic algorithms: principles of natural selection applied to computation. *Science* **261**, 872-878 (1993).
5. Yao, X. Evolving artificial neural networks. *Proceedings of the IEEE.* **87**, 1423-1447 (1999).
6. Szostak, J. W. Ribozymes. Evolution ex vivo. *Nature* **361**, 119-120 (1993).
7. Arnold, F. H. & Volkov, A. A. Directed evolution of biocatalysts. *Curr Opin Chem Biol* **3**, 54-59 (1999).
8. Schultz, P. G. Bringing biological solutions to chemical problems. *Proc Natl Acad Sci U S A* **95**, 14590-14591 (1998).
9. Bartel, D. P. & Szostak, J. W. Isolation of new ribozymes from a large pool of random sequences [see comment]. *Science* **261**, 1411-1418 (1993).
10. Ekland, E. H., Szostak, J. W. & Bartel, D. P. Structurally complex and highly active RNA ligases derived from random RNA sequences. *Science* **269**, 364-370 (1995).
11. Cadwell, R. C. & Joyce, G. F. Randomization of genes by PCR mutagenesis. *PCR Methods Appl* **2**, 28-33 (1992).
12. Stemmer, W. P. DNA shuffling by random fragmentation and reassembly: in vitro recombination for molecular evolution. *Proc Natl Acad Sci U S A* **91**, 10747-10751 (1994).
13. Shao, Z., Zhao, H., Giver, L. & Arnold, F. H. Random-priming in vitro recombination: an effective tool for directed evolution. *Nucleic Acids Res* **26**, 681-683 (1998).
14. Bornscheuer, U. T. & Pohl, M. Improved biocatalysts by directed evolution and rational protein design. *Curr Opin Chem Biol* **5**, 137-143 (2001).
15. Cull, M. G., Miller, J. F. & Schatz, P. J. Screening for receptor ligands using large libraries of peptides linked to the C terminus of the lac repressor. *Proc Natl Acad Sci U S A* **89**, 1865-1869 (1992).
16. Hanes, J. & Pluckthun, A. In vitro selection and evolution of functional proteins by using ribosome display. *Proc Natl Acad Sci U S A* **94**, 4937-4942 (1997).
17. Mattheakis, L. C., Bhatt, R. R. & Dower, W. J. An in vitro polysome display system for identifying ligands from very large peptide libraries. *Proc Natl Acad Sci U S A* **91**, 9022-9026 (1994).
18. Roberts, R. W. & Szostak, J. W. RNA-peptide fusions for the in vitro selection of peptides and proteins. *Proc Natl Acad Sci U S A* **94**, 12297-12302 (1997).
19. Fraser, T. H. & Rich, A. Synthesis and aminoacylation of 3'-amino-3'-deoxy transfer RNA and its activity in ribosomal protein synthesis. *Proc Natl Acad Sci U S A* **70**, 2671-2675 (1973).
20. Robins, M. J., Hawrelak, S. D., Hernandez, A. E. & Wnuk, S. F. Nucleic Acid Related Compounds. 71. Efficient general synthesis of purine (amino, azido, and triflate)-sugar nucleoside. *Nucloes, Nucloet.* **11**, 821-834 (1992).
21. Ludwig, J. A New Route to Nucleoside 5'-triphosphates. *Acta Biochim et Biophys Acad Sci Hung* **16**, 133-133 (1981).

22. Wolf, E. & Kim, P. S. Combinatorial codons: a computer program to approximate amino acid probabilities with biased nucleotide usage. *Protein Sci* **8**, 680-688 (1999).
23. Cho, G., Keefe, A. D., Liu, R., Wilson, D. S. & Szostak, J. W. Constructing high complexity synthetic libraries of long ORFs using in vitro selection. *J Mol Biol* **297**, 309-319. (2000).
24. Arnold, T. E., Yu, J. & Belasco, J. G. mRNA stabilization by the ompA 5' untranslated region: two protective elements hinder distinct pathways for mRNA degradation. *Rna* **4**, 319-330 (1998).
25. Varani, G., Cheong, C. & Tinoco, I., Jr. Structure of an unusually stable RNA hairpin. *Biochemistry* **30**, 3280-3289 (1991).

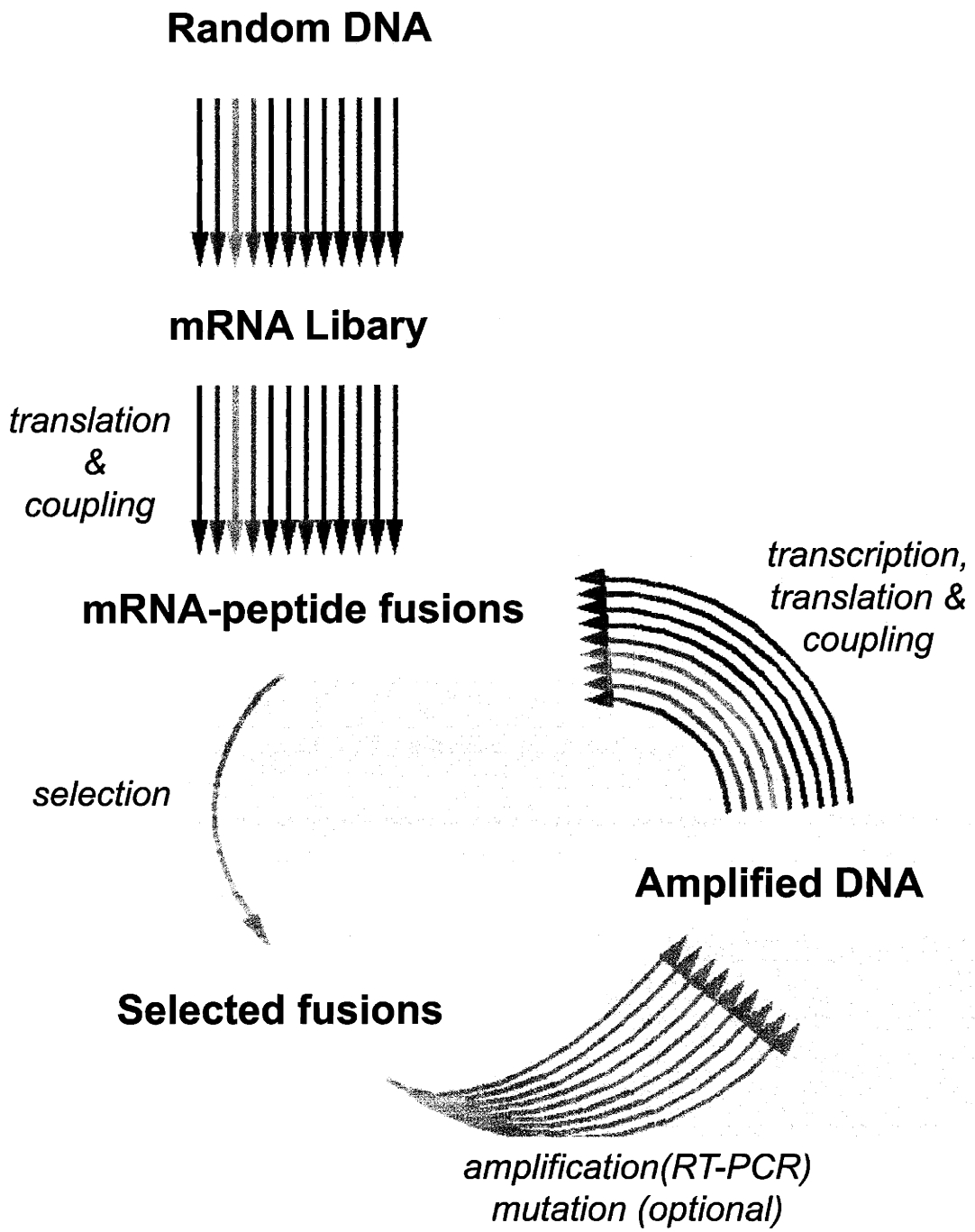


Figure 1. Schematic of a protein directed evolution experiment.
An initial high-complexity library of DNA with randomized open reading frames is transcribed into mRNA. Coding mRNAs are then translated and linked in cis to their respective peptides via the photoactivatable bifunctional tRNA. Selective pressure for desired peptide molecules is applied and the mRNA linked to selected peptides is amplified and optionally randomized using standard reverse-transcription polymerase chain reaction methods (RT-PCR). Selected cDNAs are then used as input to the next round of selection. This process is iterated until a desired functionality is reached.

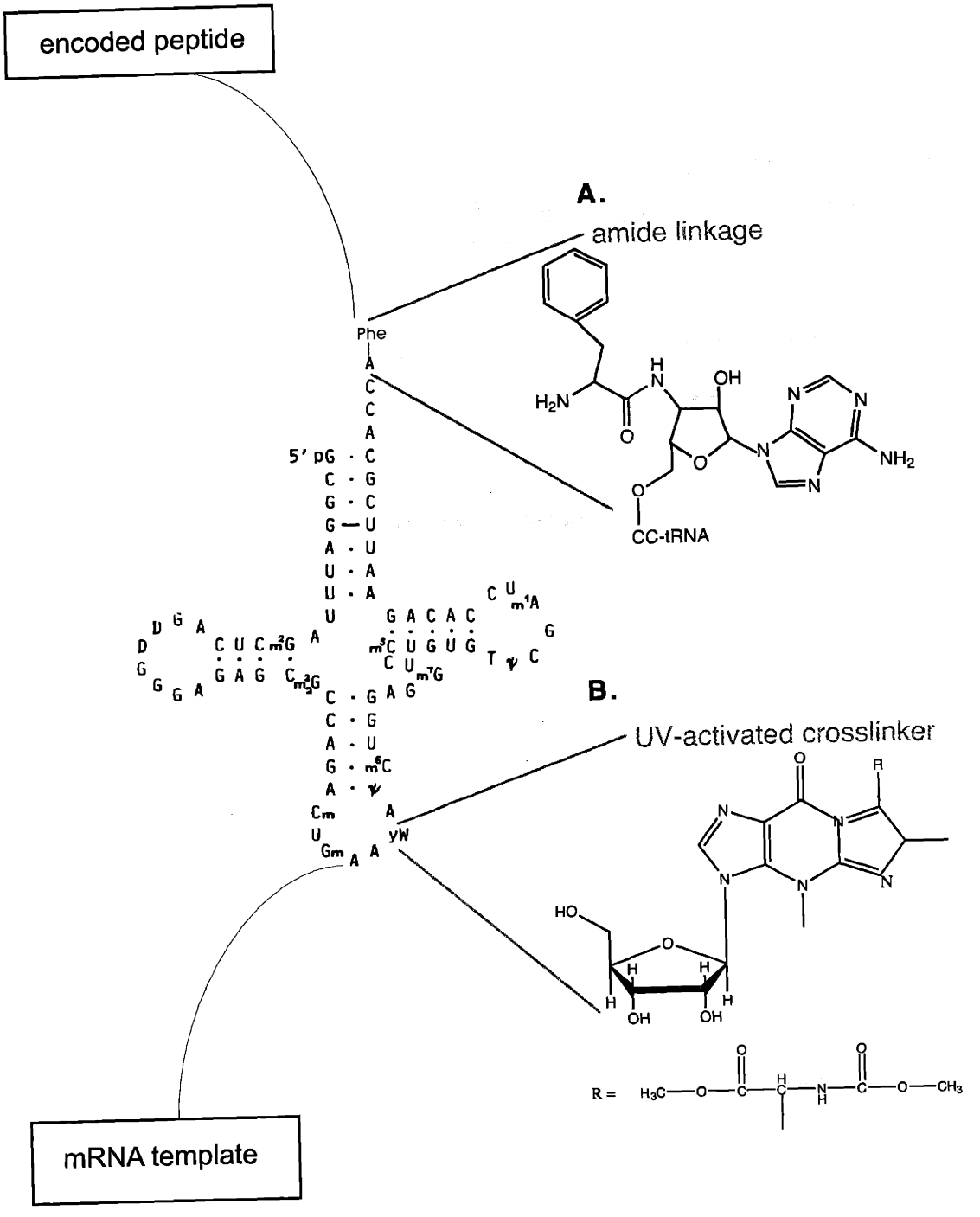


Figure 2. Structure of the photoactivatable bifunctional PHE-N-tRNA.
PHE-N-tRNA is used to link a peptide in cis to its encoding mRNA template.
(A.) At the three prime end, a stable amide linkage (shown in red) replaces the normal base-labile ester linkage to phenylalanine.
(B.) A naturally occurring wybutine base (Y base) near the anticodon loop comprises the photoactivatable crosslinking moiety used to link the tRNA/peptide to the encoding mRNA template.

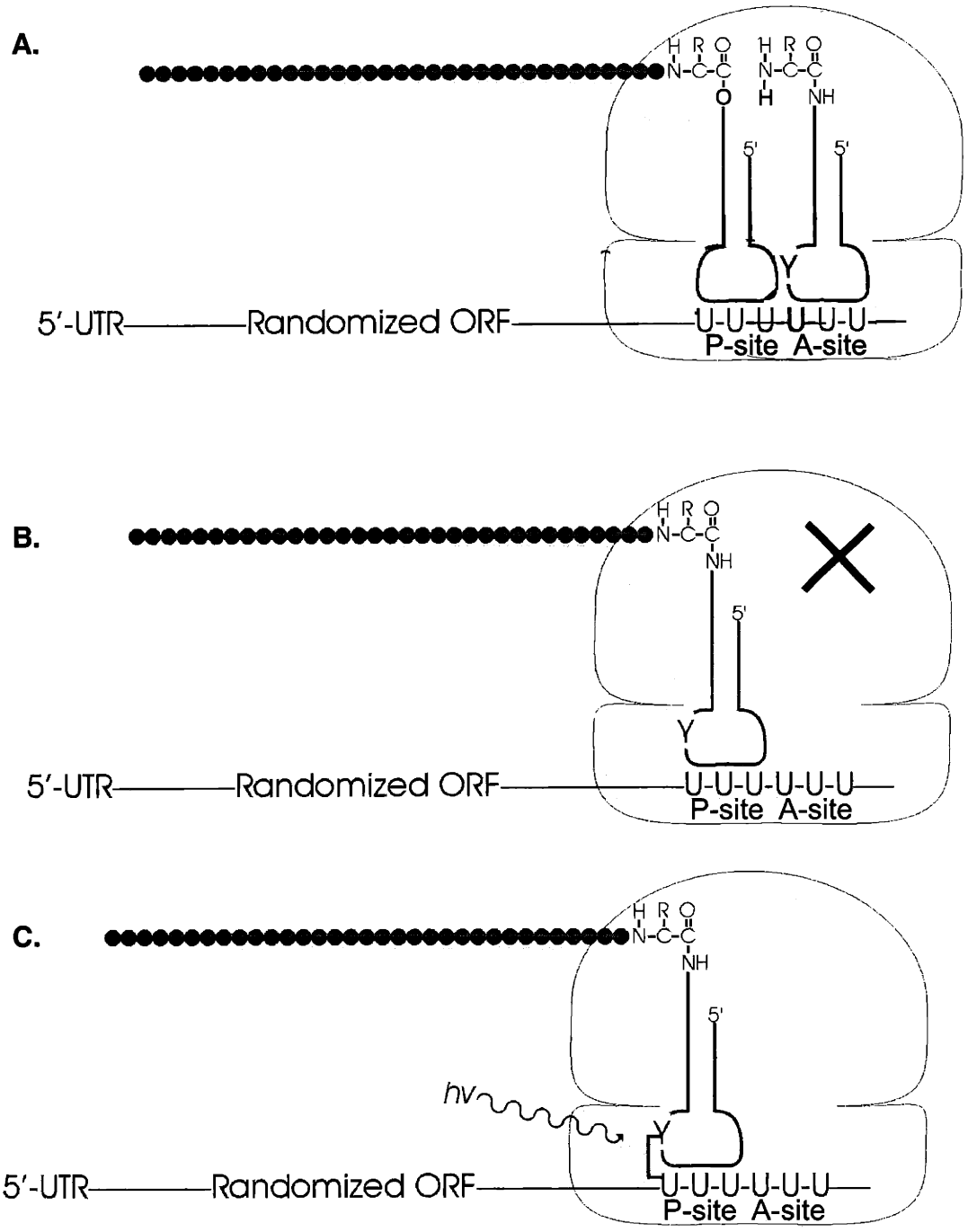


Figure 3. Translation steps using the bifunctional tRNA.
(A) Bifunctional tRNA enters the A site and accepts the peptide from tRNA in the P site.
(B) The amide-linked tRNAphe does not function as a donor tRNA and translation pauses.
(C) Bifunctional tRNAphe linked to the peptide is UV-crosslinked in cis to the encoding mRNA sequence.

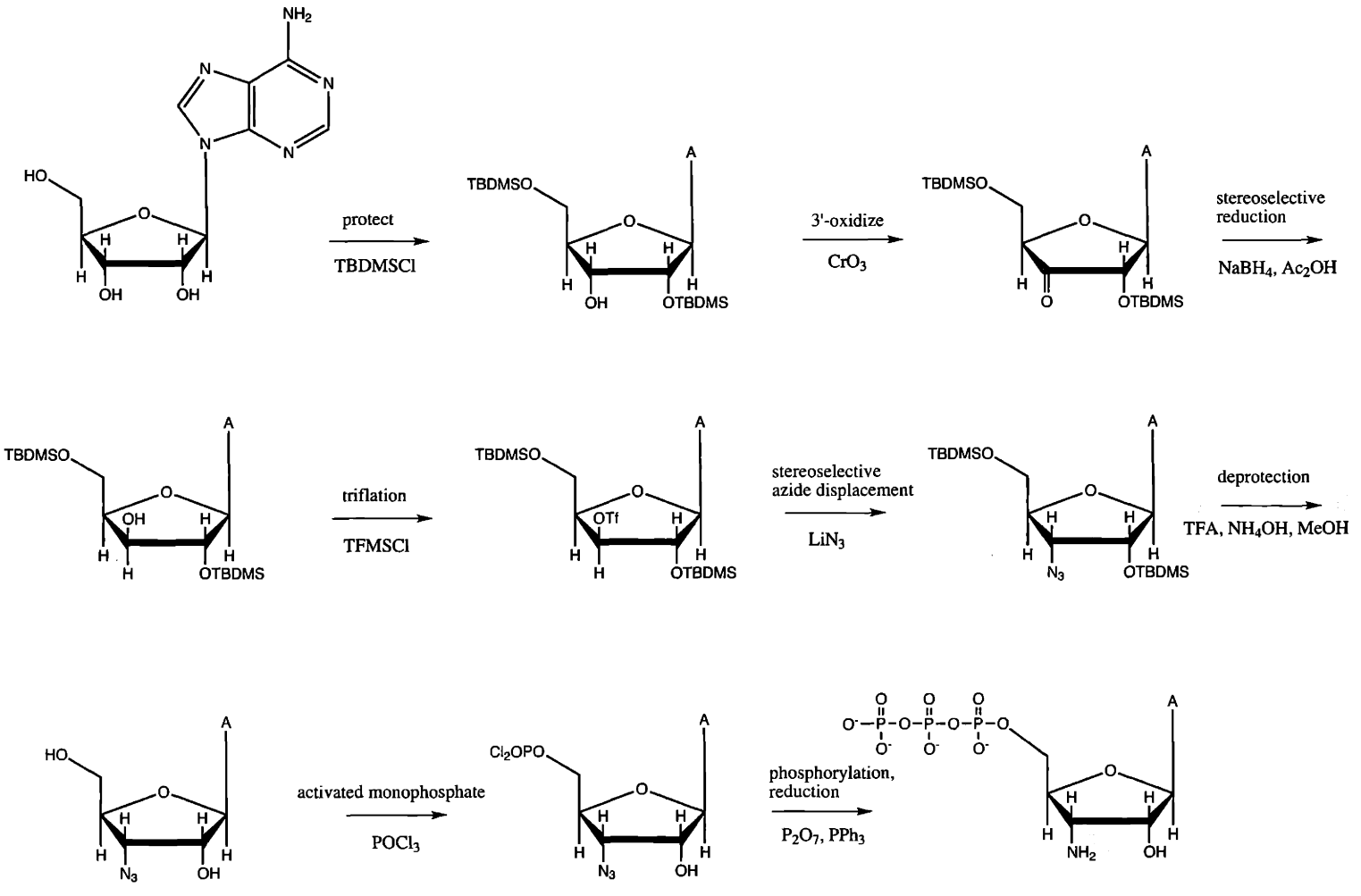


Figure 4. Synthetic scheme for 3'-NH_2 adenosine triphosphate.

