The C. elegans Heterochronic Gene lin-4 Encodes Small RNAs with Antisense Complementarity to lin-14

Rosalind C. Lee,* 1 Rhonda L. Feinbaum,* 1 and Victor Ambros1
Harvard University
Department of Cellular and Developmental Biology
Cambridge, Massachusetts 02138

Summary

lin-4 is essential for the normal temporal control of diverse postembryonic developmental events in C. elegans. lin-4 acts by negatively regulating the level of LIN-14 protein, creating a temporal decrease in LIN-14 protein starting in the first larval stage (L1). We have cloned the C. elegans lin-4 locus by chromosomal walking and transformation rescue. We used the C. elegans clone to isolate the gene from three other Caenorhabditis species; all four Caenorhabditis clones functionally rescue the lin-4 null allele of C. elegans. Comparison of the lin-4 genomic sequence from these four species and site-directed mutagenesis of potential open reading frames indicated that lin-4 does not encode a protein. Two small lin-4 transcripts of approximately 22 and 61 nt were identified in C. elegans and found to contain sequences complementary to a repeated sequence element in the 3′ untranslated region (UTR) of lin-14 mRNA, suggesting that lin-4 regulates lin-14 translation via an antisense RNA-RNA interaction.

Introduction

A genetic pathway of heterochronic genes in Caenorhabditis elegans acts to specify the temporal fates of cells during larval development, thereby controlling the timing and sequence of events in diverse postembryonic cell lineages (Ambros and Horvitz, 1984; Ambros and Horvitz, 1987; Ambros, 1989). Mutations in the heterochronic genes can cause either precocious development, in which normally late developmental programs are expressed at early larval stages, or retarded development, in which normally early developmental programs are reiteratcd at later stages (Chalfie et al., 1981; Ambros and Horvitz, 1984). It is likely that the expression of stage-specific developmental programs by particular cells requires the accurate temporal regulation of the products of the heterochronic genes.

lin-4 acts early in C. elegans larval development to affect the timing of developmental events at essentially all larval stages and in diverse cell types (Chalfie et al., 1981; Ambros and Horvitz, 1987). Animals carrying a lin-4 loss-of-function (lf) mutation, lin-4(e912), display reiterations of early fates at inappropriately late developmental stages; cell lineage patterns normally specific for the L1 are reiterated at later stages, and the animals execute extra larval molts (Chalfie et al., 1981). The consequences of these heterochronic developmental patterns include the absence of adult structures (such as adult cuticle and the vulva) and the prevention of egg laying.

lin-14 null (0) mutations cause a phenotype opposite to that of lin-4(0) and are completely epistatic to lin-4(0), which is consistent with lin-4 acting as a negative regulator of lin-14 (Ambros and Horvitz, 1987; Ambros, 1989). lin-14(0) mutants skip the expression of L1-specific events and precociously execute programs normally specific for the L2, L3, L4, and adult stages. lin-14 gain-of-function (gf) mutations, which cause inappropriately high lin-14 activity at late stages of development, result in a retarded phenotype virtually identical to that of lin-4(0) (Ambros and Horvitz, 1987). These observations indicate that in wild-type development a high level of lin-14 activity in the early L1 stage specifies L1-specific programs, and lower levels of lin-14 activity in the late L1 specify later stage-specific programs. Thus, the normal developmental progression from the execution of L1 programs to later programs depends critically on the lin-4-dependent decrease in lin-14 activity.

The temporal decrease in lin-14 activity reflects a decrease in the level of LIN-14 protein. LIN-14 protein is normally abundant in the nuclei of late-stage embryos and younger L1 larvae and then is barely detectable by the L2 (Ruvkun and Giusto, 1989). lin-14 transcripts are constant throughout development, indicating that lin-14 is negatively regulated posttranscriptionally (Wightman et al., 1993 [this issue of Cell]). In lin-4 mutant animals, as in lin-14(gf) mutants, the level of LIN-14 protein remains abnormally high late in development (Aranu et al., 1991). Mapping of the lin-14(gf) mutations to the 3′ UTR of lin-14 mRNA (Wightman et al., 1991), and gene fusion experiments (Wightman et al., 1993), define the lin-14 3′ UTR as a necessary and sufficient cis-negative regulatory element of LIN-14 protein level. The temporal decrease in LIN-14 protein levels requires both lin-4 in trans and lin-14 3′ UTR sequences in cis. These observations suggest that the lin-4 gene product or products act via the lin-14 3′ UTR to inhibit, directly or indirectly, the translation of lin-14 mRNA. To determine the mechanism by which lin-4 developmentally regulates the level of LIN-14 protein, we cloned the lin-4 locus by chromosomal walking and transformation rescue. Our analysis of the lin-4 genomic sequence indicates that lin-4 does not encode a protein product. We have identified two small lin-4 transcripts of approximately 22 and 61 nt. These lin-4 RNAs are complementary to a repeated sequence in the 3′ UTR of lin-14 (Wightman et al., 1991, 1993), supporting a model in which these lin-4 RNAs could regulate lin-14 translation by an antisense mechanism.

*A The first two authors contributed equally to this work.
†Present address: Dartmouth College Department of Biology, Hanover, New Hampshire 03755.
‡Present address: Massachusetts General Hospital Department of Molecular Biology, Boston, Massachusetts 02114.
Recombination distance between lin-4 and mapP12 is based on RFLP mapping as described in Experimental Procedures. The lin-4-dpy-10 and lin-4-bli-3 distances are approximately 0.5% and 2% recombination, respectively (Wood et al., 1988). YAC (Y prefix) and cosmid (T, 6, Z, and C prefixes) clones are from the C. elegans physical mapping project (Coulson et al., 1986, 1988) and were obtained from A. Coulson. Genomic clone pVT2D was identified by hybridization to Y18C1, as described in Experimental Procedures and in the text. Derivatives of pVT2D were constructed by treatment with restriction enzymes or ExoIII nuclease (see Experimental Procedures). Dotted lines indicate a change in scale. Plus: the cloned DNA rescued the mutant phenotype of lin-4(e972), as described in Experimental Procedures. Minus: no rescue was observed in any of at least three transgenic lines. Nonrescuing lines were confirmed to contain the plasmid DNA by PCR. Plus/minus: weak rescue was observed, i.e., the appearance of vulval morphogenesis and lateral alae on the adult cuticle was observed, but without complete rescue of egg-laying defects. The e972 lesion deletes most of pVT2D sequences. One breakpoint was determined to lie between two closely positioned BamHI and EcoRI sites at one end of pVT2D by Southern blot analysis (Figure 2). The e972 deletion also extends to sequences corresponding to two other clones, pVT6G and pVT1C (data not shown). Presumably, pVT6G and pVT1C lie to the left of pVT2D, as drawn here. The orientation of pVT2D with respect to the genetic and physical maps has not been established.

