doi:10.1038/nature09616

Supplementary Discussion

C. elegans miRNAs

To get a more comprehensive picture of the miRNAs in *C. elegans*, we analyzed the sequences of ~23 million small RNAs with perfect matches to the *C. elegans* genome, which had been generated using Illumina sequencing[15,37]. These included small RNAs from the same or comparable samples as those used for our 3P-Seq analyses (embryo, L1, L2, L3, L4, adult, dauer L3, and germline-deficient *glp-4(bn2)* mutant adults).

Previously annotated miRNAs. Previous miRNA annotations were extracted from version 14.0 of miRBase[38]. All of the 115 miRNAs sequenced earlier by using pyrosequencing[39], including the three mirtrons annotated later (*mir-1018~1020*)[17], were detected, and miRNA* sequences were detected for 104 of these. Two of those genes from which no miRNA* species were observed, *mir-78* and *mir-798*, have since been re-classified as 21U-RNAs, based on the presence of the 21U-RNA–associated upstream motif[39] and the correlation of their expression patterns with other RNAs of that class[15]. The deeper coverage also provided direct evidence for the expression of miRNAs from the *mir-356*, *mir-360*, and *lsy-6* loci, which were three early-annotated miRNA genes that had not been supported by the limited coverage of our pyrosequencing experiment[39]. The mature miRNA deriving from the *lsy-6* locus was observed with 521 reads and matched the existing annotation[40], which begins 2 nt downstream of the originally proposed miRNA[41]. However, miRNAs deriving from the *mir-360* and *mir-356* loci did not match prior annotations[40], with miR-360 processed from the opposite arm of the hairpin and miR-356 transcribed from the opposite genomic strand (Supplementary Table 6). Of the other 19 early annotations that were not supported by our pyrosequencing experiment (*mir-256/257/258-1,-2/260~273/353/354*), none were supported by the deeper coverage (Supplementary Table 6). When considering the hundreds of reads from *lsy-6*, a locus with very restricted expression, the absence miRNA reads from these loci speaks compellingly against their authenticity as miRNA genes.

Two miRNA annotations (*mir-1021/1022*) derive from traditional sequencing of small RNAs that co-purify with miRISC[42]. While *mir-1022* was validated by our analysis, no miRNA was observed deriving from the *mir-1021* locus (Supplementary Table 6). Additional miRNA annotations derive from the deep sequencing of small RNAs that co-immunoprecitate with miRISC-associated P-body components[43]. Of these eight miRNA genes, six were clearly supported by our analysis (*mir-1819~1824*). For the other two (*mir-1817/1818*) the data were consistent with the prior annotations but did not satisfy our requirements for inclusion in the confidently annotated set (Supplementary Table 6). Yet more miRNA genes have been predicted and annotated in *C. elegans* based on automated re-analysis of scarce reads from two published pyrosequencing datasets[44]. Of those 13 miRNA genes, our analysis clearly supported nine of the annotations (*mir-1820~1822/1828/1829a-c/1830/1834*; Supplementary Table 6). The data were consistent with four prior annotations (*mir-1817/1832/1832b/1833*) but did not satisfy our criteria for inclusion in the confidently annotated set (Supplementary Table 6). Small RNAs tiling across the *mir-1831* locus were inconsistent with its annotation as a miRNA gene (Supplementary Table 6).

The most extensive sequencing of *C. elegans* small RNAs, with almost twice as many genome-matching reads as were analyzed here, was reported by Kato et al.[45]. That study, which included small RNAs sampled from across a developmental profile similar to the one described here, employed an automated analysis pipeline to identify candidate miRNA genes[44] and reported 66 such candidates, 18 of which appeared sufficiently confident for miRbase annotation and were reported uniquely from that study (*mir-2207/2208a-b/2209a-c/2210~2217/2218a-b/2219/2220*)[38,45]. Of those 18 miRNA genes, our analysis clearly supported four of the annotations (*mir-2208b/2210/2215/2220*; Supplementary Table 6). For 11 of the annotations (*mir-2208a/2209a-c/2211-2213/2216/2217/2218a-b*) the data did not pass our

thresholds for confident annotation but did not specificly contradiction the existing annotations. For three annotations (*mir-2207/2214/2219*), the data were inconsistent with their status as miRNA genes. The *mir-2207* locus generated only a single read offset from the miRNA annotation (Supplementary Table 6). The *mir-2214* locus gave rise to hundreds of reads distributed across the locus in pattern more consistent with non-specific RNA degradation than with specific RNase III cleavage (Supplementary Table 6). Six reads from the *mir-2219* locus agreed in part with the prior miRNA annotation, but the two reads deriving from the annotated star strand were staggered, and both were inconsistent with the expected position of a miRNA*. Furthermore, this locus overlapped an annotated mRNA and spanned a splice junction, with the more-abundant reads from the annotated miRNA arm overlapping the exon and the less-abundant reads from the annotated miRNA* arm overlapping the intron (Supplementary Table 6). Taken together, these data indicate that the small RNAs deriving from the *mir-2219* locus were the degradation products of an mRNA.