Adenosine is used as starting material and protected at the 5' and 2' positions. Oxidation at the 3' OH, followed by stereoselective reduction and triflate displacement is used to introduce an azide group. The deprotected nucleoside is then phosphorylated and reduced to form the desired 3' amine group.

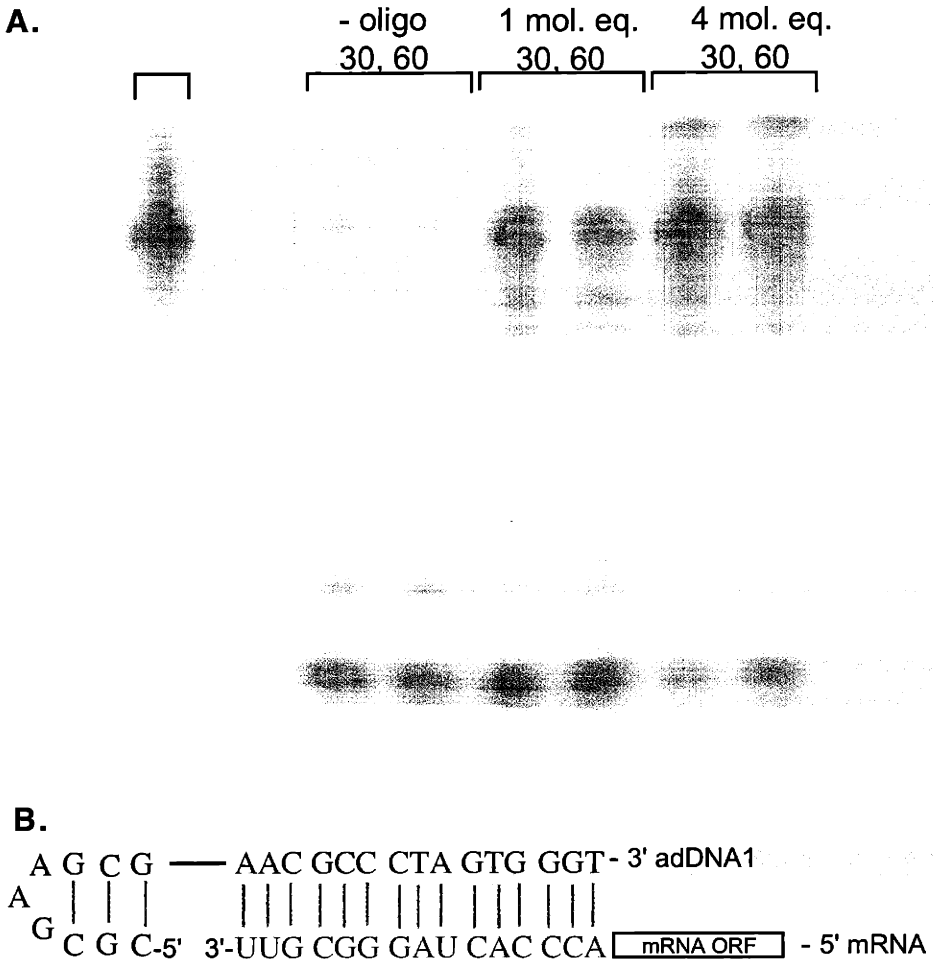


Figure 5. mRNA library stabilization using a structured oligonucleotide.
(*A*) Degradation of radiolabelled mRNA after 30 and 60 minutes in a translation reaction. In the absence of the adDNA1 oligo (-oligo), the mRNA is rapidly degraded into small fragments (lower bands).
(*B*) The adDNA1 oligo bound to the 3' end of an mRNA. The adDNA1 priming region (red) is complementary to the 3' end of the mRNA (green). The 5' end of the adDNA1 (blue) forms a nuclease-resistant secondary structure.

	1	2	3	4
AUGUUU mRNA	+	+	+	+
30s, 50s subunits	+	+	+	+
PHE-N-tRNA	-	-	+	+
UV	-	+	-	+

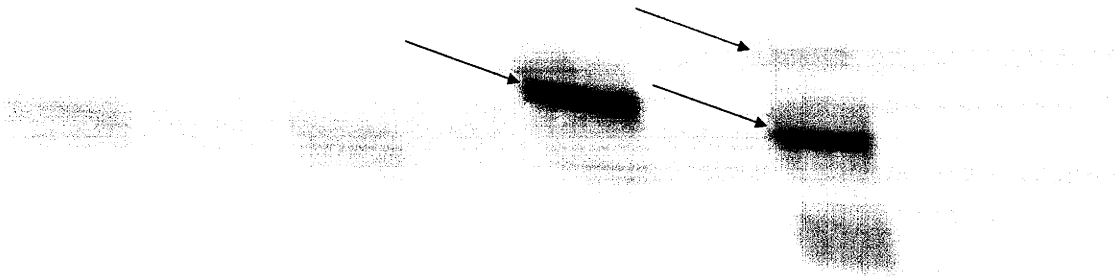


Figure 6. Crosslinking AUGUUU mRNA to 35S-MET-PHE peptides.
No stable tRNA/peptide molecules are observed in the absence of PHE-N-tRNA (lanes 1-2).
A band corresponding to a stable tRNA-peptide (blue arrow) is observed in the presence of PHE-N-tRNA (lanes 3-4).
A super shifted band (red arrow) is observed in the presence of UV and corresponds to the desired tRNA-mRNA-peptide complex (lane 4).

A Bifunctional tRNA for *In Vitro* Selection

The work presented in this chapter was a collaborative effort between myself and Chuck Merryman. Specifically, I synthesized the 3'-NH2 adenosine triphosphate, used this nucleoside to synthesize the bifunctional tRNA, developed and carried out experiments to show that the bifunctional tRNA was used in translation and specifically linked to the encoded peptide, developed the methods for stabilizing mRNA molecules in the library during translation, and carried out a proof of principle selection for a six-his peptide.

ABSTRACT: *In vitro* selection is a powerful approach for generating novel aptamers and catalysts. Currently, several methods are being developed to extend this technique to proteins. In principal, selection methods could be applied to any library whose members can be replicated. Here, we describe a bifunctional tRNA that fuses translation products to their mRNAs. The utility of peptide-tRNA-mRNA fusions for *in vitro* selection was illustrated by the selective enrichment of tagged peptides—together with their mRNAs—by affinity chromatography. Our system can generate libraries larger than 10^{11}. Because library members can be copied and amplified, they provide a means for applying *in vitro* selection procedures to peptides and proteins. Furthermore, because the system is amenable to translation with misacylated tRNAs, a wide range of unusual monomers could be used to make libraries of non-standard polymers for selection experiments.

Introduction

In vitro selection is regularly used to search libraries of nucleic acids for rare molecules with desirable functions [see refs. 1-4 for reviews]. Molecules with specific functions are isolated from libraries with more than 10^{15} sequences through iterative rounds of selection and amplification; molecules that fulfill the selective criteria increase in representation, and amplification increases their number. Thus, with each round, the functional molecules replace less capable members of the initial population. Although nucleic acids are endowed with recognition and catalytic potential, the application of *in vitro* selection to polymers with greater chemical diversity would be beneficial. Toward this end, several methods have been developed to attach peptides and proteins to their encoding DNAs or mRNAs [5-9]. Such fusions contain the essential elements needed for selection and amplification: a potentially active protein and a corresponding nucleic acid sequence that stores the information needed to make copies of the protein. In phage display [5] and plasmid display [6], fusion proteins that associate with their encoding nucleic acids are expressed *in vivo*; the cell membrane encapsulates corresponding protein and nucleic acid sequences during complex formation. Ribosome display [7] relies on the integrity of stalled translation complexes to maintain a link between an mRNA and its protein product. With mRNA display [8, 9], covalent protein-mRNA fusions are formed on stalled translation complexes; the 3'-ends of mRNAs are modified in such a way that they stall protein synthesis, enter the A site of the ribosome, and act like an aminoacyl tRNA to become attached to their protein products.

We have been developing an alternative method for generating libraries of mRNA-encoded peptides, which also has the potential to work with other peptide-like polymers. In our system, a modified tRNA ($tRNA^x$) acts as a bifunctional crosslinking agent to attach an mRNA to its translation product. For an *in vitro* selection experiment to succeed, a molecule with some degree of the desired activity must reside in the initial population, and thus library complexity is critical. Completely *in vitro* methods like ours, ribosome display [7] and mRNA display [8, 9] have an advantage over methods that require the use of cells because library complexity is not limited by transformation efficiency (bacterial transformation currently limits libraries to ~10^9 [10]). In our system as with mRNA display, physically linking a peptide to its mRNA by covalent bonds simplifies the purification of peptide-mRNA fusions and increases the range of selectable properties because the integrity of the peptide-nucleic acid complex is not susceptible

to disruption. Optimized bacterial systems are capable of translating about 10% of input mRNAs into protein [11]. Ultimately, this could increase library complexity and simplify selections by decreasing the interference from free mRNA. Other advantages of bacterial systems involve the ability to add and subtract specific components. For example, release factors could be removed and suppressor tRNAs supplied to obviate the need of removing stop codons from mRNAs [12]. Our system, as with ribosome display, uses *E. coli* extracts that could take advantage of these features. Here, we show that the bifunctional tRNA can be used to generate 10^{11} covalent peptide-$tRNA^{x}$-mRNA fusions (PTM fusions) per milliliter of translation reaction and that the fusions can be used to selectively enrich and amplify peptide libraries.

Results

Design and Synthesis of the Bifunctional tRNA.

We designed a tRNA that can be used to form a stable linkage between polypeptides and the mRNA that encodes them. Within the ribosome, the growing peptide chain is normally linked to a tRNA by a labile ester bond. In turn, the tRNA is transiently linked to the mRNA that encodes the peptide by the base pairs of the codon-anticodon interaction. If the 3'-terminal adenosine of an aminoacyl tRNA is replaced by 3'-amino-3'-deoxyadenosine, the labile ester bond is replaced by a stable amide bond [13]. Similarly, crosslinking the wybutine base (Y base) of yeast $tRNA^{phe}$ to the mRNA within the ribosome complex covalently joins the tRNA and mRNA [14]. Thus, a tRNA containing both an amide linkage to its amino acid and a Y base (Fig. 1*A*, "$tRNA^{x}$") could be useful for crosslinking proteins to their encoding mRNAs. Yeast $tRNA^{phe}$, which contains the Y base, was purified by benzoylated DEAE cellulose chromatography, and tRNA missing its 3'-terminal adenosine was repaired with tRNA nucleotidyl transferase and 3'-amino-3'-deoxyadenosine triphosphate (Fig. 1*B*, lanes 1-4). Several methods exist for removing the 3'-terminal adenosine of a tRNA, but with yeast $tRNA^{phe}$ this is generally unnecessary because the nucleotide is lost during purification [15]. Once repaired, the modified tRNA contained an intact acceptor stem with a 3'-terminal amine, and it contained the naturally occurring Y base in the anticodon loop. When the tRNA is aminoacylated with phenylalanine, the amino acid migrates to the 3'-amine forming the desired amide bond [13]. The tRNA purification, repair and aminoacylation steps were all efficient (Fig. 1*B*), and it was not necessary to purify the intermediates or product.

Synthesis of PTM fusions.

PTM fusions were made by translating synthetic mRNAs in the presence of $tRNA^{x}$ and subsequently irradiating the translation products with UV light. In the translation mix, protein synthesis proceeded normally until a phenylalanine codon (Phe codon) in the mRNA reached the A site of the ribosome. At this point, either $tRNA^{x}$ or *E. coli* $tRNA^{phe}$ could be incorporated. If $tRNA^{x}$ was selected, it was attached to the translated peptide chain (formyl-[^{35}S]-MKDYKDDDDK) (Fig. 2). No fusion products were formed if the mRNA did not code for $tRNA^{x}$ (Fig. 2, lane 1). If normal $tRNA^{phe}$ was selected at a Phe codon, translation continued, as shown by the production of multiple peptide-tRNA fusions (PT fusions) with an mRNA that

contained multiple Phe codons (Fig. 2, lane 2). Because the number of PT fusions produced was always equal to the number of in-frame Phe codons in the mRNA, the process was likely the result of normal translation (Fig. 2, lanes 2-5). Furthermore, the mobilities of PT fusions in acrylamide gels were consistent with the size and charge of the polypeptide encoded by the translated mRNA (Fig 2, lanes 2-5).

Once linked to the peptide chain, $tRNA^x$ stalls protein synthesis because the ribosome cannot break the amide bond that connects it to its amino acid [13]. The stalled ribosomal complexes were stable on sucrose gradients (data not shown). Therefore, high-salt sucrose cushions were used to purify ribosomal complexes from free mRNA and ribonucleases. The ribosome-bound PT fusions were then linked to their mRNAs by UV irradiation (Fig. 3*A*), which crosslinks the wybutine base in $tRNA^x$ to the 5'-U of the Phe codon [16] (Fig. 3*B*). As expected from a process that links a peptide to its mRNA, the product detected when the peptide was labeled comigrated with the product detected when mRNA was labeled (Fig. 3*A*, lanes 4 and 5, respectively). Controls showed that crosslinking required Phe codons in the mRNA to recruit $tRNA^x$ (Fig. 3*A*, lanes 1 and 8), inclusion of $tRNA^x$ in the translation mix (Fig. 3*A*, lanes 2 and 7), and exposure to UV light (Fig. 3*A*, lanes 3 and 6). Quantitation of the PT and PTM fusion bands (Fig. 3*A*, lane 4) indicated that about 0.2% of the mRNA was decoded by $tRNA^x$ and that 1% of the PT fusions formed PTM fusions. Although the bulk of input mRNA is degraded by contaminating ribonucleases, 10^{11} PTM fusions were made in a 1-ml translation reaction. These results open the prospect of using $tRNA^x$ to make complex pools of mRNA-encoded polypeptides.

Enrichment of Mixed Populations by Peptide Selection.

Mock *in vitro* selections were performed to show that PTM fusions could be used to enrich RNA sequences that encode peptides with specified properties. In the first experiment, a synthetic mRNA coding for a Cys-containing peptide and an mRNA coding for 6 consecutive histidines (poly-His) were mixed at a ratio of 1:10. The mRNA mixture was added to a translation reaction that contained $tRNA^x$, and translated peptides were crosslinked to their encoding mRNAs. PTM fusions were then removed from the ribosome and partially purified by urea-LiCl precipitation. To selectively isolate PTM fusions that contained Cys, the purified translation reaction was subjected to thiol-affinity chromatography (Fig.4*A*). Quantitation of the

band intensities from the initial (Fig.4*A*, lane 1) and selected populations (Fig.4*A*, lane 3), indicated that the initially under-represented Cys fusion was enriched about 15-fold. To control for inadvertent skewing of the makeup of the mixed population by mechanisms other than peptide selection, we performed the inverse experiment; the mRNA ratio was switched and used to generate a second population of PTM fusions which were subjected to metal-affinity chromatography. Again, the initially under-represented fusion—in this case the one containing poly-His—was enriched, but by about 5-fold (Fig.4*A*, lanes 4 and 6).

To show that peptide selection was reflected at the genetic level, the mRNAs contained in the initial and enriched populations were subjected to RT-PCR and compared (Fig. 4*B*). When thiol-affinity chromatography was used as the selective step, the PCR product encoding Cys was enriched (Fig. 4*B*, lanes 1-3), whereas when metal-affinity chromatography was used, the PCR product encoding poly-His was enriched (Fig. 4*B*, lanes 4-6). Thus, the intended selective step drives the evolution of the nucleic acid sequences that encode PTM fusions; if another mechanism was dominant—such as preferential RT-PCR amplification of a specific template—the same species would have overtaken both populations. Because RT-PCR products could be used to produce a new population of fusions, selection and amplification could be repeated to provide exponential enrichment of target molecules.

Discussion

By fusing a peptide to its mRNA, $tRNA^x$ linked corresponding functional and replicable sequences in a single molecule. Two simple libraries of mRNAs were translated and fused to their peptide products, and the mRNAs coding for the selected peptide were amplified. These results indicate that the system can be used for the *in vitro* selection of peptides and proteins from complex libraries.

Library production requires the translation of mRNA pools that contain randomized coding regions. Phe codons within the randomized region could recruit $tRNA^x$ early, which would produce truncated peptides. However, selection of a normal $tRNA^{phe}$ at a Phe codon allows translation to proceed. Thus, by adding $tRNA^x$ at a low effective concentration and placing a large number of Phe codons after the randomized region, most fusions will be formed near the end of an mRNA. Another concern is the presence of stop codons in the randomized region. Of the existing methods for dealing with this problem, perhaps the easiest is to use translation mixes that contain suppressor tRNAs but no release factors [17, 18]. Suppressor tRNAs that are chemically misacylated would have the added advantage of allowing the introduction of unnatural amino acids and other monomers [19, 20].

In a selection experiment, a large population is critical because it increases the likelihood that desirable molecules are represented. Currently, we can make 10^{11} fusions in a 1-ml translation reaction, which already surpasses the complexity achieved by *in vivo* methods. Furthermore, up to a 1000-fold improvement in fusion efficiency might be possible; if fully realized, a 1-ml reaction would yield 10^{14} PTM fusions. The bulk of this anticipated increase comes from improving crosslinking efficiency and increased utilization of the mRNA. For example, only 0.2 percent of the mRNA was translated and decoded by $tRNA^x$ in our experiments, whereas in experiments using more highly purified bacterial translation systems over 10 percent of the mRNA is utilized for protein synthesis, perhaps because purified systems have less ribonuclease contamination [11]. If still larger libraries are desired, bacterial translation extracts can be scaled up without undue expense.

In conjunction with a more highly purified translation system, our method might offer advantages for constructing libraries synthesized from unusual monomers. Although all of the systems have the potential to incorporate unnatural amino acids by nonsense suppression, incorporating unusual monomers at sense codons is difficult in most other systems. The cognate

tRNAs would need to be specifically eliminated, perhaps by use of antisense oligonucleotides, and these tRNAs would need to be replaced with "orthogonal" tRNAs that are designed to avoid editing or charging by aminoacyl-tRNA synthetases [20]. With our system, such measures would not be necessary or would be more easily accomplished because bacterial translation systems are more readily customized. For example, fusions bearing non-standard polymers have been generated in translation mixes that use misacylated tRNA, EFG and EFTu rather than total tRNA and S150 (data not shown). In principal, ribosome display and mRNA display could use similar translation systems. However, mRNA display has not been shown to work with bacterial ribosomes, and ribosome display requires the translation of much longer peptides, as over 40 residues must be translated before the peptide begins to emerge from the exit channel of the ribosome [7]. With PTM fusions, even short open reading frames can satisfy the requirements of complexity and accessibility. Thus, because the ribosome can use hundreds of monomers [e.g., 21-24], it could be possible to build low-molecular-weight libraries that have desirable properties such as protease resistance, permeability, and conformational rigidity. In conjunction with *in vitro* selection methods, such libraries could open the door to a vast array of useful molecules that could serve as leads for the development of therapeutics and other useful reagents.

Significance

We anticipate that the flexibility of our system with respect to the types of polymers that could be produced will broaden the number of applications to which *in vitro* selection can be applied. In principal, fusion libraries that are larger than 10^{11} could be constructed from any combination of monomers that the ribosome can polymerize. This large complexity has the potential to generate rare-functional molecules, while the wide range of acceptable monomers could be used to adjust the overall physical and chemical properties of a library. Thus, the method could be useful for evolving non-biological polymers whose properties depart from those accessible to peptides and proteins.