Figure 1. Partial Genetic and Physical Map of the lin-4 Region of LGII
Recombination distance between lin-4 and mapP12 is based on RFLP mapping as described in Experimental Procedures. The lin-4-dpy-10 and lin-4-bli-3 distances are approximately 0.5% and 2% recombination, respectively (Wood et al., 1988). YAC (Y prefix) and cosmid (T, B, Z, and C prefixes) clones are from the C. elegans physical mapping project (Coulson et al., 1986, 1988) and were obtained from A. Coulson. Genomic clone pVT2D was identified by hybridization to Y18C1, as described in Experimental Procedures and in the text. Derivatives of pVT2D were constructed by treatment with restriction enzymes or ExoIII nuclease (see Experimental Procedures). Dotted lines indicate a change in scale. Plus: the cloned DNA rescued the mutant phenotype of lin-4(e972), as described in Experimental Procedures. Minus: no rescue was observed in any of at least three transgenic lines. Nonrescuing lines were confirmed to contain the plasmid DNA by PCR. Plus/minus: weak rescue was observed, i.e., the appearance of vulval morphogenesis and lateral alae on the adult cuticle was observed, but without complete rescue of egg-laying defects. The e972 lesion deletes most of pVT2D sequences. One breakpoint was determined to lie between two closely positioned BamHI and EcoRI sites at one end of pVT2D by Southern blot analysis (Figure 2). The e972 deletion also extends to sequences corresponding to two other clones, pVT6G and pVT1C (data not shown). Presumably, pVT6G and pVT1C lie to the left of pVT2D, as drawn here. The orientation of pVT2D with respect to the genetic and physical maps has not been established.

Figure 2. Southern Blot of Genomic DNAs from Wild-Type and lin-4(e912) Animals Digested with Various Restriction Endonucleases and Hyridized with pVT2D Probe
Similar amounts of wild-type and lin-4(e912) DNA were loaded. For each enzyme, the hybridization in e912 DNA is to a band of weaker intensity and of altered size (smaller for each enzyme except Clal) than that in wild-type DNA, consistent with a deletion in e912 DNA of sequences corresponding to the pVT2D probe.

Results
Identification of C. elegans lin-4 Genomic DNA
lin-4 is located on the C. elegans genetic map in a region of LGII between bli-3 and dpy-10 (Wood et al., 1988) (Figure 1). We mapped the positions of restriction fragment length polymorphisms (RFLPs) in the bli-3-dpy-10 interval with respect to lin-4, as described in Experimental Procedures. Among the RFLPs that we analyzed, mapP12, detected by cosmid ZK459, mapped closest to lin-4, and to the right (Figure 1). We used cosmid and yeast artificial chromosome (YAC) clones to the left of ZK459 as probes to southern blots of wild-type and lin-4 DNA. The YAC clone Y18C1 detected a restriction fragment alteration associated with lin-4(e912) DNA. This lesion, which results in the absence of a 5 kb EcoRI hybridizing band, was not detected by an overlapping YAC clone, Y42A4.

We cloned the wild-type sequences corresponding to the e912-affected 5 kb band based on its differential hybridization to Y18C1 and Y42A4, as described in Experimental Procedures. Three classes of clones were identified, each corresponding to a unique EcoRI insert: pVT2D (containing a 5 kb insert), pVT1C (a distinct 5 kb insert), and pVT6G (a 3.5 kb insert). Probes from pVT2D (Figure 2), pVT1C, pVT6G, and an overlapping cosmid clone, C02B6 (data not shown), detected restriction fragment aberrations on southern blots of lin-4(e912) DNA, indicating that the e912 lesion must extend over several kb of DNA. We have not characterized the e912 lesion in detail, except to determine that sequences corresponding to both pVT2D and pVT1C are at least partially deleted, and that the pVT6G insert fragment seems to be rearranged and possibly duplicated (data not shown; see legend to Figure 1).

Functional lin-4 sequences were localized to pVT2D by transformation rescue. pVT2D, pVT2DCla (a 3.2 kb subclone of pVT2D), and C02B6 (a cosmid clone that overlaps pVT2D) rescued the mutant phenotype of lin-4(e912) animals in transgenic lines generated by microinjection of the cloned DNA (Figure 1; see also Experimental Procedures). These results indicate that the lin-4 gene lies within the 3.2 kb EcoRI/C1al insert fragment of pVT2DCla. The e912 lesion is a deletion of most of the sequences of pVT2D, including about half of the sequences in pVT2DCla (Figure 1). Subclones and ExoIII deletion derivatives of pVT2DCla (see Experimental Procedures) were tested for rescue of lin-4(e912) defects in transgenic animals. A subclone (pVTSal3) containing a 693 bp SalI fragment (see...
Developmental Regulation via Natural Antisense RNA

Figure 3. Sequence Analysis of lin-4

(A) Comparison of lin-4 sequences from four Caenorhabditis species: C. elegans, C. briggsae, C. remanei, and C. vulgaris. Genomic lin-4 clones were identified and prepared as described in Experimental Procedures. DNA sequence was determined for both strands of each species; only one strand is shown here. Numbering refers to the 693 bp SalI insert of the C. elegans rescuing clone pVTSal3 (see Figure 1). The right end of the C. remanei sequences shown here is the EcoRI site at the end of the C. remanei genomic DNA insert in pVTreA1. The right ends of the C. briggsae and C. vulgaris sequences are defined by sequence alignment with the end of the C. elegans pVTSal3 clone. Dots indicate absence of nucleotides. Sequences conserved in all four species are shown in bold. Note the conserved TTTCAG beginning at nucleotide 662, which is likely the 3' splice acceptor for an exon of the host gene within which lin-4 resides. The eight nucleotides that follow are the only coding sequences of the host gene on the pVTSal3 fragment. Sequences complementary to a repeated sequence in the 3' UTR of lin-14 mRNA (Wightman et al., 1991, 1993) are boxed. Lin-4S and Lin-4L transcripts are indicated according to the results of nuclease protection, primer extension, and Northern blot experiments (see Figures 4, 5, 6, and 7). The position of the ethylmethane sulfonate-induced C-to-T change associated with ma-761 is shown by an arrow below the line. The positions of mutations that were induced in vitro and retain lin-4 function (see Results and Experimental Procedures) are shown above the C. elegans line. In separate experiments, insertions of 61 and 1 bp were introduced at positions 90 and 259, respectively. Triplets that were converted to stop codons are shown by bars at positions 52 and 109. A T-to-C mutation introduced at position 547 is indicated by a C above the C. elegans sequence.