Mirtrons from noncoding transcripts. Five newly identified miRNAs (see text below) derived from mirtrons (*mir-4807~4810/4816*; Supplementary Table 6). The four previously annotated *C. elegans* mirtrons (*mir-62/1018~1020*) as well as two newly identified mirtrons (*mir-4809/4816*) derived from annotated, protein coding host mRNAs. However, the other three newly identified mirtrons (*mir-4807/4808/4810*) as well as two pre-miRNAs that we reclassified as mirtrons (*mir-255/2220*) derived from spliced host transcripts that did not appear to be protein coding (Supplementary Fig. 10; Supplementary Table 6). Most notable of these was *mir-255*, for which the predicted pre-miRNA hairpin structure, the miRNA sequence and the presumed splice sites were all broadly conserved, but the base of the previously proposed pri-miRNA hairpin, which would normally be required for Drosha processing, was not conserved (Supplementary Fig. 10). A search of RNA-seq data[10,46] found reads spanning mirtron splice junctions for each of the 11 *C. elegans* mirtrons, except for *mir-2220*.

Untemplated nucleotides and length heterogeneity. Untemplated nucleotides are added to the 3′ ends of mature miRNAs with low efficiency in *C. elegans*[39]. Untemplated nucleotides were observed appended to previously and newly annotated (see below) mature miRNA and miRNA* species with an overall efficiency of 1.73%. MicroRNAs extended by one untemplated nucleotide were extended by a second with higher efficiency (5.53%), and so on for addition of a third (7.20%) and a fourth (11.23%) untemplated nucleotide. The most commonly appended untemplated nucleotide was U (75% of all untemplated nucleotides). The frequency of untemplated nucleotide addition was not even across all miRNAs: for 57 miRNA or miRNA* species deriving from both the 5′ and 3′ arms of their hairpin precursors, ≥5% of the mature RNAs sequenced included untemplated 3′ nucleotides (Supplementary Table 6). The overall 1.73% efficiency of untemplated nucleotide addition to miRNAs was in large excess over the 0.16% efficiency observed for annotated 21U-RNAs, which was considered as the upper limit on the background for untemplated nucleotide detection detected using our sequencing method.

MicroRNAs generally exhibit more 3′ heterogeneity than 5′ heterogeneity[29,39,47,48]. However, for a handful of miRNAs, shortened isoforms were abundantly observed whose 5′ ends were truncated by 6–10 nt (miR-43/54/75/85/358/1829c; Supplementary Table 6). These products gave rise to even more reads than the full-length mature miRNAs in these datasets despite being observed only scarcely or not at all in previous datasets generated using the 454 sequencing platform[39].

Newly identified miRNAs. Additional miRNAs were sought among the remaining small RNA reads. Genomic hits matching annotated ncRNA genes (tRNA, rRNA, etc) were excluded, as were hits whose upstream sequences had high-scoring matches to the 21U-RNA-associated motif[15,39]. MicroRNA 5′ ends are more consistent than 3′ ends because of the important role that 5′ end placement plays in defining the seed, and thereby the targets, of the mature miRNA[29,39,47,48,49]. MicroRNA candidates were therefore identified as genomic loci corresponding to the 5′ nucleotide of at least 10 perfectly-matching

reads that exhibited little local 5′ heterogeneity and were found in the context of a candidate miRNA precursor hairpin. Candidates meeting these criteria were manually inspected, and 14 loci were found that were not present in miRBase 14.0[38]. These produced few reads, with nine of the 14 producing only a tenth the number as the *lsy-6* locus. Only two appeared to be conserved in other nematodes. Since the time of our analysis, *mir-2221* and *mir-2953* have been annotated by others and listed in miRBase 16.0, leaving 12 as novel miRNA genes (*mir-4805~4816*; Supplementary Table 6).

The propensity for RNA complementing a miRNA hairpin to also form a hairpin structure creates a pathway for the emergence of novel miRNA genes on the opposite genomic strand of existing miRNA genes[50,51,52]. Small numbers of reads were observed deriving from the antisense strands of 43 validated miRNA genes, but these antisense reads were generally distributed across the predicted reverse-strand hairpin structure in a manner inconsistent with Dicer/Drosha processing (Supplementary Table 6). In the rare cases where reads derived from both arms of the predicted antisense hairpin (*mir-38/58/67/232*), the inconsistencies of 5′ and 3′termini, deviations from the expected 2 nt 3′ overhang, and frequency of G as the 5′ terminal nucleotide all conflicted with a the hypothesis of a miRNA biogenesis for these small RNAs and were more consistent with the properties of *C. elegans* endogenous siRNAs[39,53]. The locus with antisense reads most consistent with Drosha/Dicer processing was *mir-67*, and it generated only nine antisense reads, whereas the sense hairpin generated >67,000 reads (Supplementary Table 6). Taken together, these observations suggest that few if any *C. elegans* miRNA loci give rise to antisense miRNAs with regulatory roles.

MicroRNA expression patterns. The number of mature miRNA reads deriving from each gene varied widely. Some of the previously annotated miRNAs were sequenced more than a million times, whereas the newly identified miRNAs were scarce, often sequenced less frequently than the *lsy-6* miRNA, which is transcribed in only 1–9 neurons[41] and was represented by 521 reads (Supplementary Fig. 9; Supplementary Table 6). The most abundant isoform of mature miR-2953 was observed with only 11 reads, yet its seed and predicted pri-miRNA hairpin structure were conserved throughout the *Caenorhabditis* genus (Supplementary Fig. 9b,c), indicating that miRNAs scarce in deep-sequencing datasets can nonetheless impact evolutionary fitness and that additional biologically functional miRNAs might remain undetected.