Materials and Methods

Purification of Ribosomes and S150 Enzyme Fraction.

For ribosomes, 5 g of an *E. coli* ribonuclease-deficient strain (A19) were washed with 300 ml of buffer A (10 mM Tris-HCl, pH 7.5, 10 mM Mg-acetate, 22 mM NH_4Cl, 1 mM DTT), pelleted in a Sorvall SLA-3000 rotor (4600 g), and suspended in buffer A in a final volume of 20 ml. The suspension was lysed in a BeadBeater mixer (Biospec) according to the manufacturers directions with 80 ml of 0.1 mm zirconia/silica beads. The beads were washed several times with buffer A and the supernatants combined and transferred to 13 ml centrifuge tubes. The lysate was cleared by repeated 20 min centrifugations in a Sorvall SS-34 rotor (17,000 g). Ribosomes were isolated by layering 13 ml of the clarified supernatant on 13 ml of 32% sucrose in buffer A and centrifuging for 13 hours in a Beckman 70Ti rotor (120,000 g). Pellets were dissolved in a small volume of buffer A, and the concentration was adjusted to 45 μM before storage at -80°C in 5-25 μl aliquots. For the S150 enzyme fraction, 4 g of cells were washed in 38 ml of buffer B (10 mM Tris-HCl, pH 7.5, 10 mM $MgCl_2$, 30 mM NH_4Cl, 6 mM BME) and lysed as before but with buffer B. The lysate was clarified twice by centrifugation for 20 min in a Sorvall SS-34 rotor (30,000 g) and once for 30 min in a Beckman VTi50 rotor (150,000 g). The clarified supernatant was loaded on a 40 ml DEAE sepharose column (Pharmacia) that had been equilibrated with buffer B, and the column was washed with 1 L of buffer B. The S150 enzyme fraction was eluted with buffer B plus 220 mM NH_4Cl. Fractions were examined by eye, and were pooled and stored at -80°C in 20-100 μl aliquots.

Synthesis of the Bi-functional $tRNA^x$.

Yeast $tRNA^{phe}$ was purified by benzoylated DEAE cellulose chromatography [25]. 3'-Amino-3'-deoxyadenosine triphosphate was prepared by published protocols [26, 27]. The construct pQECCA, expressing *E. coli* tRNA nucleotidyl transferase as a fusion protein with a poly-His tag, was a generous gift from U. RajBhandary (MIT). Yeast $tRNA^{phe}$ missing its 3'-terminal adenosine was repaired by incubating 500 μM tRNA, 2 mM 3'-amino-3'-deoxyadenosine triphosphate, and tRNA nucleotidyl transferase in buffer (50 mM Tris-HCl, pH 8.0, 10 mM $MgCl_2$, 30 mM KCl, 5 mM DTT, 0.5 mg/ml BSA) for 10 min at 37°C. Protein was removed by phenol extraction, and the tRNA was ethanol precipitated. The intermediate (3'-amino-$tRNA^{phe}$) was charged by incubating 25 μM 3'-amino-$tRNA^{phe}$, 100 μM phenylalanine, 4

mM dATP, with 1/5 (vol/vol) S150 enzyme fraction for 30 min at 37° C in 30 mM Tris-HCl, pH 7.5, 15 mM $MgCl_2$, 25 mM KCl, and 5 mM DTT. The final product ($tRNA^x$) was purified by phenol extraction and ethanol precipitated.

Synthesis of mRNA.

DNA templates for T7 *in vitro* transcription were generated by PCR, using appropriate templates and primers. PCR products were ethanol precipitated and transcribed in half their original volume (40 mM Tris-HCl, pH 7.9, 26 mM $MgCl_2$, 2.5 mM spermidine, 0.01% triton X-100, 5 mM ATP, 5 mM CTP, 8 mM GTP, 2 mM UTP, and T7 RNA polymerase). All mRNAs (Table 1) were gel-purified. *5.4.*

Synthesis of PT Fusions.

Three μg of mRNA was translated in 40 μl of an *E. coli* S30 extract (Promega) according to the manufacturers protocol. In addition to exogenous mRNA, reactions contained [^{35}S]-methionine and 2 μM $tRNA^x$. After incubating for 30 min at 37°C, the reactions were terminated by phenol extraction and ethanol precipitated. Amino acids and peptides bound to normal tRNAs were removed by incubation in 0.5 M Ches-KOH, pH 9.5 for 1 hr at 37°C, followed by phenol extraction and ethanol precipitation.

Synthesis of PTM Fusions.

For each reaction, translation premix contained 2.5 μl of 10X translation buffer (500 mM Tris-acetate, pH 8.0, 110 mM Mg-acetate, 1 M NH_4Cl, 10 mM DTT); 2.5 μl of 100 mM phosphoenol pyruvate; 0.5 μl of 100 mM ATP:10 mM GTP; 0.5 μl of 25 μg/μl total *E. coli* tRNA; 2.5 μl of 1 mM amino acids; 0.5 μl of 1 μg/μl pyruvate kinase; 3 μl of 18 μM ribosomes; and 6 μl of S150 enzyme fraction. Premix was incubated for 10 min at 37°C after which 2.5 μl of 200 μM $tRNA^x$ was added when appropriate. Aliquots (16 μl) of the premix were distributed among tubes that contained 2 μl of [^{35}S]-methionine labeled fMet-$tRNA^{Met}$ and 2 μl of 50 μM mRNA. When the mRNA rather than the peptide was labeled, the tubes contained 2 μl of fMet-$tRNA^{Met}$ and 2 μl of 50 μM [^{32}P]-cordycepin-labeled mRNA. Incubation was continued for another 10 min and samples layered onto 500 μl sucrose cushions (32% sucrose, 50 mM Hepes, pH 7.8, 20 mM $MgCl_2$, 500 mM NH_4Cl, 6 mM BME). Stalled ribosomal complexes were

pelleted by centrifugation for 45 min at 4°C in a Beckman TLA 100.2 rotor (360,000 g). Pellets were dissolved in 25 µl of buffer (40 mM Hepes, pH 7.8, 20 mM $MgCl_2$, 80 mM NH_4Cl, 6 mM BME), spotted on a polystyrene petri-dish, covered with the supplied lid and exposed to UV for 20 min using a 450 watt Hanovia bulb with water jacket (Ace Glass). As before, phenol extraction and base treatment were used to remove undesired translation products.

In Vitro Peptide Selection.

PTM fusions were synthesized as described above with minor modifications. Briefly, [^{35}S]-methionine labeled fMet-tRNA^{Met} was generated *in situ* by adding 1/10 (vol/vol) deprotected 5,10-methenyltetrahydrofolate [28] and 1/10 (vol/vol) [^{35}S]-methionine; the amino acid mix did not contain methionine. To make room for the additional reagents, half as much amino acid mix was added, concentrated stocks of nucleotide mix and ribosomes were used, and the final concentration of tRNA^x was reduced to 10 µM. Rather than individual mRNAs, mixtures were used to direct translation; mRNA encoding the target peptide was at 0.4 µM and mRNA encoding the background peptide was at 4 µM. After crosslinking, ribosomal complexes were mixed 1:1 with lithium buffer (8 M urea, 4 M LiCl, 10 mM EDTA, pH 8.0, 6 mM BME) and precipitated overnight at 4°C. Pellets were dissolved in 20 mM Tris-HCl, pH 7.5, 2 mM $MgCl_2$, 60 mM KCl, 6 mM BME and stored at -20°C. Before selection, aliquots were precipitated from ethanol and dried by aspiration to remove BME, then dissolved in 10 µl of water. Poly-His-containing fusions from a 90 µl translation reaction were bound to Talon metal-affinity resin (Clontech) by adding 9 µl of the concentrated product to a resin slurry (100 µl resin : 100 µl 50 mM Tris-HCl, pH 7.5, 300 mM NH_4Cl, 0.25 mg/ml BSA) and rotating for 16 hr at 4°C. Unbound fusions were removed by washing the resin 5 times with 100 µl aliquots of wash buffer (50 mM Tris-HCl, pH 7.5, 1 M NH_4Cl). Bound fusions were eluted by washing the resin 3 times with 100 µl of wash buffer plus 10 mM EDTA, pH 8.0, and 8.3 µg of total *E. coli* tRNA. Cys-containing fusions were immobilized in a similar manner with 500 µl of activated Thiol-Sepharose 4B (Pharmacia) in 5 ml of buffer (25 mM Tris-HCl, pH 7.5, 300 mM NaCl, 7M urea). After binding, the resin was poured into a column and washed with 100 ml of the same buffer. The resin was then removed and eluted 3 times with 1 ml aliquots of buffer that also contained 50 mM DTT and 25 µg/ml total *E. coli* tRNA. Column washes and eluants were ethanol precipitated and dissolved in 90 µl or 9 µl of water, respectively. For both selections,

before RT-PCR, unfused mRNA was removed from the enriched population of PTM fusions by gel purification; a liberal section of the gel was excised to insure that gel purification did not influence the ratio of the two PTM fusions.

References

1. Gold, L., Polisky, B., Uhlenbeck, O. and Yarus, M. (1995). Diversity of oligonucleotide functions. Annu. Rev. Biochem. 64, 763-797.
2. Lorsch, J.R. and Szostak, J. W. (1996). Chance and necessity in the selection of nucleic acid catalysts. Acc. Chem. Res. 29, 103-110.
3. Kurz, M. and Breaker, R. R. (1999). In vitro selection of nucleic acid enzymes. Curr. Top. Microbiol. Immunol. 243, 137-158
4. Bartel D. P. and Unrau P. J. (1999). Constructing an RNA world. Trends Cell Biol. 9, M9-M13.
5. Smith, G. P. (1985). Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. Science 228, 1315-1317.
6. Cull, M. G., Miller, J. F. and Schatz, P. J. (1992). Screening for receptor ligands using large libraries of peptides linked to the C terminus of the lac repressor. Proc. Natl. Acad. Sci. USA. 89, 1865-1869
7. Mattheakis, L. C., Bhatt, R. R., and Dower, W. J. (1994). An in vitro polysome display system for identifying ligands from very large peptide libraries. Proc. Natl. Acad. Sci. USA. 91, 9022-9026.
8. Nemoto, N., Miyamoto-Sato, E., Husimi, Y. and Yanagawa, H. (1997). In vitro virus: bonding of mRNA bearing puromycin at the 3'-terminal end to the C-terminal end of its encoded protein on the ribosome in vitro. FEBS Letters. 414, 405-408.
9. Roberts, R. W. and Szostak, J. W. (1997). RNA-peptide fusions for the in vitro selection of peptides and proteins. Proc. Natl. Acad. Sci. USA. 94, 12297-12302.
10. Noren, K. A. and Noren, C. J. (2001). Construction of high-complexity combinatorial phage display peptide libraries. Methods 23, 169-178.
11. Pavlov, M. Y. and Ehrenberg, M. (1996). Rate of translation of natural mRNAs in an optimized in vitro system. Arch. Biochem. Biophys. 328, 9-16.
12. Cho, G., Keefe, A. D., Liu, R., Wilson, D. S. and Szostak, J. W. (2000). Constructing high complexity synthetic libraries of long ORFs using in vitro selection. J. Mol. Biol. 297, 309-319
13. Fraser, T. H. and Rich, A. (1973). Synthesis and aminoacylation of 3'-amino-3'-deoxy transfer RNA and its activity in ribosomal protein synthesis. Proc. Natl. Acad. Sci. USA. 70, 2671-2675.
14. Matzke, A. J., Barta, A.and Kuechler, E. (1980). Mechanism of translocation: relative arrangement of tRNA and mRNA on the ribosome. Proc. Natl. Acad. Sci. USA. 77, 5110-5114.
15. RajBhandary, U. L., Stuart, A., Hoskinson, R. M. and Khorana, H. G. (1968). Studies on polynucleotides. 78. Yeast phenylalanine transfer ribonucleic acid: terminal sequences. J. Biol. Chem. 243, 565-574.
16. Steiner, G., Luhrmann, R.. and Kuechler, E. (1984). Crosslinking transfer RNA and messenger RNA at the ribosomal decoding region: identification of the site of reaction on the messenger RNA. Nucleic Acids Res. 12, 181-191.
17. Short, G. F., Golovine, S. Y. and Hecht, S. M. (1999). Effects of release factor 1 on in vitro protein translation and the elaboration of proteins containing unnatural amino acids. Biochemistry. 38, 8808-8819.

18. Shimizu, Y., Inoue, A., Tomari, Y., Suzuki, T., Yokogawa, T., Nishikawa, K. and Ueda, T. (2001). Cell-free translation reconstituted with purified components. Nat. Biotechnol. 19, 751-755.
19. Heckler, T. G., Chang, L. H., Zama, Y., Naka, T., Chorghade, M. S. and Hecht, S. M. (1984). T4 RNA ligase mediated preparation of novel "chemically misacylated" tRNAPheS. Biochemistry. 23, 1468-1473.
20. Noren, C. J., Anthony-Cahill, S. J., Griffith, M. C. and Schultz, P. G. (1989). A general method for site-specific incorporation of unnatural amino acids into proteins. Science 244, 182-188.
21. Bain, J. D., Wacker, D. A., Kuo, E. E. and Chamberlin, A. R. (1991). Site-specific incorporation of nonnatural residues into peptides - effect of residue structure on suppression and translation efficiencies Tetrahedron 47, 2389-2400.
22. Mendel, D., Ellman, J. and Schultz, P. G. (1993). Protein-biosynthesis with conformationally restricted amino-acids. J. Am. Chem. Soc. 115, 4359-4360.
23. Thorson, J. S., Cornish, V. W., Barrett, J. E., Cload, S. T., Yano, T. and Schultz, P. G. (1998). A biosynthetic approach for the incorporation of unnatural amino acids into proteins. Meth. Mol. Biol. 77, 43-73.
24. Hohsaka, T., Kajihara, D., Ashizuka, Y., Murakami, H. and Sisido, M. (1999). Efficient incorporation of nonnatural amino acids with large aromatic groups into streptavidin in in vitro protein synthesizing systems. J. Am. Chem. Soc. 121, 34-40.
25. Gillam, I., Millward, S., Blew, D., von Tigerstrom, M., Wimmer, E., and Tener, G. M. (1967). The separation of soluble ribonucleic acids on benzoylated diethylaminoethylcellulose. Biochemistry. 6, 3043-3056.
26. Robins, M. J., Hawrelak, S. D., Hernandez, A. E. and Wnuk, S. F. (1992). Nucleic-acid related-compounds. 71. Efficient general-synthesis of purine (amino, azido, and triflate)-sugar nucleosides. Nucleosides and Nucleotides. 11, 821-834.
27. Morr, M. and Wray, V. (1994). New cyclic-derivatives of 3'-amino-3'-deoxyadenosine-5'-diphosphate, 3'-amino-3'-deoxyadenosine-5'-triphosphate, and 3'-amino-3'-deoxyadenosine-5'-methylenebis(phosphonate). Angew. Chem. Int. Ed. 33, 1394-1396.
28. Blanquet, S., Dessen, P. and Kahn, D. (1984). Properties and specificity of methionyl-tRNAfMet formyltransferase from Escherichia coli. Meth. in Enz. 106, 141-152.
29. Matzke, A. J., Barta, A. and Kuechler, E. (1980). Photo-induced crosslinking between phenylalanine transfer RNA and messenger RNA on the Escherichia coli ribosome. Eur. J. Biochem. 112, 169-178.

Table I. Sequences of the mRNAs used to direct *in vitro* protein synthesis.

mRNA	Sequence
1	GGAUCCUAGGAAGCUUGAAGGAGAUAUACC**AUG** AAA GAC UAC AAG GAC GAC GAC GAC AAG UAU AAA GUU...
2	GGAUCCUAGGAAGCUUGAAGGAGAUAUACC**AUG** AAA GAC UAC AAG GAC GAC GAC GAC AAG **UUU UUU UUU**...
3	GGAUCCUAGGAAGCUUGAAGGAGAUAUACC**AUG** AAA GAC UAC AAG GAC GAC GAC GAC AAG **UUU** AAA GUU...
4	GGAUCCUAGGAAGCUUGAAGGAGAUAUACC**AUG** AAA GAC UAC AAG GAC GAC GAC GAC AAG UAU **UUU** GUU...
5	GGAUCCUAGGAAGCUUGAAGGAGAUAUACC**AUG** AAA GAC UAC AAG GAC GAC GAC GAC AAG UAU AAA **UUU**...
6	GGGUUAACUUUAGAAGGAGGUAAAAAAA**AUG** AAA CGU GAA AAG ACA **UUU UUU UUU**
7	GGGUUAACUUUAGAAGGAGGUAAAAAAA**AUG** AAA CGU GAA AAG ACA GAA CGU ACA
8	GGGUUAACUUUAGAAGGAGGUAAAAAAA**AUG** AAA CAC CAU CAC CAC CAU CAC GGA AAU CGU **UUU UUC UUU UUC UUU UUC** CGC UAG CGU CAG GGC UAU UCA CCA UUA ACC CAC UAG GGC GUU
9	GGGUUAACUUUAGAAGGAGGUAAAAAAA**AUG** CGU UGC GAU CAC GGA AAU CGU **UUU UUC UUU UUC UUU UUC** CGC UAG CGU CAG GGC UAU UCA CCA UUA ACC CAC UAG GGC GUU
10	GGGUUAACUUUAGAAGGAGGUAAAAAAA**AUG UUU** AAA GAA AAG UUU GAA CGU ACA

For each mRNA, the codons are underlined and the initiator methionine codon (AUG) and Phe codons (UUU or UUC) are in bold. For mRNAs 1-5, part of the sequence is not shown (...). The remaining sequence is free of phenylalanine codons and is the same as nucleotides 217-444 of beta-lactamase.

Figures

Fig. 1. Design and synthesis of $tRNA^x$. (A) Schematic highlighting the important components of $tRNA^x$. The nitrogen that replaces the 3'-terminal oxygen of yeast $tRNA^{phe}$ is boxed. The amide bond between the tRNA and phenylalanine is shown as a thick line, and the naturally occurring wybutine base at position 37 of yeast $tRNA^{phe}$ is labeled (Y). (B) Synthesis of $tRNA^x$. RNAs were resolved on a denaturing 6% polyacrylamide gel (PAGE) and visualized with SYBR-gold: lane 1, total yeast tRNA; lane 2, yeast $tRNA^{phe}$; lane 3, purified ΔA-yeast $tRNA^{phe}$; lane 4, ΔA-yeast $tRNA^{phe}$ repaired with 3'-amino-3'-deoxyadenosine triphosphate; lane 5, repaired yeast $tRNA^{phe}$ aminoacylated with phenylalanine ($tRNA^x$). From top to bottom the three bands in lane 2 correspond to the full-length tRNA and truncated forms missing the 3'-terminal A and CA.

Fig. 2. Attachment of $tRNA^x$ to growing peptide chains. Denaturing 6% PAGE was used to separate the products of *in vitro* translation reactions that contained [^{35}S]-methionine and $tRNA^x$. Amino acids and peptides bound to normal tRNAs were removed by base treatment and phenol extraction. Five different mRNAs were used to direct translation in an S30 extract (mRNAs 1-5, lanes 1-5, respectively). Each of the mRNAs coded for the same leader peptide (MKDYKDDDDK). Only amino acids specified at codons 11, 12, and 13 differed, as indicated above each lane.

Fig. 3. PTM fusions. (A) Denaturing 4% PAGE of the products of *in vitro* translation reactions. Translation reactions contained [^{35}S]-methionine (lanes 1-4) or ^{32}P-labeled mRNA (lanes 5-8). In lanes 1 and 8, translation was directed by an mRNA that contained no Phe codons (mRNA 6), whereas the mRNA in lanes 2-7 contained Phe codons at positions 7, 8, and 9 (mRNA 7). Transfer RNA^x was not added to lanes 2 and 7. Lanes 3 and 6 were not exposed to UV. The mobilities of peptide-$tRNA^x$ (PT) and peptide-$tRNA^x$-mRNA (PTM) fusions are indicated. The band below the PT fusion in lane 4 probably resulted from UV dependent degradation of the Y base [27]. In the absence of translation, mRNA 7 ran as two separate bands which could also account for the two PTM fusion products seen in lanes 4 and 5. (B) Denaturing 6% PAGE of the primer-extension stops for reverse transcriptase on mRNA 10 that was crosslinked to $tRNA^x$. A, C, G and U are dideoxy-sequencing lanes using mRNA as the template. Lane 1, mRNA; lane

2, crosslinking reaction; lane 3, gel-purified PTM fusions from the crosslinking reaction. The 5'-UTR of the mRNA and codons 1-4 (MFKE) are indicated.