(B) Predicted open reading frames (in all six frames) for the wild-type lin-4 sequences of four Caenorhabditis species. Short bars are predicted AUG codons, and tall bars are predicted stop codons. The top three lines and bottom three lines for each species correspond to the three rightward and three leftward reading frames, respectively.

Phylogenetic Comparison of lin-4 Genomic Sequence

C. briggsae and C. remanei lin-4 clones were identified by hybridization to a C. elegans lin-4 probe, and a C. vulgaris lin-4 clone was constructed using polymerase chain reaction (PCR)-amplified genomic DNA as described in Experimental Procedures. The C. briggsae, C. remanei, and C. vulgaris clones were tested for rescue of the mutant phenotypes of lin-4(e912) and were found to be fully functional for rescue in C. elegans. It is likely that the clones from C. briggsae, C. remanei, and C. vulgaris correspond to the lin-4 locus of each species, because they all function in C. elegans, and because lin-4 seems to be a single copy gene in C. elegans (Figure 2).

The C. elegans lin-4 genomic clone pVT2DCla and the C. briggsae, C. remanei, and C. vulgaris lin-4 clones were sequenced as described in Experimental Procedures. Among the four Caenorhabditis species, two main blocks of phylogenetically conserved DNA sequence were observed within the lin-4 rescuing (pVTSal) region, one from approximately base pair 85 to base pair 350 of pVTSal3, and another from about base pair 500 to base pair 600 (Figure 3A). The sequence conservation consists of stretches of complete identity, interspersed with short stretches of divergence, and insertions and deletions of 1 or more bp. Outside the sequences corresponding to the 693 bp pVTSal3 insert, the sequences of these four species are less well conserved (data not shown). The fact that the lin-4 clones from four Caenorhabditis species function in C. elegans indicates that they encode similar functionally related gene products.

lin-4 Is Unlikely to Encode a Protein

A probe from pVT2DCla was used to screen a cDNA library. The clones that were isolated and sequenced all contained DNA corresponding to genomic sequences that were almost entirely outside the 693 bp SalI (pVTSal) rescuing region. Comparison of these cDNA sequences with
pVT2D sequences indicated that the corresponding mRNA is spliced from a primary transcript at conventional splice donor and acceptor sites within pVT2D. Therefore, it appears that the lin-4 gene lies within an intron of another gene. The normal function of the host gene is unknown but is apparently unrelated to lin-4 function; pVTSal fully rescues the lin-4(e912) mutant phenotype in spite of the fact that the e912 lesion deletes large regions of the host gene outside of pVTSal (data not shown).

To determine whether the lin-4 rescuing region might encode a protein from a rare mRNA not represented in the cDNA library that we screened, we compared the potential protein coding sequences of the lin-4 rescuing region from the four species. These four lin-4 genes contain no conserved protein sequence that begins and ends with conventional start and stop codons or that can be predicted to be assembled using conventional C. elegans splice donor and acceptor sites (Wood et al., 1988) (Figure 3B). One relatively long predicted open reading frame of 143 amino acids in the region from base pair 1 to base pair 429 of the C. elegans sequence was tested by inserting 82 bp at an MluI site (base pair 90) and by inserting one base at a Tth1111 site (base pair 259) (Figure 3A). Similarly, three mutations were introduced by oligonucleotide-mediated in vitro mutagenesis (see Experimental Procedures) that each disrupt this and other putative open reading frames in the C. elegans clone: nonsense codons were created at base pairs 53 and 111, and a conserved ATG at base pair 546 was converted to ACG (Figure 3A). All of these mutant constructs fully rescued lin-4(e912). These observations are inconsistent with a LIN-4 protein product.

Two Small lin-4 Transcripts

The fact that lin-4 negatively regulates lin-14 activity in trans strongly implies a lin-4 gene product. Northern blot and RNAase protection analysis using strand-specific probes (Figure 4) that covered the 693 bp lin-4 rescuing sequence detected two small transcripts (Figure 5). Both transcripts mapped to the same region by Northern blots and RNAase protection experiments (see below) and are transcribed in the same orientation (See Figures 3 and 4). The larger transcript, lin-4L, appears to be 61 nt, and the
Three S1 digestion products of 40, 41, and 42 nt were obtained. The largest product (arrow) corresponded to the same 5' end as that identified by primer extension. The estimated error in the size of the S1-protected RNAs is ± 1 nt. (C) S1 mapping of the 3' end of lin-4L. The rMGH5 probe (Figure 4) was labeled at the 3' end as described in Experimental Procedures. S1 analysis of total RNA from VT510 is shown, but an identical pattern of S1 products was also detected in N2 RNA. The two most abundant species were 44 and 45 nt long; the 45 nt product (arrow) corresponds to a 3' end of lin-4L at the C located at base pair 573 of pVTSa13 (Figures 3 and 4).

Figure 6. Mapping 5' and 3' Ends of lin-4L

(A) Primer extension analysis of total RNA from wild-type N2, lin-4(e912), and the rescued line VT510. rMGH3 (see Figure 4) was used as a primer. An M18 sequencing ladder provides size markers. The estimated error in the size of the extension product (arrow), determined by comparison with the sequencing ladder as well as to labeled oligonucleotide markers (data not shown), is ± 1 nt. The single extended product of 49 nt corresponds to lin-4L, transcription initiating at the T located at position 513 of the C. elegans DNA sequence (Figures 3 and 4).

(B) S1 mapping of the 5' end of lin-4L. The probe used was rMGH1 (Figure 4). S1 analysis of total RNA from VT510 is shown, but an identical pattern of S1 products was also detected in wild-type N2 RNA. The size of 5' end-labeled oligonucleotide markers is indicated to the left.