Differences in miRNA read frequencies observed between the various developmental stages from which small RNAs were sequenced were used to construct a developmental expression profile for each miRNA gene (Supplementary Fig. 11a, Supplementary Table 6). The profiles determined based on read frequencies closely matched those that have been previously determined by northern blot (Supplementary Fig. 11b; compare to northern blots from Lau et al.[29]). As previously described in *C. elegans* as well as other systems[29,50,54,55], the expression patterns of genomically adjacent miRNAs were highly correlated, suggesting derivation from a common primary transcript (Supplementary Fig. 11c). In human, the correlation of expression patterns diminishes as the intervening distance surpasses 50 kb[55], whereas in *Drosophila* it diminishes as the intervening distance surpasses 10 kb[17]. We found that in *C. elegans*, the correlation diminished as the distance surpassed ~1 kb. The closely correlated expression patterns of neighboring miRNA genes suggested a substantial role for transcriptional regulation in defining miRNA expression profiles. However, many more counterexamples to this trend were observed in *C. elegans* than were observed previously in *Drosophila* (Supplementary Fig. 11c).

MicroRNA targeting in nematodes

The types of seed-matched sites preferentially conserved in nematodes. To begin to investigate miRNA targeting, we used an algorithm used previously to detect miRNA site conservation in vertebrate genome alignments[18]. Briefly, the method quantified the extent to which any *k*-mer was conserved using a branch-length score over phylogenies controlled for local conservation rates. Because a

sequence can be conserved for many reasons other than microRNA targeting, conservation scores were interpreted by comparing them to background conservation estimated from cohorts of control *k*-mers, selected for similar expected conservation based on their dinucleotide content. Thus, after controlling for local conservation rates, sequence composition, site type, and phylogenetic structure, any difference between the conservation of a miRNA site and that of its background could be attributed to selective maintenance of miRNA targeting by natural selection. Although there were fewer nematode genomes available than vertebrate genomes, the method is not strongly sensitive to the number of genomes, and the evolutionary time covered by the phylogeny was comparable to that for the vertebrates (Supplementary Fig. 12a).

We applied this method to sequences complementary to the 60 *C. elegans* miRNA families that were conserved among sequenced nematodes (Supplementary Table 7). As expected, systematic examination of the conservation above background for hexamer matches starting at each miRNA position revealed statistically significant and specific conservation of sequences matching the miRNA seeds, i.e., miRNA nucleotides 2–7 (Fig. 4a, Supplementary Fig. 12). Interestingly, matches to miRNA positions 1–6 were also significantly conserved above background, even when excluding sites with matches to position seven. This result, which differed from that in vertebrates[18], indicated an important difference between target recognition in nematodes compared to that in other clades. Further analyses showed that for sites that matched miRNA nucleotides 2–6, those with an A at target position 1 were preferentially conserved, even if the miRNA did not begin with a U (Supplementary Fig. 13a,b). Therefore, we called this the "6mer-A1" site (Fig 4a). The preference for an A at target position 1 resembled 7 or 8 nt seed-matched sites in vertebrates, which are more often conserved and more effective if they have an adenosine at position 1 than if they have a Watson-Crick match[49,56,57]. We also observed statistically significant conservation for hexamer matches to nucleotides 3-8 and 4-9, although the signal-to-background ratios for these shifted hexamer seed matches were marginal (data not shown).

Analysis of 7 nt matches revealed preferential conservation of the same two seed-matched types as conserved in vertebrates[49]; these are the 7mer-m8 and the 7mer-A1 sites (Supplementary Fig. 12b). When examining 8 nt sites, however, another difference was observed between nematodes and vertebrates. In nematodes, matches to positions 2-8 followed by a U were nearly as conserved as those followed by an A, which differed from the preference for only an A observed in vertebrates (Supplementary Fig. 13c). Thus nematodes have two preferentially conserved 8 nt sites, the 8mer-A1 and the 8mer-U1 (Fig. 4a).

Validation of site efficacy using experimental datasets. We tested the efficacy of these six types, as well as that of the two offset 6mers with more marginal conservation, using two sources of data: an ALG-1 cross-linking immunoprecipitation (CLIP) experiment representing the genome-wide targeting preferences of *C. elegans* miRNAs[19], and microarray data following a miR-124 knockout experiment[20]. We found significant enrichment of previously established site types as well as of 8mer-U1 sites in ALG-1 CLIP tag clusters (Supplementary Table 8). The two shifted 6mer types (3-8 and 4-9), which were conserved only marginally, were not significantly enriched. The 6mer-A1 sites also failed to achieve statistical significance, but were enriched more than matches to nucleotides 2-5 flanked by other nucleotides opposite miRNA position 1. In miR-124 knockout cells, mRNAs containing any seed match type including the 8mer-U1 and 6mer-A1 were significantly de-repressed ($P < 0.03$, Supplementary Fig. 12d). Therefore, we performed further analyses using the set of six seed match types with strong evidence for preferential conservation as well as experimental evidence for in vivo targeting (Fig. 4a). One of the shifted 6mer sites that had a marginal conservation signal was also included for comparison. The signal-to-background ratios, or fold-enrichment of conservation, of the site types displayed the expected hierarchy: 8mers were conserved more than 7mers, which were conserved more than 6mers

(Supplementary Fig. 12b). The six major types each had >600 sites confidently conserved above background (Supplementary Fig. 12c).