Fig. 4. *In vitro* selection of mRNA-encoded peptides. (A) Denaturing 4% PAGE of peptide-$tRNA^x$-mRNA fusions following *in vitro* selection. Two mixtures of mRNA 8 (Cys mRNA) and mRNA 9 (poly-His mRNA) were translated in the presence of $tRNA^x$. In mixture A (lanes 1-3), the Cys:poly-His mRNA ratio was 1:10. In mixture B (lanes 4-6), the mRNA ratio was inverted. Following translation, PTM fusions were formed and partially purified. The products of mixture A were selected by thiol-affinity chromatography and the products of mixture B by metal-affinity chromatography. Lanes 1 and 4, PTM fusions from 1 μl of translation; lanes 2 and 5, column wash from 1 μl of translation; lanes 3 and 6, column eluant from 10 μl of translation. (B) Non-denaturing 8% PAGE analysis of the RT-PCR products produced from mixture A (lanes 1-3) or B (lanes 4-6) following translation, fusion and *in vitro* selection by thiol affinity or metal affinity chromatography, respectively. Lanes 1 and 4, starting mRNA mixtures without reverse transcriptase; lanes 2 and 5, starting mRNA mixtures; lanes 3 and 6, translated, fused and selected mRNA mixtures. In both panels, the mobilities of the Cys- and poly-His-containing PTM fusions or their RT-PCR products are indicated (C and H, respectively).

Figure 1

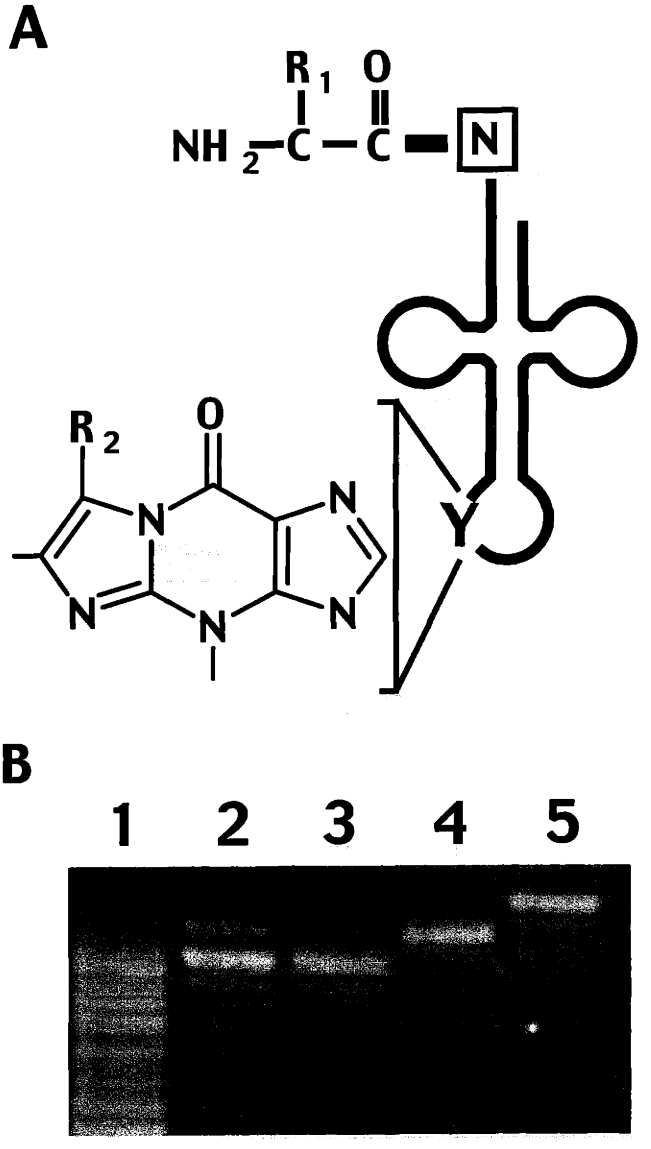


Figure 2

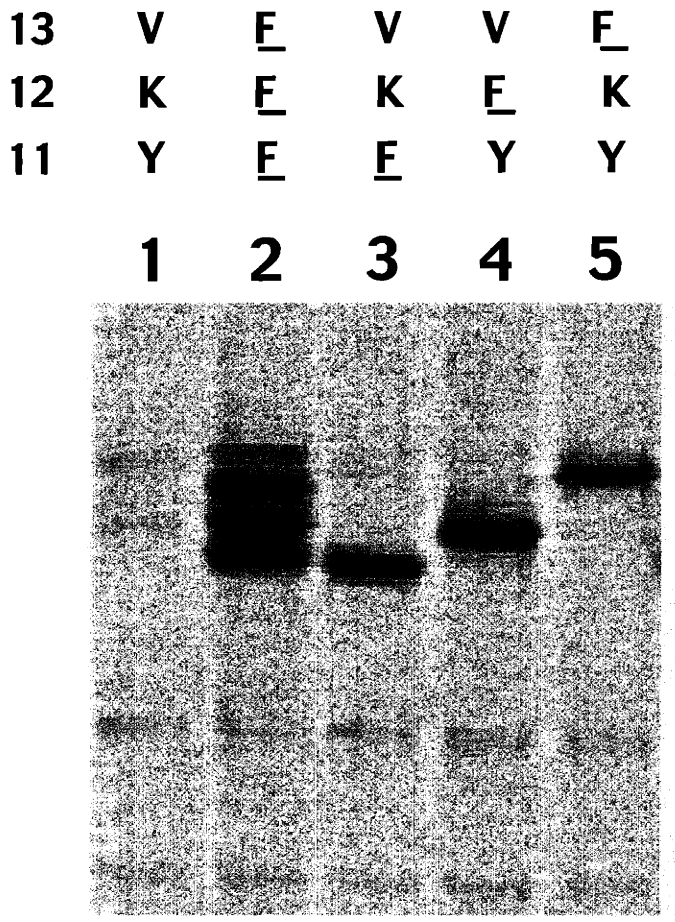


Figure 3

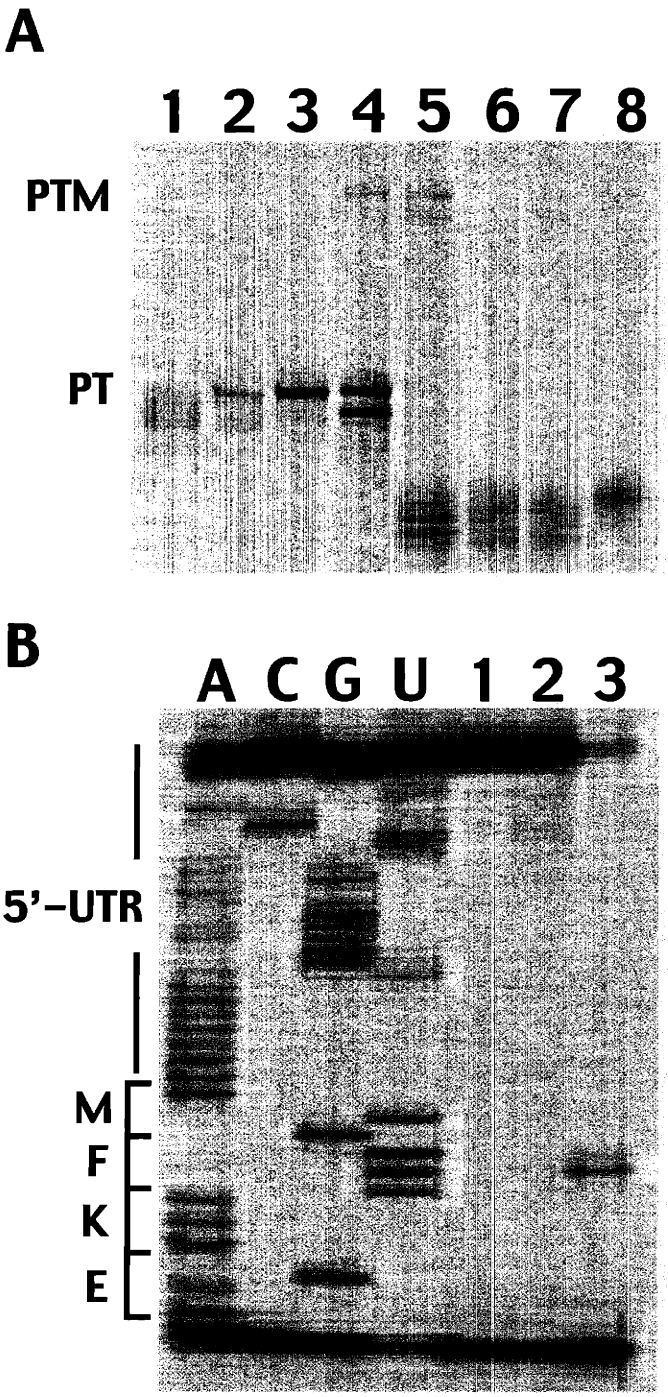
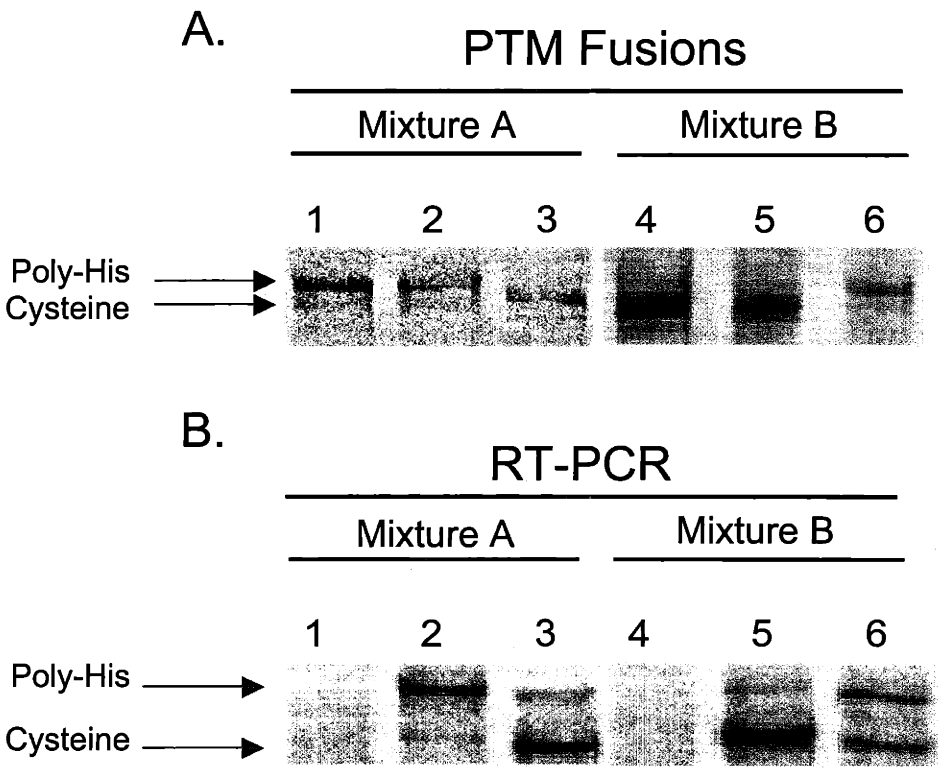


Figure 4


A.
PTM Fusions
Mixture A
Mixture B
1
2
3
4
5
6
Poly-His
Cysteine
B.
RT-PCR
Mixture A
Mixture B
1
2
3
4
5
6
Poly-His
Cysteine

miRNA gene	miRNA sequence	Length	Number of clones	Fold-back arm	Chr.	Begin	End	Nearest gene	Comments
mir-50	CAGUCCGCAC AAUUGUCAAC CU	20-25	26	5'	I	1738651	1738672	in intron of Y71G12B.11a	uRNA in an intron
D1_E02-3	AACGGCAAUU GUCGCUCAUU GUAU	23-25	5	5'	I	4599230	4599207	1.6 kb from end of U04D1.2, antisense	uRNA
N1_B06-5	AAAUCGAAAU UGUUGUGUUG AA	22-22	22	3'	I	6085222	6085201	0.7 kb from start of U09B4.7	uRNA
mir-1	AACCUCUCUA AAUGACUCUC U	20-22	70	3'	I	6095592	6095572	3.7 kb from start of U09B4.3, antisense	uRNA
*mir-1**	CUUUCCGAUC CUCUCUCACC GU	22-23	2	5'	I	6095632	6095611	3.6 kb from start of U09B4.3, antisense	uRNA*
mir-79	AAAUUAAAAU AAUCACCGUA AA	21-22	19	3'	I	9255893	9255914	2.3 kb from end of C12C8.2	uRNA
mir-2	AUCCAAUAAA AUAAUAAUUA AAA	22-24	173	3'	I	9295692	9295670	0.6 kb from start of M04C9.6a	uRNA
*mir-2**	GGAAUCGAGG CGUUACUUCA AG	22-22	1	5'	I	9295733	9295712	0.7 kb from start of M04C9.6a	uRNA*
*mir-71**	AUUCACAUUG CAUUCCAACG CC	21-22	3	3'	I	9302855	9302834	7.8 kb from start of M04C9.6a	uRNA*
mir-71	UAAUUCAUGU CUUGAUACG	18-24	122	5'	I	9302892	9302874	7.8 kb from start of M04C9.6a	uRNA
mir-72	AGGCAAGAUG UUGGCAUAGC UGA	19-24	64	3'	II	2452852	2452874	0.2 kb from end of F53G2.4, antisense	uRNA
lin-4	UCCCUGAGAC CUCAAGUGUG A	19-22	88	5'	II	5902232	5902252	in coding sequence of F59G1.6	uRNA?
mir-60	UAUUAUGCAC AUUUUCUAGU UCA	22-24	28	3'	II	6328684	6328662	1.5 kb from end of C32D5.5	uRNA
N2_B02-14	UAAUACUGUC AGGUAAUGAC GCU	21-24	6	3'	II	7030144	7030122	0.3 kb from end of C52E12.1, antisense	uRNA
mir-57	UACCCUGUAG AUCGAGCUGU GUGU	20-25	45	5'	II	7850498	7850475	0.9 kb from start of AF187012-1.U09A5	uRNA
mir-85	UACAAAGUAU UUGAAAAGUC GUGC	24-25	20	3'	II	8393532	8393555	in coding sequence of F49E12.8, antisense	uRNA?
mir-35	UCACCGGGUG GAAACUAGCA GU	21-22	22	3'	II	11537564	11537585	1.3 kb from end of F54D5.12, antisense	uRNA
mir-36	UCACCGGGUG AAAAUUCGCA UG	21-23	28	3'	II	11537669	11537690	1.2 kb from end of F54D5.12, antisense	uRNA
mir-37	UCACCGGGUG AACACUUGCA GU	22-22	8	3'	II	11537789	11537810	1.1 kb from end of F54D5.12, antisense	uRNA
mir-38	UCACCGGGAG AAAAACUGGA GU	21-22	10	3'	II	11537886	11537907	1.0 kb from end of F54D5.12, antisense	uRNA
mir-39	UCACCGGGUG UAAAUCAGCU UG	22-22	14	3'	II	11538039	11538060	0.8 kb from end of F54D5.12, antisense	uRNA
mir-40	UCACCGGGUG UACAUCAGCU AA	21-23	18	3'	II	11538135	11538156	0.7 kb from end of F54D5.12, antisense	uRNA
mir-41	UCACCGGGUG AAAAAUCACC UA	21-23	3	3'	II	11538264	11538285	0.6 kb from end of F54D5.12, antisense	uRNA
mir-45	UGACUAGAGA CACAUUCAGC U	20-22	28	3'	II	11880873	11880853	0.7 kb from end of K12D12.1, antisense	uRNA
*mir-45**	CUGGAUGUGC UCGUUAGUCA UA	22-22	1	5'	II	11880919	11880898	0.7 kb from end of K12D12.1, antisense	uRNA*
mir-42	UCACCGGGUU AACAUCUACA GA	20-22	17	3'	II	11889764	11889785	1.2 kb from end of ZK930.2, antisense	uRNA
mir-43	UAUCACAGUU UACUUGCUGU CGC	22-23	12	3'	II	11889864	11889886	1.1 kb from end of ZK930.2, antisense	uRNA
mir-44	UGACUAGAGA CACAUUCAGC U	20-22	28	3'	II	11889977	11889997	1.0 kb from end of ZK930.2, antisense	uRNA
mir-77	UUCAUCAGGC CAUAGCUGUC CA	21-22	18	3'	II	12519222	12519243	1.5 kb from start of U21B4.9, antisense	uRNA
L1_G01-4	UUAUUGCUCG AGAAUACCCU U	21-21	1	3'	II	14461944	14461924	1.6 kb from end of Y54G11B.1, antisense	uRNA
mir-64	UAUGACACUG AAGCGUUACC GAA	21-24	22	5'	III	2218821	2218843	0.2 kb from start of Y48G9A.1	uRNA
mir-65	UAUGACACUG AAGCGUAACC GAA	21-23	28	5'	III	2218971	2218993	0.1 kb from start of Y48G9A.1	uRNA