Primer extension and S1 analysis of total RNA from mixed-stage N2, lin-4 rescued animals, or both was used to determine the 5' end of lin-4L (Figures 6A and 6B). Primer extension experiments gave a single extended product of 49 nt (Figure 6A), indicating that lin-4L transcription initiates at the T at position 513 of the C. elegans DNA sequence (see Figures 3 and 4). The oligonucleotide primer used to map the 5' end of lin-4L (rMGH3, see Figure 4) should not have detected lin-4S transcripts, owing to the limited complementarity (6 bp) between the primer and lin-4S. S1 analysis confirmed the 5' end of lin-4L determined by the above primer extension experiment. Three S1 digestion products of 40, 41, and 42 nt were obtained (Figure 6B). The largest product corresponded to the same 5' end as that identified by primer extension. The smaller bands presumably are due to breathing of the DNA:RNA hybrid during the S1 digestion.

The 3' end of lin-4L was also mapped by S1 analysis. A set of S1 digestion products ranging from 42 to 46 nt was obtained (Figure 6C). The two most abundant species were 44 and 45 nt long; the 45 nt product corresponded to termination of the transcript at the C located at base pair 573 of pVTSa13 (see Figures 3 and 4). The multiple products obtained in this experiment may be in part due to probe heterogeneity caused by incomplete labeling of the four terminal bases of the probe (see Figure 6; also Experimental Procedures) or by breathing of the RNA: DNA hybrid during the S1 digestion, or both. It is also possible that there is natural variation in the 3' end of lin-4L.

lin-4S appears to be more 5' end overlapping lin-4L. RNAase protection experiments using nested probes (Figure 7A) demonstrated that lin-4S spans a Ddel site that is located 5 nt downstream of the 5' end of lin-4L (at T 513). The length of the protected product corresponding to lin-4S is approximately 5 nt shorter in the case of the Ddel probe compared with the EcoRI probe (Figure 7A), indicating that lin-4S starts at the same position as lin-4L.

The 3' end of lin-4S was mapped by S1 analysis. An oligonucleotide probe (rMGH8) labeled at the two 3' terminal A residues that are complementary to the proposed 5' end of both lin-4 transcripts (see Figure 4) was hybridized to total N2 RNA and then digested with S1 nuclease at different temperatures. At 15°C, the S1 digestion produced protected products that ranged from 19–22 nt, while at 37°C, protected products were predominantly 17–21 nt long (see Figure 7B). These mapping data agree with the estimated size for lin-4S of 20 ± 2 nt from Northern blots and RNAase protection experiments (see Figure 5). This suggests that lin-4S is identical in sequence to the first 22 nt of lin-4L (see Figures 3 and 4).

lin-4 Mutations Affect lin-4L and lin-4S

As mentioned previously, neither lin-4 transcript was detectable in the single previously isolated lin-4 mutant, e912. This was not unexpected, since the e912 lesion is a dele-
tion or rearrangement that removes at least 5 kb in the lin-4 region, including the entire lin-4S- and lin-4L-transcribed sequences. To test further the functional significance of the lin-4 transcribed sequences, we used a noncomplementation screen (see Experimental Procedures) to isolate a novel lin-4 mutation and then identified the corresponding molecular lesion. Over 20,000 mutagenized chromosomes were screened, and a single novel lin-4 allele, ma161, was identified. DNA from ma161 animals was amplified by PCR and sequenced. The only sequence alteration in the 693 bp lin-4 region of ma161 DNA was a C to T transition at base pair 517 (see Figure 3). This point mutation would presumably alter nucleotide 5 in lin-4L and lin-4S.

lin-4 Transcripts Are Complementary to the 3' UTR of lin-14 mRNA

The lin-4 transcribed sequence was combined in tandem to the sequence of the lin-14 3' UTR (Wightman et al., 1991), and this sequence was used for the formation of lin-4: lin-14 hybrid RNA structures, using the MULFOLD program of the GCG sequence analysis package (Devereux et al., 1984), as described in Experimental Procedures. Two short blocks of lin-4 sequence were identified (Figure 3A) that are complementary to an element repeated seven times in the 3' UTR of lin-14 (Figure 8; Wightman et al., 1991, 1993). The first block of complementary sequence, 5'-GUCCCGAG-3', is (with the exception of the first G) at the very 5' end of both lin-4L and lin-4S. The second block of sequence, 5'AAUGUGAGA3', would lie internal to lin-4L and at the 3' end of lin-4S, with the 3' G perhaps missing from lin-4S (see Figures 3, 4, and 8).

Discussion

The lin-4 Locus

The C. elegans heterochronic gene lin-4 was identified by chromosomal walking from linked polymorphisms, char-

Figure 7. Mapping 5' and 3' Ends of lin-4s

(A) RNase protection of total RNA from wild-type N2, lin-4(e972), and VT510, using nested antisense probes to map the 5' and 3' ends of lin-4s. Probes were runoff transcripts from pMspl that had been previously digested with either EcoRI, which cuts in the polylinker (T7RI, Figure 4), or Ddel, which cuts within lin-4 sequences (T-7Dde, Figure 4). The size of one RNA marker is indicated to the left. Arrows indicate the probe fragments protected by lin-4s. The length of the protected product is approximately 5 nt longer in the case of the EcoRI probe compared with the Ddel probe, indicating that lin-4s starts 5 nt upstream of the end of the Ddel probe.

(B) S1 analysis of total RNA from wild-type N2 using 3' end-labeled rMGH8 as a probe. The temperature at which the S1 digestion was performed is indicated above each lane. S1 digestion was for 1 hr. The size of 5' end-labeled oligonucleotide markers is indicated to the right.

Figure 8. lin-4 Transcripts and Complementarity between lin-4 and lin-14

(A) Sequences for lin-4L and lin-4S RNAs, and a proposed secondary structure for lin-4L, predicted by the MULFOLD program (see Experimental Procedures). A secondary structure for lin-4S is not shown, owing to the uncertainty of the precise 3' and 5' nucleotides of lin-4S, which significantly affect structural predictions (see text). Sequences complementary to the lin-14 3' UTR (Wightman et al., 1991, 1993) are bold. Nucleotides that differ from the C. elegans sequence in one of the other three species (Figure 3) are indicated by the alternative nucleotide in smaller type. Also in smaller type and indicated by an asterisk is a U-to-C mutation introduced in vitro that retains lin-4 function (see Results and Figure 3A). The position of the ma161 C-to-U loss-of-function mutation is shown by an arrow.