Sites with seed mismatches. We next searched for preferential conservation of imperfect seed matches. As in vertebrates[18], there was detectable conservation of seed matches with G:U wobbles or bulges in the target, but fewer than 400 of these imperfect seed match sites were conserved above background levels (Supplementary Fig. 14a). We found no significant conservation above background for sites either with other mismatches or with bulges on the miRNA side of the duplex (data not shown).

Although pairing to the 3′ end of the miRNA can supplement seed matches or compensate for imperfect seed matches, only a small fraction of sites conserved in flies and vertebrates have preferentially conserved 3′-supplementary or 3′-compensatory pairing[18,49,56,58]. We found that the same was true for nematode sites. An established metric for detecting conserved 3′ pairing[18] indicated that only 365 ± 220 3′-supplementary sites were preferentially conserved (Supplementary Fig. 14b), and the number of 3′-compensatory sites preferentially conserved was estimated to be <50 (Supplementary Fig. 14c). The most compelling evidence for conservation of 3′-compensatory pairing was at sites with extremely favorable 3′-pairing scores. We found seven conserved instances with 3′-pairing score ≥6, compared to zero for chimeric control cohorts (Supplementary Table 9). The top three of these predicted 3′-compensatory sites were the two *let-7* sites in *lin-41* and a *let-7* site in *hbl-1*, each of which has strong experimental support[22,23,59]. Thus, despite their influence on early notions of miRNA target recognition, 3′-compensatory sites represent only ~1% of all sites conserved in nematodes. Taken together, our analyses revealed that with respect to mismatches and wobbles to the seed nucleotides, nematodes have stringency comparable to that observed for vertebrates but nonetheless have more permissive miRNA targeting because of two additional seed-matched types.

Rationale for detectable site enrichment. What might explain the substantial enrichment for miRNA sites in *C. elegans* 3′UTRs but not human 3′UTRs? Both neutral sites (i.e., sites that function but impart inconsequential mRNA repression or miRNA titration) and non-functional sites not will be either enriched or depleted. In contrast, consequential sites that emerge over the course of UTR evolution will be preferentially retained if they are beneficial and preferentially lost (in a process sometimes called anti-targeting) if they are harmful[60,61,62]. Thus, overall site enrichment is a function of these two competing processes. The emergence by chance of both deleterious and beneficial miRNA sites should be directly proportional to 3′UTR length, but the selective pressure acting on sites should be similar regardless of UTR length. Because the loss of deleterious sites should scale directly with their rate of emergence, anti-targeting should also be proportional with UTR length. In contrast, the fraction of mRNAs whose output could benefit from miRNA-mediated targeting is presumably similar in humans and nematodes, which would lead to comparable optimal numbers of functional sites in the two species. The selective maintenance of beneficial sites would tend to increase the number of functional sites towards this optimum. The extent to which the optimum for *C. elegans* (with its shorter UTRs) approaches that of human (with its longer UTRs) would create a greater density of beneficial sites in *C. elegans* UTRs, as reflected in part by the greater density of conserved sites in nematode UTRs (Fig. 4b). Thus in humans, the extensive anti-targeting balances out the selective retention of miRNA sites, resulting in no net enrichment, whereas *C. elegans*, with its short UTRs, the loss of sites through antitargeting is not sufficient to balance the selective retention of beneficial sites, resulting in a net enrichment.

Supplementary References

37 Seo, T. S. *et al.* Photocleavable fluorescent nucleotides for DNA sequencing on a chip constructed by site-specific coupling chemistry. *Proc Natl Acad Sci U S A* **101**, 5488-5493 (2004).

38 Griffiths-Jones, S., Grocock, R. J., van Dongen, S., Bateman, A. & Enright, A. J. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34**, D140-144 (2006).

39 Ruby, J. G. *et al.* Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell* **127**, 1193-1207 (2006).

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

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

42 Gu, S. G. *et al.* Distinct ribonucleoprotein reservoirs for microRNA and siRNA populations in *C. elegans*. *RNA* **13**, 1492-1504 (2007).

43 Zhang, L. *et al.* Systematic identification of *C. elegans* miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2. *Mol Cell* **28**, 598-613 (2007).

44 Friedlander, M. R. *et al.* Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol* **26**, 407-415 (2008).

45 Kato, M., de Lencastre, A., Pincus, Z. & Slack, F. J. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during *Caenorhabditis elegans* development. *Genome Biol* **10**, R54 (2009).

46 Celniker, S. E. *et al.* Unlocking the secrets of the genome. *Nature* **459**, 927-930 (2009).

47 Basyuk, E., Suavet, F., Doglio, A., Bordonne, R. & Bertrand, E. Human let-7 stem-loop precursors harbor features of RNase III cleavage products. *Nucleic Acids Res* **31**, 6593-6597 (2003).

48 Lim, L. P. *et al.* The microRNAs of *Caenorhabditis elegans*. *Genes Dev* **17**, 991-1008 (2003).

49 Lewis, B. P., Burge, C. B. & Bartel, D. P. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *Cell* **120**, 15-20 (2005).

50 Ruby, J. G. *et al.* Evolution, biogenesis, expression, and target predictions of a substantially expanded set of *Drosophila* microRNAs. *Genome Res* **17**, 1850-1864 (2007).

51 Stark, A. *et al.* A single Hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands. *Genes Dev* **22**, 8-13 (2008).