mir-66	CAUGACACUG AUUAGGGAUG UGA	18-25	94	5'	III	2219076	2219098	in coding sequence of Y48G9A.1	uRNA?
mir-76	UUCGUUGUUG AUGAAGCCUU GA	22-22	8	3'	III	3188059	3188080	3.0 kb from start of C44B11.3, antisense	uRNA
mir-67	UCACAACCUC CUAGAAAGAG UAGA	24-24	3	3'	III	5977395	5977372	4.4 kb from end of EGAP1.1	uRNA
D2_B02-2	UAAGCUCGUG AUCAACAGGC AGAA	23-24	3	3'	III	7591066	7591043	10.4 kb from start of C07H6.7	uRNA
D2_B08-2	UGGUGAGACA CGUCGUAACG AAU	23-23	1	5'	III	7906207	7906185	0.2 kb from end of ZK652.2, antisense	uRNA
mir-80	UGAGAUCAUU AGUUGAAAGC CGA	20-25	142	3'	III	8911259	8911237	4.7 kb from end of F44E2.2, antisense	uRNA
D2_F03-4	UUUGUACUCC GAUGCCAUUC AGA	23-23	2	3'	III	8913317	8913295	6.7 kb from end of F44E2.2, antisense	uRNA
mir-90	UGAUAUGUUG UUUGAAUGCC CCU	20-24	41	3'	III	8919975	8919953	9.0 kb from end of ZK637.1	uRNA
*mir-90**	CGGCUUUCAA CGACGAUAUC AAC	23-23	3	5'	III	8920021	8919999	8.9 kb from end of ZK637.1	uRNA*
mir-86	UAAGUGAAUG CUUUGCCACA GUC	22-24	112	5'	III	11982594	11982572	in intron of Y56A3A.7	uRNA in an intron
mir-46	UGUCAUGGAG UCGCUCUCUU CA	21-23	28	3'	III	13706089	13706110	3.0 kb from end of ZK525.1, antisense	uRNA
mir-58	UGAGAUCGUU CAGUACGGCA AU	19-24	206	3'	IV	3244918	3244939	in intron of Y67D8A.1	uRNA in an intron
L1_D07-2	UUGCGUAGGC CUUUGCUUCG A	21-21	1	5'	IV	4285862	4285882	0.9 kb from end of F36A4.14, antisense	uRNA
N3_D05-3	CGGUACGAUC GCGGCGGGAU AUC	23-23	1	3'	IV	4462439	4462461	1.0 kb from start of R08C7.1	uRNA
D1_A01-1	AAUGGCACUG CAUGAAUUCA CGG	23-24	16	5'	IV	5573632	5573654	0.2 kb from end of U12E12.5, antisense	uRNA
mir-83	UAGCACCAUA UAAAUUCAGU AA	20-23	32	3'	IV	7853279	7853300	5.0 kb from start of C06A6.2	uRNA
D1_F11-24	UAAAUGCAUC UUAACUGCGG UGA	23-23	10	3'	IV	11005780	11005758	1.2 kb from end of F13H10.5, antisense	uRNA
mir-51	UACCCGUAGC UCCUAUCCAU GUU	22-23	25	5'	IV	11037670	11037648	0.4 kb from end of F36H1.6, antisense	uRNA
mir-53	CACCCGUACA UUUGUUUCCG UGCU	21-26	26	5'	IV	11039250	11039227	1.9 kb from end of F36H1.6, antisense	uRNA
mir-59	UCGAAUCGUU UAUCAGGAUG AUG	23-23	1	3'	IV	11320766	11320744	1.8 kb from start of B0035.1a, antisense	uRNA
*N1_H09-5**	GCAUGCACCC UAGUGACUUU AGU	23-23	1	5'	IV	11883328	11883350	1.1 kb from end of C29E6.6	uRNA*
N1_H09-5	UAAGGCACGC GGUGAAUGCC A	21-21	26	3'	IV	11883371	11883391	1.2 kb from end of C29E6.6	uRNA
mir-52	GAGUUCUCAU CUGAUUUUGA AUUG	21-26	311	5'	IV	14045594	14045617	4.6 kb from end of Y37A1B.6, antisense	uRNA
mir-78	UAGAAUUUCG UUUUCCUUUA U	21-21	5	3'	IV	15177270	15177290	2.0 kb from start of Y40H7A.3, antisense	uRNA
mir-68	AUUUUCAACU GUUUCAUGUU G	21-21	1	3'	IV	16708152	16708172	3.3 kb from start of Y51H4A.22	uRNA
mir-70	UAAUACGUCG UUGGUGUUUC CAU	22-23	19	3'	V	6667160	6667138	in intron of U10H9.5	uRNA in an intron
mir-61	UGACUAGAAC CGUUACUCAU C	21-21	15	3'	V	11771676	11771656	0.4 kb from end of F55A11.3, antisense	uRNA
mir-48	UGAGGUAGGC UCAGUAGAUG CGA	21-24	50	5'	V	14366074	14366052	6.1 kb from start of Y49A3A.4	uRNA
N5_F02-7	UGAGGUAGGU GCGAGAAAUG A	21-21	8	5'	V	14367848	14367828	7.1 kb from start of F56A12.1, antisense	uRNA
D2_H10-2	UACACGUGCA CGGAUAACGC UCA	23-23	1	3'	X	2296712	2296690	in intron of AH9.3	uRNA in an intron
mir-73	UGGCAAGAUG UAGGCAGUUC AGU	23-23	19	3'	X	2403128	2403150	2.9 kb from start of U24D8.6, antisense	uRNA
mir-74	UGGCAAGAAA UGGCAGUCUA CA	21-24	51	3'	X	2403405	2403426	3.2 kb from start of U24D8.6, antisense	uRNA
mir-75	UUAAAGCUAC CAACCGGCUU CA	22-22	18	3'	X	2406816	2406837	3.5 kb from start of F47G3.3	uRNA

mir-81	UGAGAUCAUC GUGAAAGCUA GU	20-22	51	3'	X	2465441	2465462	in coding sequence of U07D1.2, antisense	uRNA?
mir-82	UGAGAUCAUC GUGAAAGCCA GU	21-23	52	3'	X	2469597	2469576	0.1 kb from start of U07D1.2	uRNA
*mir-34**	ACGGCUACCU UCACUGCCAC CC	21-22	8	3'	X	3004095	3004074	2.1 kb from end of Y41G9A.4, antisense	uRNA*
mir-34	AGGCAGUGUG GUUAGCUGGU UG	19-23	39	5'	X	3004132	3004111	2.1 kb from end of Y41G9A.4, antisense	uRNA
D1_F11-5	UCACAGGACU UUUGAGCGUU GC	22-22	1	3'	X	3040775	3040754	2.7 kb from start of Y41G9A.6	uRNA
H1_H05-5	GUAUUAGUUG UGCGACCAGG AGA	23-23	1	3'	X	5838220	5838242	0.4 kb from end of F13D11.3, antisense	uRNA
H1_G07-2	UACUGGCCCC CAAAUCUUCG CU	22-22	3	3'	X	7916929	7916950	1.7 kb from start of C39D10.3	uRNA
N1_C08-5	UCCCUGAGAA UUCUCGAACA GCUU	23-25	3	3'	X	8188341	8188364	0.3 kb from start of F22F1.t2	uRNA
mir-49	AAGCACCACG AGAAGCUGCA GA	22-22	2	3'	X	10023574	10023595	2.7 kb from end of F19C6.1, antisense	uRNA
N2_H06-10	UUUGUACUAC ACAUAGGUAC UGG	22-23	5	5'	X	11826683	11826705	6.1 kb from start of C34E11.1	uRNA
D1_F02-2	UUGAGCAAUG CGCAUGUGCG G	21-23	9	3'	X	12112846	12112826	in intron of W03G11.4	uRNA in an intron
mir-62	UGAUAUGUAA UCUAGCUUAC AG	21-22	10	3'	X	12726862	12726883	in intron of U07C5.1a	uRNA in an intron
mir-56	UACCCGUAAU GUUUCCGCUG AG	21-23	68	5'	X	13178947	13178926	5.2 kb from end of F09A5.2, antisense	uRNA
*mir-56**	UGGCGGAUCC AUUUUGGGUU GUA	23-23	2	3'	X	13178990	13178968	5.2 kb from end of F09A5.2, antisense	uRNA*
mir-55	UACCCGUAUA AGUUUCUGCU GAG	21-25	88	3'	X	13179084	13179062	5.3 kb from end of F09A5.2, antisense	uRNA
mir-54	UACCCGUAAU CUUCAUAAUC CGAG	21-25	88	3'	X	13179254	13179231	5.5 kb from end of F09A5.2, antisense	uRNA
mir-47	UGUCAUGGAG GCGCUCUCUU CA	21-23	29	3'	X	13955532	13955553	1.8 kb from end of K02B9.2, antisense	uRNA
let-7	UGAGGUAGUA GGUUGUAUAG UU	22-23	15	5'	X	14778474	14778453	3.1 kb from start of C05G5.2, antisense	uRNA
mir-84	UGAGGUAGUA UGUAAUAUUG UAGA	22-24	14	5'	X	16052268	16052245	0.8 kb from end of B0395.1, antisense	uRNA
mir-63	UAUGACACUG AAGCGAGUUG GAAA	22-24	8	3'	X	17628598	17628575	1.7 kb from start of C16H3.2, antisense	uRNA
N4_B09-3	GGGAGCUGUC AAUAACGGGG GG	22-22	1	-	I	227128	227149	in intron of Y48G1BM.4	mRNA?
N4_B09-3	GGGAGCUGUC AAUAACGGGG GG	22-22	1	-	I	238896	238917	in coding sequence of Y48G1BM.7, antisense	siRNA?
H1_F10-12	GAGCCGAUCC GUGACCUUC	19-19	1	-	I	252850	252832	1.3 kb from start of Y48G1BM.8	?
N0_2-19	CAAGUCCAUC AACGGCCAGA	20-20	1	-	I	4550757	4550738	in coding sequence of C44E4.6	mRNA
H1_F10-4	GAGAAAAAGC UACGACAUUU G	21-21	1	-	I	5368405	5368425	2.2 kb from end of U05E8.1, antisense	?
N2_F05-11	AGCUUCGCAG AUGCGCGGCC CAA	23-23	1	-	I	5535183	5535205	in coding sequence of F46F11.2	mRNA
N2_G06-2	UAAAUAUGGG CGCUAGCUCA	20-20	1	-	I	5975460	5975441	1.3 kb from start of C27A12.t1	?
H1_B12-2	CCAGAUUAGU AUUAAUUCAU G	21-21	1	-	I	7208991	7209011	0.3 kb from start of F21C3.6, antisense	?
D1_E09-2	UUUUAAUUAG UCUGAUAAAA CUUUGA	26-26	1	-	I	12581645	12581670	in coding sequence of H28O16.1	mRNA
H1_D12-3	UAAUGGCAUU GGCGAUAGA	19-19	1	-	I	14813541	14813523	in intron of ZK270.2b	mRNA?
H1_D12-3	AUUGGUAAGA AGAUUUCAC	19-19	1	-	I	14814150	14814132	in intron of ZK270.2b	mRNA?
H1_D12-3	AAGGAAUAAA AAAUUAUAU	19-19	1	-	I	14814759	14814741	in intron of ZK270.2b	mRNA?

H1_D12-3	CCCUUUAUAA UAAAGUUGG	19-19	1	-	I	14815368	14815350	in intron of ZK270.2b	mRNA?
H1_D12-3	AGUAAAAGGA GUUUUAAAC	19-19	1	-	I	14815977	14815959	3.1 kb from start of ZK270.2c	?
H1_D12-3	AACCAUUUUA CGUAGCGCU	19-19	1	-	I	14816586	14816568	3.7 kb from start of ZK270.2c	?
H1_D12-3	GCAUCUCCUG GAGGGCUCG	19-19	1	-	I	14817195	14817177	4.3 kb from start of ZK270.2c	?
H1_D12-3	UAUGAGAAUA UCUAAAGAA	19-19	1	-	I	14817804	14817786	4.9 kb from start of ZK270.2c	?
H1_D12-3	AAAAUUUGAA CUAGACAGU	19-19	1	-	I	14818413	14818395	5.5 kb from start of ZK270.2c	?
H1_D12-3	UUGCCAAGCU UUGGCAAAA	19-19	1	-	I	14819022	14819004	6.1 kb from start of ZK270.2c	?
H1_G10-6	UCCUUUGAUU GAGUUUUUUG U	21-23	5	-	I	14989176	14989196	2.4 kb from start of F31C3.6, antisense	?
H1_G10-6	UCCUUUGAUU GAGUUUUUUG U	21-23	5	-	I	14996372	14996392	9.6 kb from start of F31C3.6, antisense	?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	II	2024362	2024343	1.0 kb from start of U20902-1.F59H6	?
N4_A05-10	GAACAGAUAU GGUGACUUUU GUUGAC	26-26	1	-	II	2094887	2094912	in coding sequence of U16A1.9, antisense	siRNA?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	II	2551162	2551181	1.0 kb from end of K02F6.3	?
D2_H12-5	UAGCGAUGGU UUUAUGA	17-17	1	-	II	3306632	3306648	in coding sequence of F39E9.7, antisense	siRNA?
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	II	3521550	3521532	in coding sequence of W09G10.t1	mRNA
N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	II	6968440	6968416	in intron of C15F1.5	mRNA?
N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	II	7171814	7171838	0.5 kb from end of C27H5.8, antisense	siRNA?
L1_G01-2	UUUGCCGAUG UUUCUGAGAU GUC	23-23	1	-	II	7367517	7367539	in coding sequence of F43E2.6b, antisense	siRNA
N4_A03-6	UGCGGACAAU ACUUUUUUUG G	21-21	1	-	II	8464626	8464606	0.1 kb from start of F22B5.9, antisense	siRNA
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	II	8862692	8862673	in coding sequence of U07D4.1	mRNA?
D2_C07-7	AUAAGUAGUA GUGCCGCAGG UAA	23-23	1	-	II	11446836	11446814	1.9 kb from end of VW02B12L.4, antisense	?
N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	II	12944809	12944785	in intron of F58G1.7	mRNA?
H1_F10-12	GAGCCGAUCC GUGACCUUC	19-19	1	-	II	13044566	13044548	in coding sequence of F18A11.6, antisense	siRNA
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	II	14160813	14160794	in coding sequence of Y48B6A.3, antisense	siRNA
N1_B08-2	GGCCUUGGCG ACAUCGGCGA CU	22-22	1	-	II	14250773	14250752	in coding sequence of Y48B6A.14, antisense	siRNA
N1_F04-5	AAUGACACUG GUUAUCUUUU CCAUCGU	25-27	3	-	III	2218411	2218437	0.7 kb from start of Y48G9A.1	?
H1_H10-5	AGAAAGGUAU CGGGUGUCAU	18-20	3	-	III	2218491	2218510	0.6 kb from start of Y48G9A.1	?
D2_G10-5	CCCUGCCUGG GUCACCA	17-17	1	-	III	2477055	2477039	1.7 kb from end of W04B5.1, antisense	?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	III	3490911	3490891	0.7 kb from start of K01A11.4	?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	III	4047586	4047566	in intron of C36E8.1	mRNA
D1_A02-3	UCAGACCACU UGUCACU	17-17	1	-	III	4415786	4415802	in intron of B0285.9	mRNA?
N2_H12-15	CAUAAGGUAG GUCCGUGGG	19-19	1	-	III	4488766	4488784	in intron of C28A5.3	mRNA?
D1_H08-4	CUCUCCAGGA UGUCUUACCA ACCA	24-24	1	-	III	4799694	4799671	in coding sequence of B0393.1	mRNA
D2_B06-2	CGCAGACACG GGUGAGACUC AG	22-22	1	-	III	6426177	6426198	in coding sequence of C16A3.6	mRNA?

N0_151-1	GGAAAACGGG UUGAAAGGGA	20-20	1	-	III	7131264	7131283	0.1 kb from end of ZK418.9, antisense	siRNA
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	III	7353260	7353241	in intron of B0361.2	mRNA?
H1_A02-1	AGUACAAGCC AAUUGAUCUU CGCCU	25-25	1	-	III	7901405	7901429	in intron of ZK652.4	mRNA
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	III	9620936	9620917	in intron of C48B4.4b	mRNA
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	III	10989706	10989686	in intron of M142.6	mRNA?
L1_H09-2	UGCAAGGAAU UAUCAUCUAC C	21-21	1	-	III	13509501	13509521	in coding sequence of X76112-1.U27E9	mRNA?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	IV	280626	280606	in intron of U21D12.9b	mRNA?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	IV	2906845	2906825	0.5 kb from start of Y54G2A.19	?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	IV	3147245	3147265	4.3 kb from start of Y67D8C.10, antisense	?
D2_G10-5	CCCUGCCUGG GUCACCA	17-17	1	-	IV	5034590	5034574	in intron of C09B9.3	mRNA?
N4_D08-6	UAAAAUAGUC ACCAUUACAA A	21-21	1	-	IV	5117206	5117186	6.0 kb from start of AC7.3, antisense	?
H1_G12-3	GAUCACAGUA ACCGUGCGUA GAAUGA	26-26	1	-	IV	5321810	5321785	0.0 kb from start of ZK354.3, antisense	siRNA?
N4_D05-6	GAUUGUGGAC GAAACAUAGA CAUUGG	26-26	1	-	IV	5392175	5392200	1.1 kb from start of F41H10.2	?
N4_A11-2	UAUCUAAUAA UCGGCCCCUU C	21-21	1	-	IV	5439133	5439153	9.5 kb from start of ZK616.10, antisense	?
H1_F10-12	GAGCCGAUCC GUGACCUUC	19-19	1	-	IV	5497070	5497088	in coding sequence of B0273.3, antisense	siRNA
N3_D02-7	UUCUAAGGGC UUUGAAAAAU C	21-21	1	-	IV	5516738	5516758	in intron of B0273.4a	mRNA
N0_104-10	UCAUUUCUUG AUACAUUACU U	21-21	1	-	IV	5782106	5782086	1.5 kb from end of F44E8.1	?
D2_C10-6	UAACUUUCCG UCUACGUUAA C	21-21	1	-	IV	5932458	5932438	in intron of H32C10.3	mRNA?
H1_B10-6	UCUGUGUUCA AUUUUCACUG C	21-21	1	-	IV	6137304	6137284	in intron of M03D4.4	mRNA?
H1_G07-4	UCCGUGAACU UUUUUGCUUG C	21-21	1	-	IV	6190507	6190487	6.2 kb from start of C11D2.2, antisense	?
H1_D03-8	GUACUCGAUC GCAACAAUGG UUCCA	25-25	2	-	IV	6805789	6805813	in coding sequence of C17H12.3, antisense	siRNA?
N0_104-7	GGGCUCGGUC CGAGUUUCAU GGUCU	25-25	1	-	IV	7511097	7511121	in intron of F55G1.14	mRNA?
D2_E09-2	UGCCAGCCCG AUUCUGAAC	19-19	1	-	IV	8440219	8440201	0.1 kb from end of U26A8.2, antisense	siRNA?
L1_D07-8	CCCGCAAGAG UGAGAC	16-16	1	-	IV	9319775	9319760	in coding sequence of F49C12.6	mRNA?
N1_D02-8	UAUAGAUUUA UAAAUUAUGU G	21-21	1	-	IV	9488366	9488346	in intron of F38E11.7	mRNA?
L1_A07-3	UUUCUGGGUG UGAUGGCGUC AG	22-22	1	-	IV	11605375	11605354	1.0 kb from start of F40F11.3	?
L1_C07-5	UCCGAGGACA AGUCGAG	17-17	2	-	IV	12154621	12154605	in coding sequence of B0001.5	mRNA
N1_C11-8	GGAGCACGCU GACGAACUGG	20-20	1	-	IV	13106193	13106174	in coding sequence of C39E9.12, antisense	siRNA
H1_B12-2	GUUGAUAACU GAAACAAUAA A	21-21	1	-	IV	13600279	13600259	0.3 kb from start of Y45F10B.11, antisense	?
D2_E01-7	AAUUAAUCUA CUCCACAAUC C	21-21	1	-	IV	13728222	13728242	in coding sequence of C47A4.1, antisense	siRNA?
H1_F10-7	GUUAGGCUUC GCAGUGGGGU U	21-21	1	-	IV	13823290	13823310	4.0 kb from start of Y45F10D.1	?
N3_F06-3	AAUAUUUAUC GGCUUCGUCC A	21-21	1	-	IV	13829597	13829617	1.2 kb from end of Y45F10D.1	?
H1_G04-3	UCAAAGCUUU GGCAAAUUAC C	21-21	1	-	IV	13844490	13844470	1.4 kb from end of H08M01.1	?