(B) Complementarity between copies of a repeated element in the 3' UTR of lin-14 RNA (Wightman et al., 1993). Dots indicate absence of a nucleotide; dashes indicate one or more noncomplementary nucleotides. Only lin-4:lin-14 complementarity that is conserved between C. elegans and C. briggsae (Wightman et al., 1993) is represented. The lin-4:complementary elements in lin-14 vary somewhat among themselves and contain internal blocks of noncomplementary nucleotides of variable length and sequence. Note the five conserved noncomplementary nucleotides (S' ACCUC) of lin-4 RNA. RNA sequences are deduced from DNA sequence and transcript mapping, and not from direct sequence analysis of RNA.
acterization of the molecular lesion associated with a lin-4 allele, e972, and transformation rescue. lin-4(e972) is most likely a lin-4 null allele, because its lesion removes most of the sequences corresponding to the 693 bp clone that rescues lin-4 function. This observation, together with the fact that lin-14 null alleles are completely epistatic to lin-4 function of h-4 depends on the host gene. This is because genomic clones that contain essentially only intron sequences of the host gene rescue lin-4(e972) extrachromosomally. These rescued animals appear fully wild type, despite the fact that e972 deletes a substantial portion of the host gene in addition to the lin-4 rescuing region (R. C. L., R. L. F., and V. A., unpublished data), indicating that the activity of the host gene is probably not required for lin-4 function or for other aspects of normal development. Although our data indicate that the 693 bp Sall rescuing fragment contains the entire lin-4 locus, all our rescue experiments were performed using high copy extrachromosomal arrays (Mello et al., 1991). Since we have not tested the 693 bp Sall clones in single copy, we cannot rule out the possibility that some additional sequences may be required for full lin-4 activity. Generally, rescue of lin-4 retarded defects by high-copy plasmid arrays was complete, and we did not consistently observe additional developmental defects that might be associated with overexpression of lin-4 from the arrays.

lin-4 Is Unlikely to Encode a Protein

Two lines of evidence strongly suggest that none of the lin-4 genomic sequence encodes protein. First, genetically located frameshift and nonsense mutations, as well as an ATG to ACG mutation, were introduced into several putative open reading frames of the C. elegans genomic clone without affecting rescuing activity. Second, comparison of the C. elegans, C. briggsae, C. remanei, and C. vulgari lin-4 genomic sequences (which are functionally interchangeable in C. elegans) revealed a high level of nucleic acid sequence conservation but no conserved open reading frames. Regions of conserved sequence do not show a degeneracy at the third base positions of putative codons that would be expected for protein coding sequences, and they contain deletions or insertions that would cause frameshifts. There are no splice donor or acceptor sites that would compensate for these frameshifts. Thus, if a conserved protein product were encoded by the lin-4 locus, its synthesis would need to be directed by unorthodox translation start and stop signals. The 693 bp C. elegans fragment rescues in either orientation, arguing against the generation of a lin-4 protein from the rescuing fragment using translational signals from the vector.

lin-4 Encodes Small RNAs

Northern blots and RNase protection experiments with lin-4 genomic probes detected the lin-4S and lin-4L small RNAs in total RNA of wild-type and rescued lin-4(e972) animals. Several observations argue that either lin-4L RNA or lin-4S RNA or both are the functional lin-4 gene products. First, the ma161 mutation, which results in loss of lin-4 function, affects a conserved C of the lin-4L and lin-4S transcripts. Second, the DNA sequence encoding lin-4L and lin-4S is virtually identical in four nematode species (Figure 3A). Finally, the sequence complementarity between the lin-4 transcripts and the lin-14 mRNA (Wightman et al., 1991, 1993) suggests a direct base-pairing interaction between lin-4 and lin-14 RNAs and lends credence to the idea that lin-4S, lin-4L, or both may be the active lin-4 gene product. Regions of very high sequence conservation were identified outside the lin-4L- and lin-4S-transcribed regions of the lin-4-rescuing clones. One of these is located between approximately 140 and 420 bp upstream of the transcribed sequences. This region is essential for lin-4 function, but its orientation is not critical, because when it is inverted with respect to the transcribed region, partial rescuing activity is observed (Figure 1). These upstream conserved regions could contain regulatory sequences or encode additional small RNAs of an abundance too low to be detected in our experiments. No other lin-4 products besides lin-4L and lin-4S were consistently detected in our experiments.

lin 4L and lin 4S may be transcribed from the same promoter, given that their 5' ends appear to be identical. lin-4S could be generated by a posttranscriptional processing of lin-4L, or lin-4L may be produced by read-through of a lin-4S termination site. It is possible that either lin-4L or lin-4S is a nonfunctional byproduct of the biogenesis of the other, or that both are derived by the processing of a longer precursor transcript. lin-4L and lin-4S seem to be synthesized in approximately normal levels in a strain lacking the lin-14 3'UTR (R. L. F., unpublished data), suggesting that the generation of the two products likely occurs independently of an interaction with the lin-14 mRNA. The fact that the length of lin-4S and lin-4L by RNase protection and S1 mapping agree with the lengths estimated by Northern blot analysis indicates that the RNAs are not heavily modified posttranscriptionally (for example by significant splicing, polyadenylation, or covalent linkage to protein). However, these experiments would not necessarily detect slight modification or processing of the RNAs, such as modified nucleotides or cap structures. Further experiments are required to develop a complete description of lin-4 RNA biogenesis and structure.

Antisense Complementarity to the lin-14 3'UTR

Sequence comparison of the lin-4 transcripts with the 3'UTR of lin-14 revealed that a sequence element repeated seven times in the lin-14 3'UTR is complementary to the lin-4 transcripts. The significance of this complementarity to lin-4 function is supported by phylogenetic and genetic observations. First, the seven repeated elements are conserved between the C. elegans and C. briggsae lin-14 sequences (Wightman et al., 1993), consistent with a conserved lin-4:lin-14 RNA base pairing and the observed conservation of lin-4 function between these species. Second, the lin-4(ma161) loss-of-function mutation is located within a block of lin-14-complementary sequences, and so could affect lin-4 activity by altering or destabilizing a lin-4:
lin-14 RNA hybrid. Third, the seven lin-4-complementary elements occur in the lin-14 region deleted by lin-14(gf) alleles (Wightman et al., 1991). These observations strongly support the hypothesis that lin-4 downregulates LIN-14 protein levels through a direct RNA–RNA interaction with the lin-14 3′UTR.