52 Tyler, D. M. *et al.* Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev* **22**, 26-36 (2008).

53 Ambros, V., Lee, R. C., Lavanway, A., Williams, P. T. & Jewell, D. MicroRNAs and other tiny endogenous RNAs in *C. elegans*. *Curr Biol* **13**, 807-818 (2003).

54 Sempere, L. F. *et al.* Expression profiling of mammalian microRNAs uncovers a subset of brain-expressed microRNAs with possible roles in murine and human neuronal differentiation. *Genome Biol* **5**, R13 (2004).

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

56 Grimson, A. *et al.* MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Mol Cell* **27**, 91-105 (2007).

57 Nielsen, C. B. *et al.* Determinants of targeting by endogenous and exogenous microRNAs and siRNAs. *RNA* **13**, 1894-1910 (2007).

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

59 Vella, M. C., Choi, E. Y., Lin, S. Y., Reinert, K. & Slack, F. J. The *C. elegans* microRNA *let-7* binds to imperfect *let-7* complementary sites from the *lin-41* 3'UTR. *Genes Dev* **18**, 132-137 (2004).

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

61 Farh, K. K. *et al.* The widespread impact of mammalian MicroRNAs on mRNA repression and evolution. *Science* **310**, 1817-1821 (2005).

62 Stark, A., Brennecke, J., Bushati, N., Russell, R. B. & Cohen, S. M. Animal MicroRNAs confer robustness to gene expression and have a significant impact on 3'UTR evolution. *Cell* **123**, 1133-1146 (2005).

63 Marzluff, W. F., Wagner, E. J. & Duronio, R. J. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet* **9**, 843-854 (2008).

64 Stricklin, S. L., Griffiths-Jones, S. & Eddy, S. R. *C. elegans* noncoding RNA genes. *WormBook*, 1-7 (2005).

65 Lanzotti, D. J., Kaygun, H., Yang, X., Duronio, R. J. & Marzluff, W. F. Developmental control of histone mRNA and dSLBP synthesis during *Drosophila* embryogenesis and the role of dSLBP in histone mRNA 3' end processing in vivo. *Mol Cell Biol* **22**, 2267-2282 (2002).

66 Hajarnavis, A., Korf, I. & Durbin, R. A probabilistic model of 3' end formation in *Caenorhabditis elegans*. *Nucleic Acids Res* **32**, 3392-3399 (2004).

67 Hu, J., Lutz, C. S., Wilusz, J. & Tian, B. Bioinformatic identification of candidate cis-regulatory elements involved in human mRNA polyadenylation. *RNA* **11**, 1485-1493 (2005).

68 Brown, K. M. & Gilmartin, G. M. A mechanism for the regulation of pre-mRNA 3' processing by human cleavage factor Im. *Mol Cell* **12**, 1467-1476 (2003).

69 Kaufmann, I., Martin, G., Friedlein, A., Langen, H. & Keller, W. Human Fip1 is a subunit of CPSF that binds to U-rich RNA elements and stimulates poly(A) polymerase. *EMBO J* **23**, 616-626 (2004).

70 Chen, F., MacDonald, C. C. & Wilusz, J. Cleavage site determinants in the mammalian polyadenylation signal. *Nucleic Acids Res* **23**, 2614-2620 (1995).

71 Cali, B. & Anderson, P. Genetic evidence that *smg-6* is an essential gene. *Worm Breeder's Gazette* **12**, 26 (1993).

72 Lim, L. P. & Burge, C. B. A computational analysis of sequence features involved in recognition of short introns. *Proc Natl Acad Sci U S A* **98**, 11193-11198 (2001).

73 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863-14868 (1998).

Supplementary Table 1. Classification of reads mapping to the genome. Reads were aligned to the genome using Bowtie[32]. Those reads for which the best alignment was to one position in the genome were further categorized based on the nucleotide identities of their 3′ ends. Reads with two or more 3′-terminal adenylates were classified as 3P tags if at least one of the terminal A's was a mismatch to the genome. Reads with two or more 3′-terminal adenylates were classified as "templated terminal A" reads if all of the terminal adenylates matched the genome. Reads with less than two 3′-terminal adenylates were classified as "templated terminal N" reads. 3P clusters were built by iterative clustering of 3P tags, as described in the text, and the number of reads for each category within 10 nucleotides of the center of each cluster was tallied. Less than 100% of 3P tags overlapped a cluster because we required clusters to be supported by multiple independent tags.

Library	Mapped to > 1 locus	Templated terminal N	Templated terminal A	3P tags
Egg	1,943,019	2,679,102	727,787	4,425,380
L1	992,018	2,224,680	661,798	3,066,385
L2	1,480,095	3,120,221	623,763	2,479,287
L3	1,791,489	4,002,636	899,853	3,512,260
L4	2,104,059	4,000,728	972,357	3,691,607
Adult	2,458,264	4,660,708	313,591	1,372,600
glp-4 adult	2,080,217	4,531,197	889,187	3,561,702
Dauer	1,773,824	4,252,531	920,056	3,924,815
Mixed stage	2,203,199	4,946,767	1,643,635	5,834,657
Total	16,826,184	34,418,570	7,652,022	31,868,693
Number overlapping a 3P cluster		19,379,311	6,579,270	29,494,909
Fraction overlapping a 3P cluster		0.56	0.86	0.93

Supplementary Table 2. Direct comparison between UTRs identified by 3P-Seq and cleavage sites reported by Mangone et al., who used oligo(dT)-based methods[8]. To avoid differences attributable to microheterogeneity, UTRs were considered supported if a cleavage site identified by 3P-Seq was within 20 nucleotides of a representative site identified by oligo(dT)-based methods. To avoid differences attributable to isoforms from OERs, OER isoforms were assigned to their respective 3P-Seq UTRs, and any oligo(dT)-based cleavage site mapping between the ends of these OER isoforms or up to 20 upstream or downstream was considered as support for the UTR.