H1_B12-2	UGUGGAACAC GAUAUUACAA U	21-21	1	-	IV	13895911	13895891	0.3 kb from end of B0513.2a, antisense	siRNA
H1_B12-2	ACAUAUUUGU UGAAGAGUGU U	21-21	1	-	IV	13896640	13896620	1.0 kb from end of B0513.2a, antisense	?
N3_E04-7	AAUUGAGAAU AAAUAUAUGC A	21-21	1	-	IV	13902876	13902856	in intron of B0513.1	mRNA?
N1_F01-7	GUUUCAAACA GUUGUGAAAU U	21-21	1	-	IV	13965841	13965861	1.1 kb from end of C27H2.2	?
N1_G02-12	UUUUAUUUUU GUUUUUAAAU G	21-21	1	-	IV	14036476	14036496	0.3 kb from end of Y37A1B.8	mRNA?
N4_C12-8	GCGCAGUUUG GAGGACGAAA U	21-21	1	-	IV	14058160	14058180	2.8 kb from start of Y37A1B.7, antisense	?
L1_H03-4	UAAAAGCCAU UAGCAACCGA A	21-21	1	-	IV	14151304	14151324	in coding sequence of H12I19.2	mRNA?
N2_G08-3	UCUGUAACUU UUUCUCAAAA U	21-21	1	-	IV	14176434	14176454	0.9 kb from start of R05A10.6	?
N3_D09-5	UUGCAUUCUC AACAAUUUCU G	21-21	1	-	IV	14178188	14178168	in intron of R05A10.6	mRNA?
N0_175-2	UUUUGAAUUU GUAUUUUCUU A	21-21	1	-	IV	14250887	14250907	0.2 kb from start of Y64G10A.1, antisense	siRNA?
N3_H03-4	AUAAUUUGGU AUUAUUUUGA A	21-21	1	-	IV	14279557	14279537	8.6 kb from start of Y64G10A.6, antisense	?
N1_H08-2	GUUCCGAUUG ACGCGUCGGA C	21-21	1	-	IV	14309606	14309586	4.2 kb from start of Y64G10A.7, antisense	?
H1_A03-1	CACAGAUCAA CUAUUCGAGA A	21-21	1	-	IV	14317937	14317957	in intron of Y64G10A.7	mRNA
N4_A09-12	AUUAUUGGUA CGUUUGUGGG A	21-21	1	-	IV	14351766	14351746	in intron of Y67A10A.2	mRNA?
N0_90-8	UUUUUUUAAU UCUAAGAUGA U	21-21	1	-	IV	14368898	14368878	3.9 kb from start of Y67A10A.6, antisense	?
N5_C02-6	UUUUGUCUUU GUCUGUUUUU U	21-21	1	-	IV	14486865	14486845	6.5 kb from end of LLC1.3, antisense	?
N3_E08-6	AAAAAAGGCU UCAAUCAGAU U	21-21	1	-	IV	14500747	14500767	in coding sequence of Y57G11A.1, antisense	siRNA?
N3_D10-1	CUUGUACAAA AUACGUCUCC C	21-21	1	-	IV	14502635	14502615	in intron of Y57G11A.1	mRNA?
N2_E07-7	CUUUUCUUGU UAUAAACCAG A	21-21	1	-	IV	14507026	14507046	in intron of Y57G11A.1	mRNA
N1_B05-4	UCAGAUUUUA AGUACCACAA A	21-21	1	-	IV	14643430	14643450	1.3 kb from start of F13G11.2	?
N1_A03-5	ACAGCUGUUG AUUGAAAAAG U	21-21	1	-	IV	14658600	14658580	in coding sequence of Y57G11C.32	mRNA
N5_F11-5	AAAAUUUGGA GCAACAUCAC A	21-21	1	-	IV	14665171	14665151	3.5 kb from start of Y57G11C.33, antisense	?
N2_C11-3	ACAGCUGGAA CAUAUUUUUG A	21-21	1	-	IV	14692648	14692668	2.0 kb from start of Y57G11C.31	?
D1_A09-3	GAUACAUUUU UAAAGAUAAC U	21-21	1	-	IV	14696053	14696033	in intron of Y57G11C.31	mRNA?
N3_B02-4	CCAGAAUACC UCAAUAUCUC U	21-21	1	-	IV	14725125	14725105	in coding sequence of Y57G11C.2, antisense	siRNA?
N4_D03-9	UUCGACAGUA CUCAAAAUAA U	21-21	1	-	IV	14726023	14726003	in coding sequence of Y57G11C.2, antisense	siRNA?
N1_E05-7	AAACCAGUGU GUGUAAGAAA U	21-21	2	-	IV	14729765	14729745	in coding sequence of Y57G11C.2, antisense	siRNA?
N4_D01-7	CAGAAAUGUA AGAAGAGUUU C	21-21	2	-	IV	14733924	14733904	1.9 kb from end of Y57G11C.2, antisense	?
N2_E01-4	UGUUUCACCG UGUUAAUAUC A	21-21	2	-	IV	14734634	14734654	2.6 kb from end of Y57G11C.2	?
H1_H04-3	UAUUUAAAAA GACGGAAAU	19-19	1	-	IV	14900657	14900639	in coding sequence of Y57G11C.t1	mRNA
N0_133-2	UUGACAUUGC UCGUCUGAAU A	21-21	1	-	IV	14904452	14904432	in intron of Y57G11C.24c, antisense	siRNA?
N1_H05-3	AUUUUUUUAU UGAGUCCUUU C	21-21	1	-	IV	14907238	14907258	in intron of Y57G11C.24a	mRNA?
H1_H02-14	UAUUAAGAUA UUAUUUUACC A	21-21	2	-	IV	14908381	14908401	0.6 kb from end of Y57G11C.24a	?

H1_C12-6	UUAUGCAGAC AUUGAAACAA G	21-21	1	-	IV	15074935	15074915	1.9 kb from end of Y41E3.15	?
H1_F12-5	UUGGACGAUU UUACCAUUUC U	21-21	1	-	IV	15079077	15079097	in intron of Y41E3.15	mRNA?
N2_E12-1	AGUUUUUGCA GAAAUUAAAA U	21-21	1	-	IV	15114034	15114014	1.2 kb from start of M199.1, antisense	?
N0_164-3	CGUUGAUAAA GCUGUACUCU G	21-21	1	-	IV	15145750	15145770	4.9 kb from end of Y40H7A.8, antisense	?
N0_117-6	CACUCCCCCC AGUACAAAAU U	21-21	3	-	IV	15147823	15147803	2.8 kb from end of Y40H7A.8	?
H1_H04-3	UGUUAUCAGA UCAAUAUCA	19-19	1	-	IV	15198906	15198924	in coding sequence of Y40H7A.t1	mRNA?
H1_H04-3	AAGUAAAAAC AAAAUGAAU	19-19	1	-	IV	15199289	15199271	in coding sequence of Y40H7A.t2	mRNA?
H1_F10-10	UGUGAAUAAU GAAAAAUUGA A	21-21	1	-	IV	15259306	15259326	2.4 kb from end of Y73F8A.3, antisense	?
N2_A03-3	UAUUCUUUUG UGAUUUGAAA A	21-21	1	-	IV	15346104	15346084	2.5 kb from start of Y73F8A.18, antisense	?
N4_G02-4	AAAAACAAAC UUGCAAAGUA G	21-21	2	-	IV	15363027	15363047	10.1 kb from start of Y73F8A.19	?
N5_C12-5	AAAACAAACU UGCAAAGUAG U	21-21	2	-	IV	15363028	15363048	10.1 kb from start of Y73F8A.19	?
N0_123-2	UCUCGCCACG AUUGCAAAAU U	21-21	1	-	IV	15431829	15431809	12.3 kb from end of Y73F8A.21, antisense	?
L1_B12-9	AACGGUCCCC UUUGCAGAAU U	21-21	1	-	IV	15494197	15494177	1.7 kb from end of Y73F8A.28	?
L1_F04-3	UUGUGAAAAA UUUGUAAUGU C	21-21	1	-	IV	15555870	15555850	4.4 kb from start of Y73F8A.35	?
H1_C08-8	AUAAGACGUG UAAUUAAUUA G	21-21	1	-	IV	15557021	15557001	5.5 kb from start of Y73F8A.35	?
N2_A10-6	AAUAAAACCG AGUGUUCCUG U	21-21	1	-	IV	15561253	15561273	1.5 kb from end of Y105C5A.1, antisense	?
N3_E02-8	AGUAAUCCUU GAUUUACCUC A	21-21	1	-	IV	15670204	15670224	6.8 kb from start of Y105C5A.14, antisense	?
N4_H10-1	UCGAAUUUUU UUCUGUAAAU U	21-21	1	-	IV	15670987	15671007	7.6 kb from start of Y105C5A.14, antisense	?
N3_E02-6	CAAACAUGUG UUUCAAGUUU G	21-21	1	-	IV	15674001	15674021	10.6 kb from start of Y105C5A.14, antisense	?
N2_D05-11	UAUACGGCAA AUCAGCAGUU U	21-21	1	-	IV	15682695	15682715	19.3 kb from start of Y105C5A.14, antisense	?
N0_164-3	UUCUUGCGAA AAGCACUUUA A	21-21	1	-	IV	15693295	15693315	29.2 kb from start of Y105C5A.15	?
N0_184-10	UUCAGGAUGU CUCUUUAUCA U	21-21	1	-	IV	15789291	15789311	4.7 kb from end of Y105C5A.21, antisense	?
N5_B12-9	UCACGGAAAU UCUUGAAAUU G	21-21	1	-	IV	15818080	15818100	6.8 kb from start of Y105C5A.22, antisense	?
N3_A06-4	UGGAUCAUGG AUCAUGAGAA U	21-21	1	-	IV	15826790	15826810	4.8 kb from start of Y105C5A.23	?
N0_177-6	CUAAAAUCAC UGAAUAAUAC A	18-18	1	-	IV	15850578	15850598	11.0 kb from start of Y105C5A.24	?
N0_9-9	GCAAAAUGUU AUUAUUUUUU A	21-21	1	-	IV	15903562	15903582	0.0 kb from end of Y105C5B.4, antisense	siRNA?
L1_D07-11	AACCCGGGUC UCGCAUGUGA C	21-21	1	-	IV	15909891	15909871	1.7 kb from end of Y105C5B.5	?
N3_F01-6	CAUGUUUUCG CAGUUUAUAA C	21-21	1	-	IV	15999656	15999676	4.1 kb from start of Y105C5B.19, antisense	?
N2_F03-2	AUGAAACAGC ACGAUUUUUU U	21-21	1	-	IV	16094808	16094828	19.1 kb from start of Y105C5B.24, antisense	?
N4_A01-10	UUCUACUGUU UCAUACUUUU U	21-21	2	-	IV	16112602	16112622	36.9 kb from start of Y105C5B.24, antisense	?
N4_A01-10	GUUAAUUGCA AGUUUUUAUG G	21-21	2	-	IV	16119576	16119556	34.5 kb from end of Y105C5B.25	?
N4_D11-4	GAUCAAUUUA AAAUAAAACA C	21-21	1	-	IV	16126690	16126710	27.4 kb from end of Y105C5B.25, antisense	?
H1_E10-1	UGAAAUAGAU UACUGAG	17-17	1	-	IV	16144833	16144817	9.3 kb from end of Y105C5B.25	?

D1_H04-4	ACAAUCCUUU UAUGGCCUAA A	21-21	2	-	IV	16146249	16146229	7.9 kb from end of Y105C5B.25	?
L1_A10-2	AAAAUAAUAA AUGUAACGCA A	21-21	1	-	IV	16192057	16192037	9.8 kb from start of Y105C5B.28	?
N3_A01-9	GUGAAUUUUU AUCGUUUCAA C	21-21	1	-	IV	16204207	16204187	4.4 kb from start of Y7A9A.1, antisense	?
H1_B12-2	GCUCAUUUGC AAGUCUUUGG U	21-21	1	-	IV	16236188	16236208	1.4 kb from start of H25K10.4	?
D1_B11-2	AUAGACAGAG AAGCGGAUG	19-23	9	-	IV	16400017	16400035	in coding sequence of Y7A9D.t1	mRNA
D1_B11-2	AAAUAACAUU GAAACAGUA	19-23	9	-	IV	16412457	16412475	in coding sequence of Y65A5A.t4	mRNA
N0_74-9	AAAUGAGAUG UUAUUUUUAU U	21-21	1	-	IV	16454228	16454248	2.0 kb from start of Y10G11A.c	?
N4_G06-8	GUAAUUAAGU UCAAUCAAAU C	21-21	1	-	IV	16526908	16526888	in intron of Y51H4A.7	mRNA?
N3_E10-7	AAAUUUUCAG CAGUUUGCUA A	21-21	1	-	IV	16594293	16594313	0.7 kb from start of Y51H4A.9, antisense	?
D2_H01-24	AGAAGGUUAG UUUUUUUUUC A	21-21	4	-	IV	16677880	16677900	1.4 kb from start of Y51H4A.21, antisense	?
N4_G01-7	CUACCAGCCG GAAACAAAAA G	21-21	1	-	IV	16706144	16706124	5.3 kb from start of Y51H4A.22, antisense	?
D2_G10-5	CUCUGAAAAU AAAAAAA	17-17	1	-	IV	16731501	16731485	in intron of Y51H4A.25	mRNA?
N1_E09-3	UUGUGGGGAC AAUGAGGGAG G	21-21	1	-	IV	16799458	16799478	2.2 kb from start of Y116A8A.3, antisense	?
N1_E09-3	UUGGGCGGAG GUCGGACGGA A	21-21	1	-	IV	16799787	16799767	2.6 kb from start of Y116A8A.3	?
N5_H11-5	GAAGAAAAUG GCGAAAGUUU C	21-21	1	-	IV	16821779	16821799	0.2 kb from end of Y116A8A.7	mRNA?
N2_D12-4	AAAAUCUGAA AAUUUCGAUU U	21-21	1	-	IV	16894673	16894693	1.4 kb from start of Y116A8C.1, antisense	?
H1_D02-3	UCGAGUUUUA ACGUGAAUAA U	21-21	1	-	IV	16903679	16903699	3.2 kb from start of Y116A8C.3	?
N1_B01-5	AGUGCCUAAU CAUGGAUUAU A	21-21	1	-	IV	16981500	16981480	1.0 kb from end of Y116A8C.11, antisense	?
N2_E06-2	CAACUGAAUA GGUUGAUUUG U	21-21	1	-	IV	17017571	17017551	1.7 kb from start of Y116A8C.17	?
H1_B12-2	AGUGUCGGCU GCAAAAAUUU C	21-21	1	-	IV	17139666	17139686	in intron of Y116A8C.41	mRNA?
N3_C04-3	GCCACGCUAU CAUUAAUUCC A	21-21	1	-	IV	17189616	17189596	in intron of C52D10.12	mRNA?
D2_G10-5	AUUAUCAUAA AAUGCCG	17-17	1	-	IV	17300111	17300095	in intron of U28F3.9	mRNA?
H1_A01-7	UUCCCCACCA CAAACCAACA	20-20	1	-	V	1474935	1474954	0.0 kb from start of C38C3.5c	mRNA
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	V	3440940	3440960	0.5 kb from end of F27E11.3b, antisense	siRNA
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	V	5394737	5394717	2.9 kb from start of C35A11.4	?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	6713442	6713461	0.8 kb from start of W02F12.6, antisense	?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	8811870	8811889	in intron of C01B7.6	mRNA?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	9001394	9001375	in intron of F59E11.3	mRNA?
N2_G10-3	UGGUUUUGGA UAUCUCUGUC C	21-21	2	-	V	11175507	11175487	in coding sequence of U19C4.5, antisense	siRNA?
N4_F11-5	UGAUGUCCGA UCUAUUUGAA G	21-21	1	-	V	12113870	12113850	1.3 kb from start of C13C4.4	uRNA
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	13039393	13039412	0.9 kb from end of W07G4.3, antisense	?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	13272889	13272908	1.5 kb from end of C34D1.5	?
H1_H06-6	UGUCAGGGGU GAAGAACGAC	20-20	1	-	V	13909366	13909347	in coding sequence of C06B3.4, antisense	siRNA?

N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	V	14030170	14030146	1.2 kb from start of U08G5.3	?
N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	V	14269168	14269192	in coding sequence of F40G12.4, antisense	siRNA?
N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	V	14279199	14279223	0.8 kb from end of F40G12.8, antisense	?
N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	V	14464888	14464864	0.3 kb from end of F08H9.4	mRNA?
N3_G06-5	CCCCAUGGUG AGGCCUACCC AUUGC	21-25	2	-	V	14470269	14470245	0.6 kb from start of F08H9.6, antisense	?
D1_E07-3	CACGAGGACU CUACUAACCG C	21-21	1	-	V	14689100	14689120	in coding sequence of C47E8.5	mRNA
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	15391832	15391851	0.9 kb from start of R08H2.3, antisense	?
H1_H04-3	GGCCUGCUUA GUAUAGUGG	19-19	1	-	V	15555606	15555624	in intron of F28F8.1	mRNA?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	15558470	15558451	2.2 kb from start of F28F8.1	?
L1_A01-6	GGACGAUUCG GUGUCAACGU A	21-21	1	-	V	17075393	17075373	0.5 kb from end of U05E12.1	mRNA?
N5_E08-3	GGCCAGUAGA UUCAGACGUG CCU	23-23	1	-	V	17135862	17135884	in coding sequence of Y102A5D.2, antisense	siRNA?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	V	17796166	17796186	in coding sequence of C47A10.7, antisense	siRNA?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	V	17839819	17839799	0.1 kb from end of Y59A8A.7, antisense	siRNA?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	18617738	18617719	0.4 kb from start of Y51A2D.17, antisense	?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	V	19241540	19241521	0.9 kb from start of F21D9.5, antisense	?
L1_G05-8	GAAAAGGUUA GUUGUAAGGU UUUUUC	26-26	1	-	V	19722671	19722646	0.2 kb from end of Y116F11A.1, antisense	siRNA?
N2_A11-5	GGAACAAGAU GGGUGUUCCG UUCGCU	26-26	1	-	V	20371419	20371394	in coding sequence of K02E2.6, antisense	siRNA
L1_H01-1	GAACAGAUAU UCUGUAUGAG GUGUCU	26-26	1	-	V	20374841	20374816	0.4 kb from end of K02E2.7	mRNA?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	X	688307	688288	2.3 kb from start of U19D7.3, antisense	?
N4_G05-2	CGUCAUCUGA UGCGGUUGUU CCA	23-23	1	-	X	1045970	1045992	in coding sequence of F55A4.9	mRNA?
D2_H12-5	UAGCGAUGGU UUUAUGA	17-17	1	-	X	1059386	1059402	in coding sequence of F55A4.4, antisense	siRNA?
N4_H05-5	UGUGUUUUUU GUUGAGGUAU C	21-21	2	-	X	2363291	2363311	6.8 kb from start of U01B6.1	?
N4_G11-3	AUCCUCGGUA GUAUAGUGGU GAGUAUCCG	27-29	3	-	X	2587866	2587894	in coding sequence of F52H2.t1	mRNA
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	X	4902618	4902598	in intron of F39C12.3	mRNA
H1_H04-3	GGCCUGCUUA GUAUAGUGG	19-19	1	-	X	5057167	5057149	in coding sequence of ZC449.t1	mRNA?
N5_A02-3	UCCGCUUCUA ACUUCCAUUU GCAG	23-24	9	-	X	8519601	8519578	9.1 kb from start of U01B10.4	?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	X	8609714	8609694	0.4 kb from end of F18E9.5b, antisense	siRNA?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	X	11645197	11645217	5.6 kb from start of U05038-1.F52D10, antisense	?
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	12027815	12027797	in coding sequence of C49F5.t1	mRNA?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	X	12889984	12890003	0.9 kb from start of U14G8.t1, antisense	?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	X	13295653	13295633	0.2 kb from end of K08H2.t1, antisense	siRNA?
H1_B12-2	UCAAAGAUCA AACUACAUAG G	21-21	1	-	X	14153555	14153535	in intron of F28H6.6	mRNA?

D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	14169265	14169283	in coding sequence of F28H6.t2	mRNA
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	14169601	14169583	in coding sequence of F28H6.t7	mRNA
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	14170094	14170112	in coding sequence of F28H6.t3	mRNA?
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	14170417	14170399	in coding sequence of F28H6.t6	mRNA?
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	14170910	14170928	in coding sequence of F28H6.t4	mRNA?
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	14171232	14171214	in coding sequence of F28H6.t5	mRNA?
N5_G09-4	CAACCAUUGG AAUUUCUCUA U	21-21	1	-	X	15622552	15622532	in coding sequence of K09A9.5	uRNA?
H1_F09-13	UGGAAUGAUU GAGCUUGAUG GAU	23-23	1	-	X	16237802	16237824	0.3 kb from end of R11.t4, antisense	siRNA?
D1_B11-2	GGCCUGCGUA GUAUAGUGG	19-23	9	-	X	16664698	16664680	in intron of C06G1.1	mRNA?
N2_G06-2	CUACAAAGAA GAAAACAAGG	20-20	1	-	X	16985245	16985264	3.3 kb from start of F52G3.4, antisense	?