Although the longest stretch of contiguous sequence complementarity between lin-4 and lin-14 is 10 nt (Figure 8; Wightman et al., 1993), there is precedence for specific RNA–RNA interactions by such short regions of complementarity. For example, interactions among spliceosomal RNAs and between spliceosomal RNAs and the pre-mRNAs involve similarly short duplexes (Datta and Weiner, 1991; Madhani and Guthrie, 1992; Wassarman and Steitz, 1992; reviewed by Green, 1991; Guthrie, 1991). Since both lin-4S and lin-4L span the ma161 lesion, and both include sequences complementary to the 3′UTR of lin-14, either could in principle interact with lin-14 RNA. Although our 5′ and 3′ end-mapping experiments indicate that lin-4S might not contain all the lin-14-complementary sequences (Figure 8), the ambiguity in these data (+1 to −2 nt at each end) leaves open the possibility that lin-4S contains the entire lin-14-complementary sequence. Since lin-4S is significantly more abundant than lin-4L, it seems likely that lin-4S plays the major role in base pairing with lin-14 RNA, especially if the interaction is stoichiometric. Furthermore, the predicted secondary structure of lin-4L (Figure 8) would sequester within a stem most of the bases that are complementary to lin-14, perhaps rendering lin-4L inactive for base-pairing with lin-14 mRNA. We cannot evaluate how the secondary structure of lin-4S might affect its function, since the precise 5′ and 3′ ends of lin-4S are uncertain, and this variability would significantly alter any proposed secondary structure (data not shown; see also legend to Figure 8).

A definitive proof of the proposed antisense pairing between lin-4 and lin-14 RNAs will require the construction of compensatory base-pairing mutations in lin-14 and lin-4. Furthermore, determination of whether the interaction involves a stable complex or a transient (perhaps catalytic) interaction will require careful measurement of lin-4 and lin-4 RNA levels and direct tests for a complex in vivo.

Temporal Regulation of lin-14 by lin-4

If lin-4 RNA inhibits lin-14 translation by an antisense interaction with the lin-14 3′UTR, then the temporal decrease of LIN-14 protein during C. elegans development is probably generated by a stage-specific increase in lin-4 gene expression.

According to this view, lin-4 is inactive during late embryogenesis and the early part of the L1, permitting synthesis of a high level of LIN-14 protein, which programs the execution of L1 fates. Later in the L1 stage, lin-4 gene expression is activated, and the lin-4L transcripts, lin-4S transcripts, or both bind to the complementary sequence elements in the 3′UTR of lin-14 mRNA. This RNA–RNA interaction results, through some unknown mechanism, in a decrease in LIN-14 protein between the L1 and L2 stages. The resulting reduced level of LIN-14 protein specifies the expression of L2-specific fates in the L2, and L3-specific fates in the L3 (Ambros and Horvitz, 1987). To downregulate lin-14 globally, lin-4 would have to be activated stage-specifically in at least those cells that express lin-14. A developmental activation of lin-4 might involve transcriptional induction or posttranscriptional modification of lin-4 RNA, or both. Little is known about what developmental signal(s) activate lin-4 gene expression, but it is likely that lin-4 is not activated until after a food signal initiates postembryonic development, since LIN-14 protein does not decrease until after feeding (Arasu et al., 1991). lin-4 may be required only in the L1 to modify lin-14 mRNA into an untranslatable state, or also during later stages to inhibit LIN-14 protein synthesis continuously. Features of this model need to be tested by examining in detail the temporal profile of lin-4 transcript levels and by artificially manipulating the level of active lin-4 RNA during development.

Antisense Regulation via 3′UTRs

There are a number of known examples of natural antisense regulatory mechanisms that affect the stability or translatability of target RNAs (Kimelman and Kirschner, 1989; Hilgetag and Neef, 1992; reviewed by Simons, 1988; Eguchi et al., 1991). lin-4 RNA probably does not control the stability of lin-14 mRNA because the steady state levels of lin-14 mRNA remain relatively constant during development and are not appreciably altered in a lin-4 mutant background (Wightman et al., 1993). lin-4 could inhibit LIN-14 protein synthesis indirectly, for example by modifying lin-14 mRNA, or by localizing it to a subcellular compartment (such as the nucleus) where it is inaccessible to the ribosomes. Alternatively, lin-4 RNA may bind to lin-14 mRNA in the cytoplasm and inhibit its translation by interacting directly with components of the translational machinery. Distinguishing among these broad classes of models will require determination of the subcellular location of lin-1 and lin-14 RNAs.

Previously known natural antisense mechanisms that affect translation involve interaction of an antisense RNA with the 5′UTR of the mRNA and appear to affect ribosome binding (Liao et al., 1987; Simons, 1988; Kittle et al., 1989). By contrast, if lin-4 does directly inhibit translation, then the proposed interaction between lin-4 RNA and the 3′ end of lin-14 might represent a novel kind of antisense translational control mechanism. Structural features of the lin-4–lin-14 complex might interfere with a critical interaction between the 3′ and 5′ ends of the lin-14 mRNA (Jackson and Standart, 1990; Gallie, 1991; Sachs and Dardorff, 1992). A regulation of poly(A) length might be involved in an interaction with translation factors, such as poly(A) binding protein, that act via the 3′ end (Sachs and Davis, 1989; Jackson and Standart, 1990). lin-4 RNA may act in conjunction with proteins, but to date there is no genetic evidence for genes other than lin-4 that regulate lin-14 via its 3′UTR. It is noteworthy that lin-4 RNA sequences that are not complementary to lin-14, and hence are looped out of the predicted lin-4–lin-14 hybrid, are completely conserved (Figure 8B). This loop may bind cellular proteins involved in the control of lin-14 mRNA. Proteins that interact with lin-4 RNA might also have functions es-
sential for cell viability, and so it may not be surprising that the genes encoding them have apparently not been previously identified by a visible heterochromatin phenotype.