3P-Seq UTRs supported by an oligo(dT)-based site	15,455
3P-Seq UTRs not supported by an oligo(dT)-based site	8,581
UTRs found by 3P-Seq	24,036

Supplementary Table 3. Contributions of 3P-Seq data and the Mangone et al. (2010) dataset[8] for mRNA annotations generated and used by the modENCODE consortium[*]. Listed are the numbers of cleavage sites that define distinct 3′UTR isoforms, grouped based on the dataset(s) used to identify each site. The percentage of sites supported by an independent analysis of RNA-Seq tags with untemplated terminal A's is indicated in parentheses. The RNA-Seq approach also identified 46 proximal and 93 distal sites not identified by either of the other two approaches. Wormbase (version 170) annotations identified nine proximal and 13 distal sites not found by any of the three approaches, bringing their total number of sites to 30,737. Numbers were compiled from Table S2a of Gerstein *et al.*[*], which considers all cleavage sites within 20 nucleotides of each other as overlapping. The high percentage with RNA-Seq evidence in the first column compared to other columns can be attributed to higher expression of the mRNAs found by both approaches. However, the higher percentage with RNA-Seq evidence observed for the 3P-Seq–only UTRs compared to the Mangone-only UTRs cannot be explained by differences in expression and is consistent with the conclusion that thousands of the annotations uniquely identified from the Mangone et al. dataset are false positives (Supplementary Fig. 5c). The numbers of 3P-Seq–supported sites reported in this table and Figure 2a differ because the two analyses started with different sets of protein-coding regions and used different cutoffs to define 3P-Seq tags and sites.

	3P-Seq & Mangone		3P-Seq only		Mangone only	
Proximal	4,726	(41.2%)	3,789	(5.6%)	3,900	(1.1%)
Distal	11,364	(41.5%)	4,969	(6.7%)	1,828	(0.7%)
Total	16,090	(41.4%)	8,758	(6.2%)	5,728	(1.0%)

[*] Gerstein, M. B. *et al.* Integrative Analysis of Functional Elements in the *Caenorhabditis elegans* Genome by the modENCODE Project. *Second revision under consideration at Science.*

Supplementary Table 4. *C. elegans* Poly(A) signals.

Enriched over upstream region			Enriched over Markov Expectation		
Hexamer	Number	*P* value	Hexamer	Number	*P* value
AAUAAA	4,999	0	AAUAAA	4,999	0
AAUGAA	984	0	AAUGAA	984	0
UAUAAA	564	2.69E-166	CAUAAA	344	4.16E-128
CAUAAA	341	3.49E-97	UAUAAA	561	2.14E-147
GAUAAA	313	6.09E-96	GAUAAA	313	1.63E-154
UAUGAA	179	1.82E-48	UAUGAA	179	3.99E-80
AGUAAA	126	1.07E-34	AGUAAA	126	4.99E-61
CAUGAA	81	3.62E-22	CAUGAA	81	1.75E-40
AAAAAA	120	2.30E-22	AAUACA	67	9.07E-30
AUUAAA	96	2.00E-16	GAUGAA	64	2.49E-29
GAUGAA	63	4.17E-15	AUUAAA	100	1.12E-14
AAUACA	62	1.30E-13	AUAAUA	48	8.29E-13
AAUAAU	60	6.20E-07	AAAAAA	108	4.90E-10
ACUAAA	23	3.85E-4	AAUAUA	31	2.86E-07
AAUAUA	32	5.97E-4	ACUAAA	24	5.88E-06

Supplementary Table 5. Alternative operons. Sites are indicated using genomic coordinates (WS190/ce6).

Entrez gene ID	ALE cleavage site	SL2 splice site
181565	14,704,407	14,704,259
175480	3,725,053	3,724,956
172363	6,531,985	6,531,963
172347	6,442,512	6,442,426
182452	1,276,447	1,276,609
175919	6,457,252	6,457,357
173147	12,773,919	12,774,098
177432	6,642,151	6,642,318
181659	15,694,179	15,694,309
174654	10,726,313	10,726,414
172989	10,827,184	10,827,305
175367	2,477,878	2,478,018

Supplementary Table 6 is provided separately as an html file.

Supplementary Table 7. Conserved miRNA families used in the analysis of conserved targeting in nematodes. The miRNAs annotated as conserved in Supplementary Table 6 were grouped into families sharing nucleotides 2-8. All miRNAs in this set with *C. briggsae* annotations in miRbase release 12 were included. Of the remaining miRNAs, those with a perfectly conserved seed in two or more additional species and greater than 80% average identity over the mature sequence were also included. miR-2953 was also included as its seed and hairpin were well conserved (Supplementary Fig. 9). For miRNAs having expression of both arms (5p/3p designation), only the arm with the indicated seed was included.