21. Supplementary Web material is available on *Science* Online at www.sciencemag.org/cgi/content/full/294/5543/853/DC1.
22. S. M. Hammond, S. Boettcher, A. A. Caudy, R. Kobayashi, G. J. Hannon, *Science* **293**, 1146 (2001).
23. A. A. Aravin *et al.*, *Curr. Biol.* **11**, 1017 (2001).
24. H. Tabara *et al.*, *Cell* **99**, 123 (1999).
25. M. Fagard, S. Boutet, J. B. Morel, C. Bellini, H. Vaucheret, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 11650 (2000).
26. C. Catalanotto, G. Azzalin, G. Macino, C. Cogoni, *Nature* **404**, 245 (2000).
27. S. R. Eddy, *Curr. Opin. Genet. Dev.* **9**, 695 (1999).
28. K. M. Wassarman, F. Repoila, C. Rosenow, G. Storz, S. Gottesman, *Genes Dev.* **15**, 1637 (2001).
29. J. Cavaille *et al.*, *Proc. Natl. Acad. Sci. U.S.A.* **97**, 14311 (2000).
30. A. Hüttenhofer *et al.*, *EMBO J.* **20**, 2943 (2001).
31. L. Argaman *et al.*, *Curr. Biol.* **11**, 941 (2001).
32. D. H. Mathews, J. Sabina, M. Zuker, D. H. Turner, *J. Mol. Biol.* **288**, 911 (1999).
33. We acknowledge J. Martinez, S. M. Elbashir, C. Will, R. Rivera-Pomar, S. Baskerville, B. Reinhart, and P. D. Zamore for comments on the manuscript; G. Hernandez for the gift of total RNA isolated from staged fly populations; H. Brahms, C. Schneider, P. Kempkes, E. Raz, and A. Mansouri for providing tissues or cells; B. Reinhart for advice on Northern blot analysis; G. Dowe for sequencing; P. Bucher for bioinformatic consultations; and R. Lührmann for support. Funded by a BMBF Biofuture grant.

31 July 2001; accepted 14 September 2001

An Abundant Class of Tiny RNAs with Probable Regulatory Roles in *Caenorhabditis elegans*

Nelson C. Lau, Lee P. Lim, Earl G. Weinstein, David P. Bartel*

Two small temporal RNAs (stRNAs), *lin-4* and *let-7*, control developmental timing in *Caenorhabditis elegans*. We find that these two regulatory RNAs are members of a large class of 21- to 24-nucleotide noncoding RNAs, called microRNAs (miRNAs). We report on 55 previously unknown miRNAs in *C. elegans*. The miRNAs have diverse expression patterns during development: a *let-7* paralog is temporally coexpressed with *let-7*; miRNAs encoded in a single genomic cluster are coexpressed during embryogenesis; and still other miRNAs are expressed constitutively throughout development. Potential orthologs of several of these miRNA genes were identified in *Drosophila* and human genomes. The abundance of these tiny RNAs, their expression patterns, and their evolutionary conservation imply that, as a class, miRNAs have broad regulatory functions in animals.

Two types of short RNAs, both about 21 to 25 nucleotides (21–25 nt) in length, serve as guide RNAs to direct posttranscriptional regulatory machinery to specific mRNA targets. Small temporal RNAs (stRNAs) control developmental timing in *Caenorhabditis elegans* (*1–3*). They pair to sites within the 3′ untranslated region (3′ UTR) of target mRNAs, causing translational repression of these mRNAs and triggering the transition to the next developmental stage (*1–5*). Small interfering RNAs (siRNAs), which direct mRNA cleavage during RNA interference (RNAi) and related processes, are the other type of short regulatory RNAs (*6–12*). Both stRNAs and siRNAs are generated by processes requiring Dicer, a multidomain protein with tandem ribonuclease III (RNase III) domains (*13-15*). Dicer cleaves within the double-stranded portion of precursor molecules to yield the 21–25 nt guide RNAs.

lin-4 and *let-7* have been the only two stRNAs identified, and so the extent to which this type of small noncoding RNA normally regulates eukaryotic gene expression is only beginning to be understood (*1–5*). RNAi-related processes protect against viruses or mobile genetic elements, yet these processes are known to normally regulate only one other mRNA, that of *Drosophila Stellate* (*16–20*). To investigate whether RNAs resembling stRNAs or siRNAs might play a more general role in gene regulation, we isolated and cloned endogenous *C. elegans* RNAs that have the expected features of Dicer products. Tuschl and colleagues showed that such a strategy is feasible when they fortuitously cloned endogenous *Drosophila* RNAs while cloning siRNAs processed from exogenous dsRNA in an embryo lysate (*12*). Furthermore, other efforts focusing on longer RNAs have recently uncovered many previously unknown noncoding RNAs (*21*, *22*).

Dicer products, such as stRNAs and siRNAs, can be distinguished from most other oligonucleotides that might be present in *C. elegans* by three criteria: a length of about 22 nt, a 5′-terminal monophosphate, and a 3′-terminal hydroxyl group (*12*, *13*, *15*). Accordingly, a procedure was developed for isolating and cloning *C. elegans* RNAs with these features (*23*). Of the clones sequenced, 330 matched *C. elegans* genomic sequence, including 10 representing *lin-4* RNA and 1 representing *let-7* RNA. Another 182 corresponded to the *Escherichia coli* genomic sequence. *E. coli* RNA clones were expected because the worms were cultured with *E. coli* as the primary food source.

Three hundred of the 330 *C. elegans* clones have the potential to pair with nearby genomic sequences to form fold-back structures resembling those thought to be needed for Dicer processing of *lin-4* and *let-7* stRNAs (Fig. 1) (*24*). These 300 clones with predicted fold-backs represent 54 unique sequences: *lin-4*, *let-7*, and 52 other RNAs (Table 1). Thus, *lin-4* and *let-7* RNAs appear to be members of a larger class of noncoding RNAs that are about 20–24 nt in length and are processed from fold-back structures. We and the two other groups reporting in this issue of the journal refer to this class of tiny RNAs as microRNAs, abbreviated miRNAs, with individual miRNAs and their genes designated miR-# and *mir-#*, respectively (*25*, *26*).

We propose that most of the miRNAs are expressed from independent transcription units, previously unidentified because they do not contain an open reading frame (ORF) or other features required by current gene-recognition algorithms. No miRNAs matched a transcript validated by an annotated *C. elegans* expressed sequence tag (EST), and most were at least 1 kb from the nearest annotated sequences (Table 1). Even the miRNA genes near predicted coding regions or within predicted introns are probably expressed separately from the annotated genes. If most miRNAs were expressed from the same primary transcript as the predicted protein, their orientation would be predominantly the same as the predicted mRNA, but no such bias in orientation was observed (Table 1). Likewise, other types of RNA genes located within *C. elegans* intronic regions are usually expressed from independent transcription units (*27*).

Whereas both *lin-4* and *let-7* RNAs reside on the 5′ arm of their fold-back structures (*1*, *3*), only about a quarter of the other miRNAs lie on the 5′ arm of their proposed fold-back structures, as exemplified by miR-84 (Table 1 and Fig. 1A). All the others are on the 3′ arm, as exemplified by miR-1 (Table 1 and Fig. 1B). This implies that the stable product of Dicer processing can reside on either arm of the precursor and that features of the miRNA or its precursor—other than the loop connecting the two arms—must determine

Whitehead Institute for Biomedical Research, and Department of Biology, Massachusetts Institute of Technology, 9 Cambridge Center, Cambridge, MA 02142, USA.

*To whom correspondence should be addressed. E-mail: dbartel@wi.mit.edu

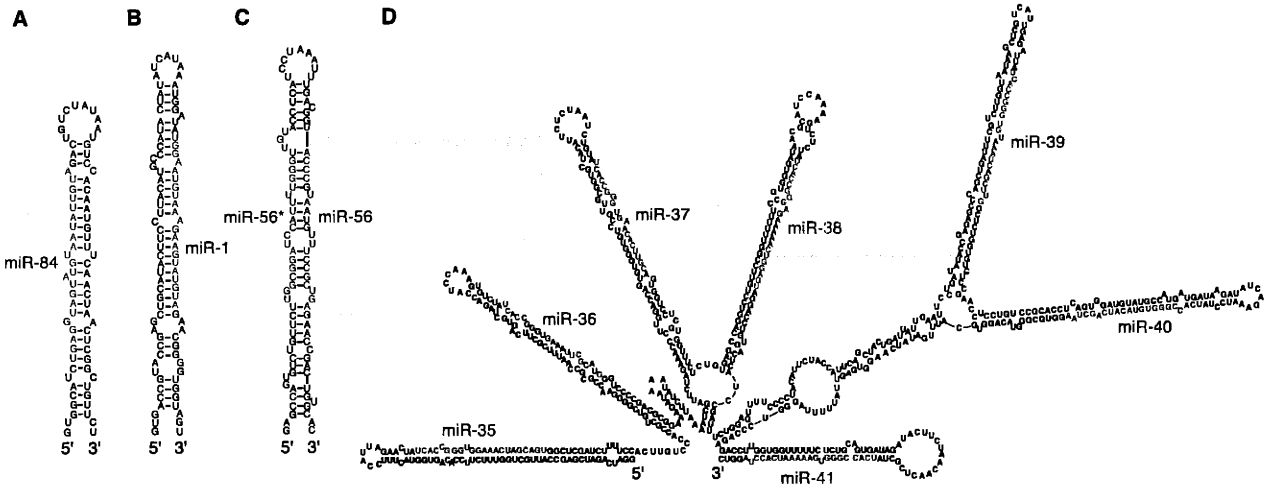


Fig. 1. Fold-back secondary structures involving miRNAs (red) and their flanking sequences (black), as predicted computationally using RNAfold (*35*). (**A**) miR-84, an miRNA with similarity to *let-7* RNA. (**B**) miR-1, an miRNA highly conserved in evolution. (**C**) miR-56 and miR-56*, the only two miRNAs cloned from both sides of the same fold-back. (**D**) The *mir-35–mir-41* cluster.

which side of the fold-back contains the stable product.

When compared with the RNA fragments cloned from *E. coli*, the miRNAs had unique length and sequence features (Fig. 2). The *E. coli* fragments had a broad length distribution, ranging from 15–29 nt, which reflects the size-selection limits imposed during the cloning procedure (*23*). In contrast, the miRNAs had a much tighter length distribution, centering on 21–24 nt, coincident with the known specificity of Dicer processing (Fig. 2A). The miRNA sequence composition preferences were most striking at the 5′ end, where there was a strong preference for U and against G at the first position and then a deficiency of U at positions 2 through 4 (Fig. 2B). miRNAs were also generally deficient in C, except at position 4. These composition preferences were not present in the clones representing *E. coli* RNA fragments.

The expression of 20 cloned miRNAs was examined, and all but two (miR-41 and miR-68) were readily detected on Northern blots (Fig. 3). For these 18 miRNAs with detectable expression, the dominant form was the mature 20–24 nt fragment(s), though for most, a longer species was also detected at the mobility expected for the fold-back precursor RNA. Fold-back precursors for *lin-4* and *let-7* have also been observed, particularly at the stage in development when the stRNA is first expressed (*1*, *14*, *15*).

Because the miRNAs resemble stRNAs, their temporal expression was examined. RNA from wild-type embryos, the four larval stages (L1 through L4), and young adults was probed. RNA from *glp-4 (bn2)* young adults, which are severely depleted in germ cells (*28*), was also probed because miRNAs might have critical functions in the germ line, as suggested by the finding that worms deficient in Dicer have germ line defects and are sterile (*14*, *29*). Many miRNAs have intriguing expression patterns during development (Fig. 3). For example, the expression of miR-84, an miRNA with 77% sequence identity to *let-7* RNA, was found to be indistinguishable from that of *let-7* (Fig. 3). Thus, it is tempting to speculate that miR-84 is an stRNA that works in concert with *let-7* RNA to control the larval-to-adult transition, an idea supported by the identification of plausible binding sites for miR-84 in the 3′ UTRs of appropriate heterochronic genes (*30*).

Nearly all of the miRNAs appear to have orthologs in other species, as would be expected if they had evolutionarily conserved regulatory roles. About 85% percent of the newly found miRNAs had recognizable homologs in the available *C. briggsae* genomic sequence, which at the time of our analysis included about 90% of the *C. briggsae* genome (Table 1). Over 40% of the miRNAs appeared to be identical in *C. briggsae*, as seen with the *lin-4* and *let-7* RNAs (*1*, *3*). Those miRNAs not absolutely conserved between *C. briggsae* and *C. elegans* might still have important functions, but they may have more readily co-varied with their target sites because, for instance, they might have fewer target sites. When the sequence of the miRNA differs from that of its homologs, there is usually a compensatory change in the other arm of the fold-back to maintain pairing, which provides phylogenetic evidence for the existence and importance of the fold-back secondary structures. *let-7*, but not *lin-4*, has discernable homologs in more distantly related organisms, including *Drosophila* and human (*31*). At least seven other miRNA genes (*mir-1*, *mir-2*, *mir-34*, *mir-60*, *mir-72*, *mir-79*, and *mir-84*) appear to be conserved in *Drosophila*, and most of these (*mir-1*, *mir-34*, *mir-60*, *mir-72*, and *mir-84*) appear to be

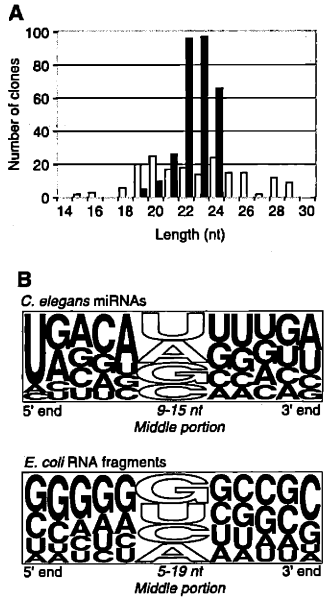


Fig. 2. Unique sequence features of the miRNAs. (**A**) Length distribution of the clones representing *E. coli* RNA fragments (white bars) and *C. elegans* miRNAs (black bars). (**B**) Sequence composition of the unique clones representing *C. elegans* miRNAs and *E. coli* RNA fragments. The height of each letter is proportional to the frequency of the indicated nucleotide. Solid letters correspond to specific positions relative to the ends of the clones; outlined letters represent the aggregate composition of the interior of the clones. To avoid overrepresentation from groups of related miRNAs in this analysis, each set of paralogs was represented by its consensus sequence.

conserved in humans (*24*). The most highly conserved miRNA found, miR-1, is expressed throughout *C. elegans* development (Fig. 3) and therefore is unlikely to control developmental timing but may control tissue-specific events.

The distribution of miRNA genes within the *C. elegans* genome is not random (Table 1). For example, clones for six miRNA paralogs

Table 1. miRNAs cloned from *C. elegans*. 300 RNA clones represented 54 different miRNAs. Also included are miR-39, miR-65, and miR-69, three miRNAs predicted based on homology and/or proximity to cloned miRNAs. miR-39 and miR-69 have been validated by Northern analysis (Fig. 3), whereas miR-65 is not sufficiently divergent to be readily distinguished by Northern analysis. All *C. elegans* sequence analyses relied on WormBase, release WS45 (*33*). Some miRNAs were represented by clones of different lengths, due to heterogeneity at the miRNA 3′ terminus. The observed lengths are indicated, as is the sequence of the most abundant length. Comparison to *C. briggsae* shotgun sequencing traces revealed miRNA orthologs with 100% sequence identity (+++) and potential orthologs with >90% (++) and >75% (+) sequence identity (*24*, *34*). Five miRNA genomic clusters are indicated with square brackets. Naming of miRNAs was coordinated with the Tuschl and Ambros groups (*25*, *26*).

miRNA gene	Number of clones	miRNA sequence	Length	*C. briggsae* homology	Fold-back arm	Chromosome and distance to nearest gene
lin-4	10	UCCCUGAGAC CUCAAGUGUG A	21	+++	5′	II
let-7	1	UGAGGUAGUA GGUUGUAUAG UU	22	+++	5′	X
mir-1	9	UGGAAUGUAA AGAAGUAUGU A	21	+++	3′	I 3.7 kb from start of T09B4.3, antisense
mir-2	24	UAUCACAGCC AGCUUUGAUG UGC	22–23	+++	3′	I 0.6 kb from start of M04C9.6b
mir-34	3	AGGCAGUGUG GUUAGCUGGU UG	22	+++	5′	X 2.1 kb from end of Y41G9A.4, antisense
mir-35	9	UCACCGGGUG GAAACUAGCA GU	22	+	3′	II 1.3 kb from end of F54D5.12, antisense
mir-36	1	UCACCGGGUG AAAAUUCGCA UG	22	+	3′	II 1.2 kb from end of F54D5.12, antisense
mir-37	2	UCACCGGGUG AACACUUGCA GU	22	++	3′	II 1.1 kb from end of F54D5.12, antisense
mir-38	1	UCACCGGGAG AAAAACUGGA GU	22	+	3′	II 1.0 kb from end of F54D5.12, antisense
mir-39	0	UCACCGGGUG UAAAUCAGCU UG	Predicted	++	3′	II 0.8 kb from end of F54D5.12, antisense
mir-40	2	UCACCGGGUG UACAUCAGCU AA	22	+	3′	II 0.7 kb from end of F54D5.12, antisense
mir-41	2	UCACCGGGUG AAAAAUCACC UA	22	+	3′	II 0.6 kb from end of F54D5.12, antisense
mir-42	1	CACCGGGUUA ACAUCUACAG	20	+++	3′	II 1.2 kb from end of ZK930.2, antisense
mir-43	1	UAUCACAGUU UACUUGCUGU CGC	23	+++	3′	II 1.1 kb from end of ZK930.2, antisense
mir-44	3*	UGACUAGAGA CACAUUCAGC U	21	+++	3′	II 1.0 kb from end of ZK930.2, antisense
mir-45				+++	3′	II 0.7 kb from end of K12D12.1, antisense
mir-46	2	UGUCAUGGAG UCGCUCUCUU CA	22	+++	3′	III 3.0 kb from end of ZK525.1, antisense
mir-47	6	UGUCAUGGAG GCGCUCUCUU CA	22	+++	3′	X 1.8 kb from end of K02B9.2, antisense
mir-48	11	UGAGGUAGGC UCAGUAGAUG CGA	22–24	+++	5′	V 6.1 kb from start of Y49A3A.4
mir-49	1	AAGCACCACG AGAAGCUGCA GA	22	+++	3′	X 2.7kb from end of F19C6.1, antisense
mir-50	2	UGAUAUGUCU GGUAUUCUUG GGUU	24	++	5′	I in intron of Y71G12B.11a
mir-51	6	UACCCGUAGC UCCUAUCCAU GUU	23	++	5′	IV 0.4 kb from end of F36H1.6, antisense
mir-52	47	CACCCGUACA UAUGUUUCCG UGCU	22–25	+++	5′	IV 4.6 kb from end of Y37A1B.6, antisense
mir-53	2	CACCCGUACA UUUGUUUCCG UGCU	24	–	5′	IV 1.9 kb from end of F36H1.6, antisense
mir-54	2	UACCCGUAAU CUUCAUAAUC CGAG	24	+	3′	X 5.5 kb from end of F09A5.2, antisense
mir-55	5	UACCCGUAUA AGUUUCUGCU GAG	23	+	3′	X 5.3 kb from end of F09A5.2, antisense
mir-56	5	UACCCGUAAU GUUUCCGCUG AG	22	+	3′	X 5.2 kb from end of F09A5.2, antisense
mir-56	2	UGGCGGAUCC AUUUUGGGUU GUA	23	+	5′	X 5.2 kb from end of F09A5.2, antisense
mir-57	9	UACCCUGUAG AUCGAGCUGU GUGU	24	+++	5′	II 0.9 kb from start of AF187012-1.T09A5
mir-58	31	UGAGAUCGUU CAGUACGGCA AU	21–22	+++	3′	I Vin intron of Y67D8A.1
mir-59	1	UCGAAUCGUU UAUCAGGAUG AUG	23	+	3′	IV 1.8 kb from start of B0035.1a, antisense
mir-60	1	UAUUAUGCAC AUUUUCUAGU UCA	23	++	3′	II 1.5 kb from end of C32D5.5
mir-61	1	UGACUAGAAC CGUUACUCAU C	21	+	3′	V 0.4 kb from end of F55A11.3, antisense
mir-62	1	UGAUAUGUAA UCUAGCUUAC AG	22	+++	3′	X in intron of T07C5.1
mir-63	1	UAUGACACUG AAGCGAGUUG GAAA	24	–	3′	X 1.7 kb from start of C16H3.2, antisense
mir-64	2	UAUGACACUG AAGCGUUACC GAA	23	–	5′	III 0.25 kb from start of Y48G9A.1
mir-65	0	UAUGACACUG AAGCGUAACC GAA	Predicted	+	5′	III 0.10 kb from start of Y48G9A.1
mir-66	10	CAUGACACUG AUUAGGGAUG UGA	23–24	–	5′	III in coding sequence of Y48G9A.1
mir-67	2	UCACAACCUC CUAGAAAGAG UAGA	24	+++	3′	III 4.4 kb from end of EGAP1.1
mir-68	1	UCGAAGACUC AAAAGUGUAG A	21	–	3′	IV 3.3 kb from start of Y51H4A.22
mir-69	0	UCGAAAAUUA AAAAGUGUAG A	Predicted	–	3′	IV 0.6 kb from start of Y41D4B.21, antisense
mir-70	1	UAAUACGUCG UUGGUGUUUC CAU	23	+	3′	V in intron of T10H9.5
mir-71	5	UGAAAGACAU GGGUAGUGA	19, 20, 22	+++	5′	I 7.8 kb from start of M04C9.6b
mir-72	9	AGGCAAGAUG UUGGCAUAGC	20, 21, 23	–	3′	II 0.21 kb from end of F53G2.4, antisense
mir-73	2	UGGCAAGAUG UAGGCAGUUC AGU	23	++	3′	X 2.9 kb from start of T24D8.6, antisense
mir-74	7	UGGCAAGAAA UGGCAGUCUA CA	22	++	3′	X 3.2 kb from start of T24D8.6, antisense
mir-75	2	UUAAAGCUAC CAACCGGCUU CA	22	++	3′	X 3.5 kb from start of F47G3.3
mir-76	1	UUCGUUGUUG AUGAAGCCUU GA	22	++	3′	III 3.0 kb from start of C44B11.3, antisense
mir-77	1	UUCAUCAgGC CAUAGCUGUC CA	22	+++	3′	II 1.5 kb from start of T21B4.9, antisense
mir-78	2	UGGAGGCCUG GUUGUUUGUG C	21	–	3′	IV 2.0 kb from start of Y40H7A.3, antisense
mir-79	1	AUAAAGCUAG GUUACCAAAG CU	22	+++	3′	I 2.3 kb from end of C12C8.2
mir-80	25	UGAGAUCAUU AGUUGAAAGC CGA	23	+++	3′	III 4.7 kb from end of F44E2.2, antisense
mir-81	7	UGAGAUCAUC GUGAAAGCUA GU	22	+++	3′	X in intron of T07D1.2, antisense
mir-82	6	UGAGAUCAUC GUGAAAGCCA GU	22	+++	3′	X 0.11 kb from start of T07D1.2
mir-83	1	UAGCACCAUA UAAAUUCAGu AA	22	++	3′	IV 5.0 kb from start of C06A6.2
mir-84	3	UGAGGUAGUA UGUAAUAUUG UA	22, 24	+	5′	X 0.8 kb from end of B0395.1, antisense
mir-85	1	UACAAAGUAU UUGAAAAGUC GUGC	24	++	3′	II in intron of F49E12.8, antisense
mir-86	6	UAAGUGAAUG CUUUGCCACA GUC	23	+++	5′	III in intron of Y56A3A.7