The involvement of 3'UTR sequences in mRNA localization, stability, or translation and the developmental roles of posttranscriptional regulatory mechanisms mediated by 3'UTR sequences are becoming increasingly clear. For example, the mRNAs of mouse protamines (Kwon and Hecht, 1991), Drosophila bicoi(d, hunchback) (Wharton and Struhl, 1991), and nanos (Gavis and Lehmann, 1992), and nematode fem-3 (Ahringer and Kimble, 1991; Ahringer et al., 1992) and tra-2 (Goodwin et al., 1993) contain developmentally significant posttranscriptional regulatory sequences in their 3'UTRs. As yet, no antisense regulatory RNAs have been shown to be involved in the above cases, but the possibility has not been ruled out for any of them. Thus, lin-14 may not be the only developmental gene that is posttranscriptionally regulated via its 3' end by a small antisense RNA. Because the regulatory affect of a small antisense RNA such as lin-4 is directed toward the translation of a specific antisense partner mRNA, translational control of gene expression by small antisense RNAs may be a particularly effective strategy for providing a developmental switch of very high specificity.

Small Regulatory RNAs

lin-4 may represent a class of developmental regulatory genes that encode small antisense RNA products, but such genes might be difficult to identify by standard genetic approaches. If lin-4S is active in lin-14 shut-off, it may be among the smallest functional RNAs so far identified. The mouse and human Xist genes are believed to encode active RNAs important in the control of X chromosome inactivation (Brockdorff et al., 1992), and the mouse H19 gene also appears to encode an untranslatable RNA (Branman et al., 1990). However, the products of both these genes are much larger than lin-4 RNA; the Xist RNAs are 15–17 kb and the H19 RNAs are 2.6 kb. Small RNA molecules play a critical role in essential cellular processes such as RNA processing (Green, 1991; Guthrie, 1991) or RNA editing (Simpson, 1990). However, even compared with such genes without altering the host gene and the lack, in many cases, of distinguishing features in their genomic sequence might contribute to the difficulty of identifying other small RNA genes like lin-4.

Experimental Procedures

Nematode Strains

C. elegans strains used in this work were the following: wild-type C. elegans var Bristol (strain N2) (Brenner, 1973); C. elegans var Bergerac; lin-4(e912) (strain VT371); lin-4(e912)var-9(e1744)mm1C1 (VT499); bli-3(e785)lin-4(e912)dpy-10(e128); lin-14(m773) (VT373); lin-4(e912)/mm1C1 (VT460); lin-4(e912)dpy-10(s1280)mm1C1; lin-14(m792s) (VT441); lin-4(e912); maEx114(pVT363, PRF4) (VT509); and lin-4(e912); maEx115(pVT366, PRF4) (VT510). Other C. elegans species used in this work are the following: C. briggsae, obtained from the Caenorhabditis Genetics Center, C. vulgarius, and C. remanei, obtained from Scott Baird.

Mutagenesis

Noncomplementation screens were performed by treating N2 males with ethylmethane sulfonate (Brenner, 1973), mating them with hermaphrodites of the strain VT441 lin-4(e912)dpy-10(e128)mm1C1; lin-14(m702s), and corecombing the non Dpy F1 for hermaphrodites that displayed a Lin phenotype (Chalfie et al., 1981).

RFLP Mapping

Physical genetic mapping of Bristol/Bergerac RFLPs in the lin-4 region was performed as described elsewhere (Ruvkun et al., 1989). A strain congenic to Bristol for all chromosomes except the lin-4 region of L641 was obtained by performing nine successive backcrosses of lin-4(e912) to Bergerac lin-4(+), isolating the lin-4(+) allele at each backcross. This strain and Bergerac were crossed with bli-3(e785)lin-4(e912)dpy-10(e128)s/+ ++ males to obtain beregerac lin-4(+)bli-3(e785)lin-4(e912)dpy-10(e128) animals. Progeny from these hermaphrodites were screened for Lin-non-Dpy, Dpy-non-Lin, and Bli-non-Lin recombinants. DNA from these recombinants was analyzed by Southern blotting (see below) using as probes clones situated to the left of dpy-10 on the C. elegans physical map (Coulson et al., 1986, 1988).

Southern Blots

Preparation of genomic DNA, agarose gel electrophoresis, and Southern blotting were performed using standard techniques (Ausubel et al., 1989). Probes were prepared from cosmid or YAC DNAs by random primed synthesis (Feinberg and Vogelstein, 1983). YAC DNA was fractionated by field inversion gel electrophoresis (MI Research) after spheroplast lysis in the gel (Carle and Olson, 1985). YAC DNA was recovered from gel slices by electroelution, followed by phenol extraction and ethanol precipitation.

Isolation of Genomic Clones

C. elegans genomic DNA was cut with EcoRI, ligated into LambdaZAP vector (Stratagene), and plated on CES200. Replicas of the library were probed, in the presence of total yeast carrier DNA, with YAC DNA labeled by random primed synthesis. Plaques that hybridized to Y1B1 clone probe to Y424A4 probe were identified, the inserts were excised in vivo as Bluescript plasmid clones, and the plasmids were reanalyzed by restriction enzyme digestion and Southern hybridization with the above YAC probes. Plasmids that retained their differential hybridization pattern were pVT2D, pVT3C, and pVT63 (see Results). A probe from pVT2D was used to identify clones containing lin-4 sequences in LambdaZAP libraries of genomic DNA from C. briggsae and C. remanei. The C. briggsae lin-4 sequence was determined for a 1.4 kb EcoRI clone pVT612, and the C. remanei lin-4 sequence was determined for a 2.0 kb EcoRI clone pVTsA1. Short stretches of conserved sequence outside the rescued region were used as a guide to design primers (rc17 and rc18) to amplify lin-4 sequences from C. vulgarius worms. The 0.9 kb PCR product from C. vulgaris was inserted into Bluescript vector, creating clone pVT644C.

DNA Sequencing

Sequencing was performed by the method of Sanger et al., 1977) using the Sequenase 2.0 kit (United States Biochemicals). Sequencing was analyzed with a Macintosh, using DNA Strider 1.2 (March, 1990) or MULFOLD (Jaeger et al., 1989a, 1989b; Zuker, 1989) software, and with a VAX, using the GCG Sequence Analysis Software Package (version 7.1) by Genetics Computer Group, Incorporated
Northern blots were stripped by washing at 90°C in 0.1 x SSC, 0.5% SDS and then reprobed with an anti-U6 oligonucleotide probe.