Seed + nt 8	*C. elegans* miRNAs in family
AAAUGCA	miR-232
AAAUGCC	miR-357
AACUGAA	miR-255
AAGCUCG	miR-231;miR-787
AAGGCAC	miR-124
AAGUGAA	miR-86;miR-785
AAUACGU	miR-70
AAUACUG	miR-236
AAUCUCA	miR-259
AAUGCCC	miR-786
ACAAAGU	miR-85
ACAGAAG	miR-2953
ACCCGUA	miR-51;miR-52;miR-53;miR-54;miR-55;miR-56
ACCCUGU	miR-57
ACUGGCC	miR-240
AGCACCA	miR-49;miR-83
AUCACAG	miR-2;miR-43;miR-250;miR-797
AUCAUCG	miR-392
AUGACAC	miR-63;miR-64;miR-65;miR-66;miR-229
AUGGCAC	miR-228
AUUAUGC	miR-60
AUUGCAC	miR-235
CACAACC	miR-67
CACAGGA	miR-249
CACCGGG	miR-35;miR-36;miR-37;miR-38;miR-39;miR-40;miR-41;miR-42
CACUGGU	miR-359
CCCUGAG	lin-4;miR-237
CCCUGCC	miR-789
CCGCUUC	miR-788
CUUUGGU	miR-244
GAAAGAC	miR-71
GACCGUA	miR-360
GACUAGA	miR-44;miR-45;miR-61;miR-247
GAGAUCA	miR-80;miR-81;miR-82;miR-1834
GAGAUCG	miR-58
GAGGUAG	let-7;miR-48;miR-84;miR-241;miR-795
GAUAUGU	miR-50;miR-62;miR-90
GCAAAUC	miR-254
GGAAUGU	miR-1;miR-796
GGCAAGA	miR-72;miR-73;miR-74;miR-266
GGCACAA	miR-784
GGCAGUG	miR-34;miR-1824
GUCAUGG	miR-46;miR-47
UAAAGCU	miR-75;miR-79
UAAGUAG	miR-251;miR-252
UACACGU	miR-248
UACAUGU	miR-246
UAGUAGG	miR-253
UAUUAGU	miR-230
UAUUGCU	miR-234
UCAAUAU	miR-4813
UCAUCAG	miR-77
UCGUUGU	miR-76
UGAGCAA	miR-87;miR-233;miR-356
UGCGUAG	miR-242
UUGGCAC	miR-790;miR-791
UUGGUCC	miR-245
UUGUACU	miR-238;miR-239a;miR-239b
UUGUUUU	miR-355
UUUGUAU	lsy-6

Supplementary Table 8. Enrichment of seed-matched types in ALG-1 CLIP tags from Zisoulis et al. (2010)[19]. For each set of sites, 1,000 cohorts of control *k*-mers were chosen to match the number of G+C nucleotides and the number of CpG dinucleotides. The observed:expected ratio compares the number of seed match occurrences to the mean of the controls, and the *P* value reports the fraction of control cohorts with more extreme observed:expected ratios. Sites types with evidence of preferential conservation are highlighted in bold.

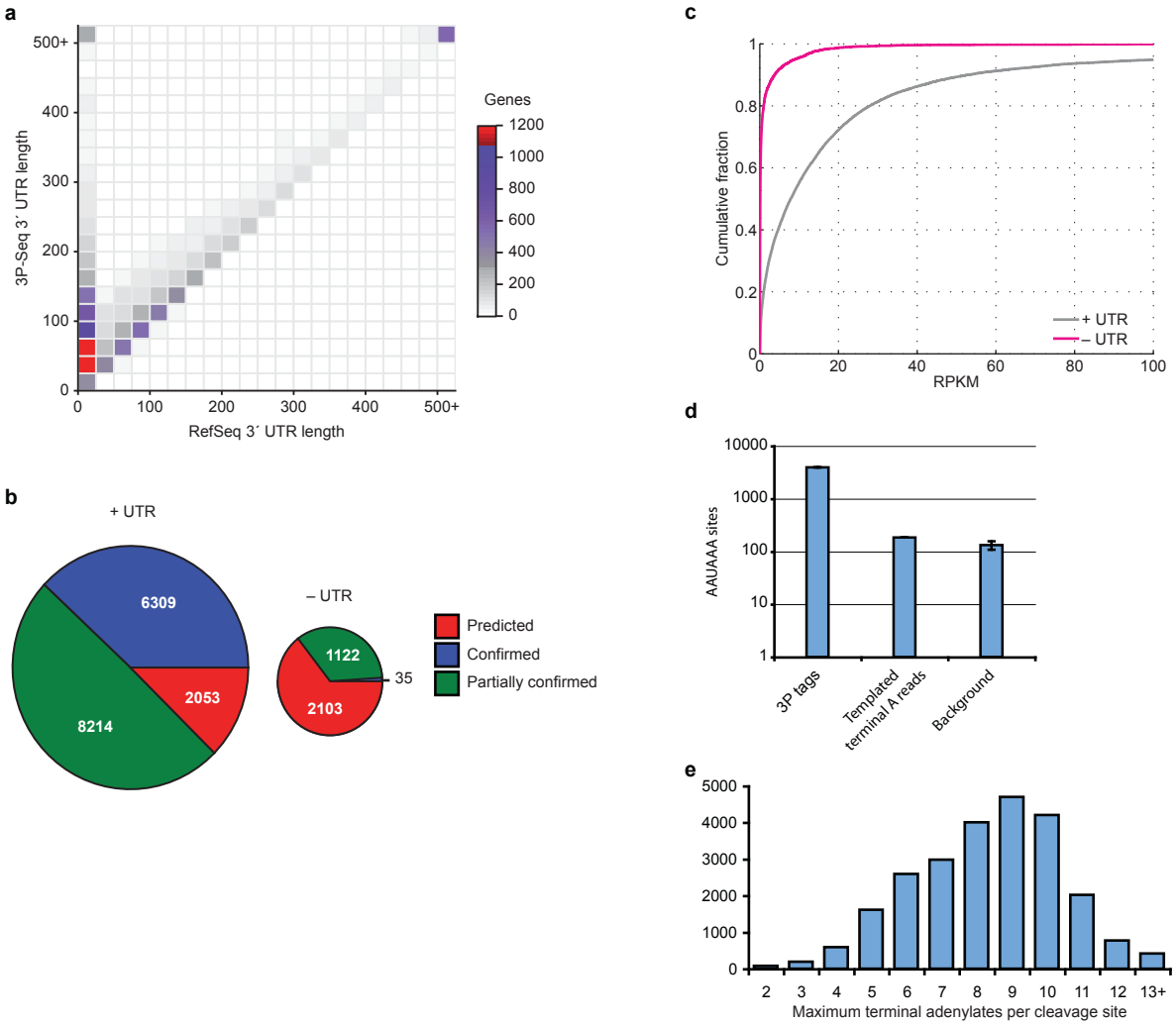
Seed match	Observed:expected	*P* value
8mer A1	**1.46**	**< 0.001**
8mer C1	1.11	0.16
8mer G1	1.13	0.12
8mer U1	**1.16**	**0.03**
7mer A1	**1.19**	**< 0.001**
7mer C1	1.05	0.30
7mer G1	1.07	0.21
7mer U1	1.05	0.22
6mer A1	**1.07**	**0.07**
6mer C1	0.99	0.57
6mer G1	0.99	0.61
6mer U1	1.03	0.29