*Because *mir-44* and *mir-45* encode identical miRNAs, the three clones represent either or both genes.

clustered within an 800–base pair (800-bp) fragment of chromosome II (Table 1). Computer folding readily identified the fold-back structures for the six cloned miRNAs of this cluster, and predicted the existence of a seventh paralog, miR-39 (Fig. 1D). Northern analysis confirmed the presence and expression of miR-39 (Fig. 3). The homologous cluster in *C. briggsae* appears to have eight related miRNAs. Some of the miRNAs in the *C. elegans* cluster are more similar to each other than to those of the *C. briggsae* cluster and vice versa, indicating that the size of the cluster has been quite dynamic over a short evolutionary interval, with expansion and perhaps also contraction since the divergence of these two species.

Northern analysis of the miRNAs of the *mir-35–mir-41* cluster showed that these miRNAs are highly expressed in the embryo and in young adults (with eggs), but not at other developmental stages (Fig. 3). For the six detectable miRNAs of this cluster, longer species with mobilities expected for the respective fold-back RNAs also appear to be expressed in the germ line; these longer RNAs were observed in wild-type L4 larvae (which have proliferating germ cells) but not in germ line–deficient mutant animals (Fig. 3) (*30*).

The close proximity of the miRNA genes within the *mir-35–mir-41* cluster (Fig. 1D) suggests that they are all transcribed and processed from a single precursor RNA, an idea supported by the coordinate expression of these genes (Fig. 3). This operon-like organization and expression brings to mind several potential models for miRNA action. For example, each miRNA of the operon might target a different member of a gene family for translational repression. At the other extreme, they all might converge on the same target, just as *lin-4* and *let-7* RNAs potentially converge on the 3′ UTR of *lin-14* (*3*).

Another four clusters were identified among the sequenced miRNA clones (Table 1). Whereas the clones from one cluster were not homologous to clones from other clusters, the clones within each cluster were usually related to each other, as seen with the *mir-35–mir-41* cluster. The last miRNA of the *mir-42–mir-44* cluster is also represented by a second gene, *mir-45*, which is not part of the cluster. This second gene appears to enable more constitutive expression of this miRNA (miR-44/45) as compared with the first two genes of the *mir-42–mir-44* cluster, which are expressed predominantly in the embryo (Fig. 3).

Dicer processing of stRNAs differs from that of siRNAs in its asymmetry: RNA from only one arm of the fold-back precursor accumulates, whereas the remainder of the precursor quickly degrades (*15*). This asymmetry extends to nearly all the miRNAs. For the 35 miRNAs yielding more than one clone, RNAs were cloned from both arms of a hairpin in only one case, miR-56 (Fig. 1C and Table 1). The functional miRNA appears to be miR-56 and not miR-56*, as indicated by sequence conservation between *C. elegans* and *C. briggsae* orthologs, analogy to the other constituents of the *mir-54–mir-56* cluster, and Northern blots detecting RNA from only the 3′ arm of the fold-back (*30*).

We were surprised to find that few, if any, of the cloned RNAs had the features of siRNAs. No *C. elegans* clones matched the antisense of annotated coding regions. Of the 30 *C. elegans* clones not classified as miRNAs, 15 matched fragments of known RNA genes, such as transfer RNA (tRNA) and ribosomal RNA. Of the remaining 15 clones, the best candidate for a natural siRNA is GGAAAACGGGUUGAAAGGGA. It was the only *C. elegans* clone perfectly complementary to an annotated EST, hybridizing to the 3′ UTR of gene ZK418.9, a possible RNA-binding protein. Even if this and a few other clones do represent authentic siRNAs, they would still be greatly outnumbered by the 300 clones representing 54 different miRNAs. Our cloning protocol is not expected to preferentially exclude siRNAs; it was similar to the protocol that efficiently cloned exogenous siRNAs from *Drosophila* extracts (*12*). Instead, we propose that the preponderance of miRNAs among our clones indicates that in healthy, growing cultures of *C. elegans*, regulation by miRNAs normally plays a more dominant role than does regulation by siRNAs.

Regardless of the relative importance of miRNAs and siRNAs in the normal regulation of endogenous genes, our results show that small RNA genes like *lin-4* and *let-7* are more abundant in *C. elegans* than previously appreciated. Results from a parallel effort that directly cloned small RNAs from *Drosophila* and HeLa cells demonstrates that the same is true in other animals (*25*), a conclusion further supported by the orthologs to the *C. elegans* miRNAs that we identified through database searching. Many of the miRNAs that we identified are represented by only a single clone (Table 1), suggesting that our sequencing has not reached saturation and that there are over 100 miRNA genes in *C. elegans*.

We presume that there is a reason for the expression and evolutionary conservation of these small noncoding RNAs. Our favored hypothesis is that these newly found miRNAs, together with *lin-4* and *let-7* RNAs, constitute an important and abundant class of riboregulators, pairing to specific sites within mRNAs to direct the posttranscriptional regulation of these genes (*32*). The abundance and diverse expression patterns of miRNA genes implies that they function in a variety of regulatory pathways, in addition to their known role in the temporal control of developmental events.

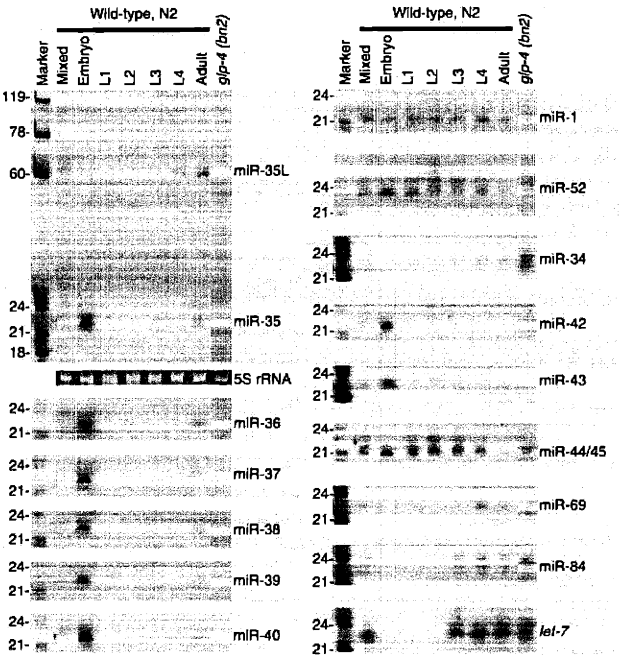


Fig. 3. Expression of newly found miRNAs and *let-7* RNA during *C. elegans* development. Northern blots probed total RNA from mixed-stage worms (Mixed), worms staged as indicated, and *glp-4 (bn2)* adult worms (*24*). Specificity controls ruled out cross-hybridization among probes for miRNAs from the *mir-35–mir-41* cluster (*24*). Other blots indicate that, miR-46 or -47, miR-56, miR-64 or -65, miR-66, and miR-80 are expressed constitutively throughout development (*30*).

References and Notes

1. R. C. Lee, R. L. Feinbaum, V. Ambros, *Cell* **75**, 843 (1993).
2. B. Wightman, I. Ha, G. Ruvkun, *Cell* **75**, 855 (1993).
3. B. J. Reinhart *et al.*, *Nature* **403**, 901 (2000).
4. E. G. Moss, R. C. Lee, V. Ambros, *Cell* **88**, 637 (1997).
5. F. J. Slack *et al.*, *Mol. Cell* **5**, 659 (2000).
6. A. J. Hamilton, D. C. Baulcombe, *Science* **286**, 950 (1999).
7. S. M. Hammond, E. Bernstein, D. Beach, G. J. Hannon, *Nature* **404**, 293 (2000).

8. P. D. Zamore, T. Tuschl, P. A. Sharp, D. P. Bartel, *Cell* **101**, 25 (2000).
9. S. Parrish, J. Fleenor, S. Xu, C. Mello, A. Fire, *Mol. Cell* **6**, 1077 (2000).
10. D. Yang, H. Lu, J. W. Erickson, *Curr. Biol.* **10**, 1191 (2000).
11. C. Cogoni, G. Macino, *Curr. Opin. Genet. Dev.* **10**, 638 (2000).
12. S. M. Elbashir, W. Lendeckel, T. Tuschl, *Genes Dev.* **15**, 188 (2001).
13. E. Bernstein, A. A. Caudy, S. M. Hammond, G. J. Hannon, *Nature* **409**, 363 (2001).
14. A. Grishok *et al.*, *Cell* **106**, 23 (2001).
15. G. Hutvágner *et al.*, *Science* **293**, 834 (2001).
16. F. G. Ratcliff, S. A. MacFarlane, D. C. Baulcombe, *Plant Cell* **11**, 1207 (1999).
17. R. F. Ketting, T. H. Haverkamp, H. G. van Luenen, R. H. Plasterk, *Cell* **99**, 133 (1999).
18. H. Tabara *et al.*, *Cell* **99**, 123 (1999).
19. S. Malinsky, A. Bucheton, I. Busseau, *Genetics* **156**, 1147 (2000).
20. A. A. Aravin *et al.*, *Curr. Biol.* 11, 1017 (2001).
21. A. Huttenhofer *et al.*, *EMBO J.* **20**, 2943 (2001).
22. K. M. Wassarman, F. Repoila, C. Rosenow, G. Storz, S. Gottesman, *Genes Dev.* **15**, 1637 (2001).
23. Short endogenous *C. elegans* RNAs were cloned using a protocol inspired by Elbashir *et al.* (*12*), but modified to make it specific for RNAs with 5′-terminal phosphate and 3′-terminal hydroxyl groups. In our protocol (*24*), gel-purified 18–26 nt RNA from mixed-stage worms was ligated to a pre-adenylylated 3′-adaptor oligonucleotide in a reaction using T4 RNA ligase but without adenosine triphosphate (ATP). Ligated RNA was gel-purified, then ligated to a 5′-adaptor oligonucleotide in a standard T4 RNA ligase reaction. Products from the second ligation were gel-purified, then reverse transcribed and amplified by using the primers corresponding to the adaptor sequences. To achieve ligation specificity for RNA with a 5′-terminal phosphate and 3′-terminal hydroxyl, phosphatase and phosphorylase treatments, useful for preventing circularization of Dicer products (*12*), were not included in our protocol. Instead, circularization was avoided by using the pre-adenylylated 3′-adaptor oligonucleotide and omitting ATP during the first ligation reaction.
24. Supplemental material describing methods and predicted fold-back secondary structures for the miRNAs of Table 1 and some of their homologs in other species is available on *Science* Online at www.sciencemag.org/cgi/content/full/294/5543/858/DC1.
25. M. Lagos-Quintana, R. Rauhut, W. Lendeckel, T. Tuschl, *Science* **294**, 853 (2001).
26. R. C. Lee, V. Ambros, *Science* **294**, 862 (2001).
27. *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
28. M. J. Beanan, S. Strome, *Development* **116**, 755 (1992).
29. S. W. Knight, B. L. Bass, *Science* **293**, 2269 (2001).
30. N. C. Lau, L. P. Lim, E. Weinstein, D. P. Bartel, data not shown.
31. A. E. Pasquinelli *et al.*, *Nature* **408**, 86 (2000).
32. This begs the question as to why more riboregulators have not been found previously. Perhaps they had not been identified biochemically because of a predisposition towards searching for protein rather than RNA factors. They could be identified genetically, which was how *lin-4* and *let-7* were discovered (*1–3*); however, when compared to mutations in protein-coding genes, point substitutions in these short RNA genes would be less likely and perhaps less disruptive of function. Furthermore, mutations that map to presumed intergenic regions with no associated RNA transcript detectable on a standard RNA blot might be put aside in favor of other mutants.
33. WormBase is available on the Web at www.wormbase.org.
34. Sequencing traces (from the Sanger Center) representing 2.5- to 3-fold average coverage of the *C. briggsae* genome were obtained at www.ncbi.nlm.nih.gov/Traces.
35. I. L. Hofacker *et al.*, *Monatsh. Chemie* **125**, 167 (1994).

36. We thank C. Ceol for guidance in culturing and staging *C. elegans*; R. Horvitz for the use of equipment and facilities; T. Tuschl, G. Ruvkun, B. Reinhart, A. Pasquinelli, C. Burge, F. Lewitter, A. Ensminger, and C. Mello for helpful discussions; P. Zamore, T. Orr-manuscript; and T. Tuschl, V. Ambros, G. Ruvkin, R. Horvitz, P. Sharp, and J. Hodgkin for discussions on nomenclature.

3 August 2001; accepted 14 September 2001

An Extensive Class of Small RNAs in *Caenorhabditis elegans*

Rosalind C. Lee and Victor Ambros*

The *lin-4* and *let-7* antisense RNAs are temporal regulators that control the timing of developmental events in *Caenorhabditis elegans* by inhibiting translation of target mRNAs. *let-7* RNA is conserved among bilaterian animals, suggesting that this class of small RNAs [microRNAs (miRNAs)] is evolutionarily ancient. Using bioinformatics and cDNA cloning, we found 15 new miRNA genes in *C. elegans*. Several of these genes express small transcripts that vary in abundance during *C. elegans* larval development, and three of them have apparent homologs in mammals and/or insects. Small noncoding RNAs of the miRNA class appear to be numerous and diverse.

Small RNAs perform diverse functions within cells, including the regulation of gene expression (*1–4*). One class of regulatory RNA includes the small temporal RNA (stRNA) products of the genes *lin-4* and *let-7* in *Caenorhabditis elegans*. The *lin-4* and *let-7* RNAs are ~22 nucleotides (nt) in length, and are expressed stage-specifically, controlling key developmental transitions in worm larvae by acting as antisense translational repressors (*2–4*).

Dartmouth Medical School, Department of Genetics, Hanover, NH 03755, USA.

*To whom correspondence should be addressed. E-mail: vambros@dartmouth.edu

lin-4 and *let-7* were identified by their mutant phenotypes (*2, 3*) and, until recently, were the only known RNAs of their class. However, the phylogenetic conservation of *let-7* RNA sequence and developmental expression (*5*), and the overlap between the stRNA and RNA interference (RNAi) pathways (*6, 7*), suggested that stRNAs are part of an ancient regulatory mechanism involving ~22-nt antisense RNA molecules (*8*).

To identify more small regulatory RNAs of the *lin-4/let-7* class in *C. elegans*, we used informatics and cDNA cloning to select *C. elegans* genomic sequences that exhibited four characteristics of *lin-4* and *let-7*: (i) expression of a mature RNA of ~22 nt in

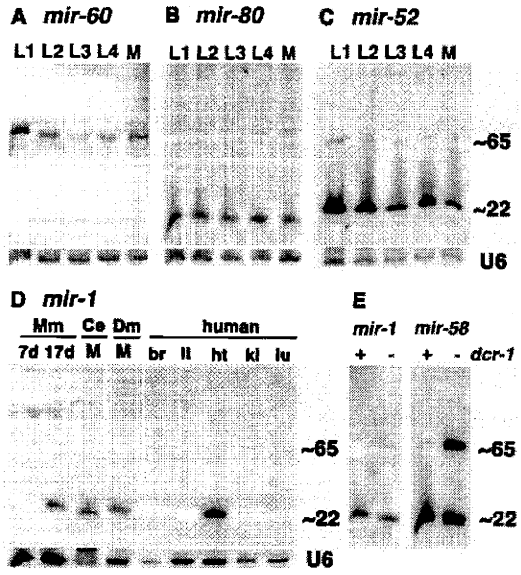


Fig. 1. Northern blots of small RNA transcripts. (A through C) Total RNA from *C. elegans* larvae (stages L1 through L4) or from mixed stage (M) populations were blotted and probed with oligonucleotides complementary to either the 5′ or 3′ half of the indicated transcript (*13*). U6 = the same filters were probed with probe to U6 snRNA as a loading control. (**A**) *mir-60* 5′ probe detects a transcript of ~65 nt. The ratio of L1 to L4 *mir-60* signal, normalized to U6, is about 5:1. The *mir-60* 3′ probe (not shown) detects a similar-sized species with a similar developmental profile. (**B**) *mir-80* 3′ probe detects a ~22-nt RNA expressed uniformly at all stages. (**C**) *mir-52* 5′ probe. The normalized *mir-52* signal is threefold greater in the L1 versus the L3. (**D**) *mir-1* 3′ probe detects a transcript of ~22 nt in total RNA from mouse (Mm) 17-day embryos, mixed-stage *C. elegans* (Ce), *Drosophila melanogaster* (Dm) mixture of embryo-larvae-pupae, and in a sample of human heart (ht) tissue. Other human tissue samples were brain (br), liver (li), kidney (ki), and lung (lu). (**E**) *mir-1* and *mir-58* probes to total RNA from mixed populations of wild-type (+) and *dcr-1(ok247)* (–) animals. An increase in the proportion of unprocessed ~65-nt precursor is observed in the *dcr-1* RNA.