S1 Nuclease Protection
S1 analysis using oligonucleotide probes was as described (Ausubel et al., 1989). Total RNA (30 µg) was analyzed in each sample. The initial hybridization of probes (1000 cpm/sample) to RNA was under standard conditions except for the rMGH6 probe, in which case hybridization was in an aqueous solution (final concentration, 1 M NaCl, 0.17 M HEPES [pH 7.5], and 0.3 M EDTA [pH 6.8]) at 50°C overnight. S1 digestion reactions contained 300 units of S1 nuclease and were at 37°C unless otherwise noted. S1 digestion products were analyzed on 10-15% urea-acrylamide gel, and the gels were exposed wet to XAR-5 film. Probes were generated as described below. rMGH1 was 5' end-labeled using [γ-32P]ATP and T4 polynucleotide kinase under standard reaction conditions. rMGH5 was labeled at its 3' end as follows: 300 ng of rMGH5 and 350 ng of rMGH6 were resuspended in 4 µl of TEN50 (10 mM Tris [pH 8.0], 1 mM EDTA, 0.5 M NaCl), heated to 75°C for 10 minutes, and then annealed at room temperature. The annealed mixture was diluted 1:3 in TEN50. Of this annealed oligonucleotide mixture, 1 µl was incubated with 10 µl of [γ-32P]dATP (6000 Ci/mmol), 10 µl of [γ-32P]dUTP (6000 Ci/mmol), 5 U of DNA polymerase I Klenow fragment in 50 mM Tris (pH 7.5), 10 mM MgCl2, 1 mM DTT, and 50 µg/ml BSA at 37°C for 15 minutes. The reaction was stopped by heating to 75°C for 10 minutes, and the oligonucleotides were ethanol precipitated in the presence of 10 µg of carrier RNA, resuspended in 1X TBE plus 90% formamide, heated to 90°C, and separated by electrophoresis through a 10% urea-acrylamide gel. The band corresponding to 3' end-labeled rMGH5 was cut from the gel, and the oligonucleotide was eluted by soaking the gel slice in 0.3 M NaOAc overnight at room temperature. The solution was extracted with phenol, CHCl3, IAA, and 3' end-labeled rMGH5 was precipitated by addition of 3 vol of ethanol. The precipitated rMGH5 was dissolved in 0.3 M NaOAc (final concentration, 0.3 M) and 3 vol of ethanol at -20°C.

Preparation of Nematode RNA
Animals were harvested from nematode growth medium agarose plates, washed several times in M9 buffer, collected by centrifugation, and frozen at -70°C. An equal volume of 2X LETS (200 mM LiCl, 20 mM EDTA, 20 mM Tris [pH 7.8], 2% SDS) and 2 vol of phenol, CHCl3, isooamyl alcohol (IAA) were added to the frozen worm pellet, which was then quickly thawed. Animals were lysed by vortexing with glass beads for several minutes. The aqueous phase was separated by centrifugation and reextracted three times with phenol, CHCl3, IAA. RNA was precipitated from the aqueous mixture by adding LiCl (final concentration 0.2 M) and 3 vol of ethanol and incubating at -20°C overnight. RNA was collected by centrifugation, and contaminating DNA was digested with RNase (Promega, 5-10 U/µg of RNA) for 30 min in 40 mM Tris (pH 7.9), 10 mM NaCl, 6 mM MgCl2, and 0.1 mM CaCl2 at 37°C. RNase was removed by extraction with phenol, CHCl3, IAA, and RNA was precipitated by the addition of NaOAc (final concentration, 0.3 M) and 3 vol of ethanol at -20°C.
Developmental Regulation via Natural Antisense RNA

Oligonucleotides

rf2 (amplification of ma167), 5' -TAGTGGAACTCTATrGT-3' .
rfb (sequencing), 5' -AGTGAAGAGACTGGGTCAT-3' .
rf3 (sequencing ma167), 5' CTAGACAATITCTAGAG-3' .
rfa (sequencing), %GACGCGTCGCCAGCGGTGGTG-3' .
rfe (sequencing), 5' -CCTG1TCCCTGAGACCTCAA-3' .
rcl2 (amplification and sequencing ma767), 5' -ACCGClTClTCTCC-.
rfd (sequencing), 5' -ACCGTCCTGGTACCCGGAGA-3' .
(TC)TCTG3' .
rcl7 (sequencing ma167). 5' -GACGAAGCGACCGAATG-3' .
CTCACAClTGAGGTCTCAGGGAACAGGCCGGAAGGCCGGAAGCATAAACTC -3' .
rfMGH5 (3' end mapping /in-4L), 5' -CTCACGAAlTCGACTACCGTCCT-.
rcl9 (mutagenesis), 5' sCGTAGACTACCTCGC-3' .
rcl18 (amplification), 5' -GGAAlTC(GT) (AG)TT(GA)TA(GA)TC(AG)AC-
TTCCAATTTTAGTATATGTTCTCCCTATAGTGAGTCGTATTA-3' .
AAGCGTCAATAG-3' .
va22 (mutagenesis), 5' GTCCTGGTACCCGGAGAGCCCAGGTGTG-
rfMGH6 (3' end labeling of rfmgh5), 5' -GTGTGAGTGTACTATTGATG-
rfMGH3 (primer extension analysis), S' CCCAGGTGTGAAGCATCAA-
rfMGH8 (3' end mapping /in-4S), 5' -GACTGCTAAGATAGTACACTC-
rfMGH7 (3' end labeling of rfmgha), 5' -TTCCCTGAGACCTCAAG-3' .
CTTCACACCT-3' .

Acknowledgments

We are grateful to Xianjie Yang for isolating the recombinants used in mapping in-4, to Amy Mills for in vitro mutagenesis, to Scott Baird and David Fitch for providing nematode strains, to Bob Barstead for his cDNA library, to Ann Rougvie for her C. elegans EcoRl genomic sequence analysis, and to Bruce Wightman and Gary Ruvkun for the C. briggsae genomic library and for communicating their results prior to publication. The product of the mouse Xist gene is a 15 kb inactive X-specific transcript containing no conserved ORF and located in the nucleus. Cell 71, 515–526.

References


GenBank Accession Numbers

The accession numbers for the Caenorhabditis sequences reported in this paper are U01829, U01830, U01837, and U01838.