Supplementary Table 9. Highly conserved predicted 3´-compensatory targeting interactions. Listed are targets with miRNA sites having any type of single-nucleotide seed mismatch or bulge that also have a 3´-pairing score ≥6.0 and a conservation branch length ≥0.5. Also listed are targets with bulged sites that have a 3´-pairing score ≥5.0 and a branch-length score >1.0.

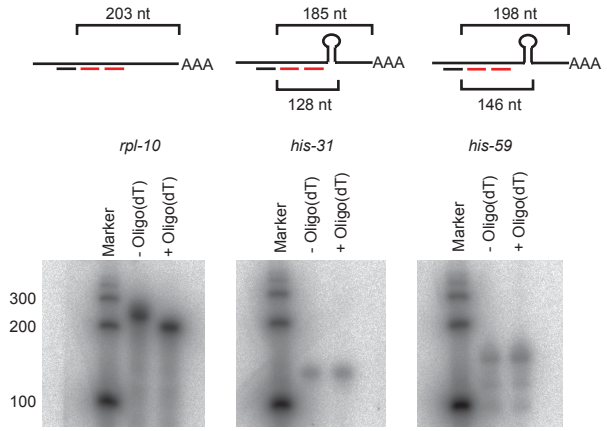
miRNA	Target mRNA	Mismatch type	Branch length	3´-pairing score
let-7	*lin-41*	Target bulge	0.9	6.5
let-7	*lin-41*	7mer GU wobble	0.9	6.5
let-7	*hbl-1*	miRNA bulge	0.9	6.5
miR-255	*unc-83*	Mismatch	1.25	6.0
miR-356	*stn-1*	miRNA bulge	1.1	6.0
miR-75	ZK1127.10	Mismatch	0.9	6.0
miR-266	*ceh-14*	GU wobble	0.5	6.0
miR-87	*pde-4*	miRNA bulge	1.25	5.5
miR-1834	*ram-2*	Target bulge	1.25	5.5
miR-45	*lin-14*	miRNA bulge	1.1	5.5
miR-90	H28G03.1	miRNA bulge	1.25	5.0
miR-239b	*icd-1*	miRNA bulge	1.25	5.0
miR-228	*acn-1*	miRNA bulge	1.25	5.0
miR-72	M02B1.2	miRNA bulge	1.05	5.0

Supplementary Table 10. Conserved *Drosophila* miRNA families. Families are based on miRBase 14.0 annotations and conserved to *Drosophila pseudoobscura*.

Seed + nt 8	*D. melanogaster* miRNAs in family
AAAGCUA	miR-79
AAAUAUC	miR-283
AAAUAUU	miR-289
AAAUGCA	miR-277
AAGGAAC	miR-5
AAGGCAC	miR-124
AAUACUG	miR-8
AAUCUCA	miR-304
ACCCGUA	miR-100
AGCACCA	miR-285;miR-995;miR-998
AGGAACU	miR-276a;miR-276b
AUCACAG	miR-2a;miR-2b;miR-6;miR-11;miR-13a;miR-13b;miR-308;miR-2c
AUCUAGC	miR-282
AUUCGAG	miR-314
AUUGCAC	miR-92a;miR-92b;miR-310;miR-311;miR-312;miR-313
CACAACC	miR-307
CACUGGG	miR-3;miR-309;miR-318
CAGGUAC	miR-275;miR-306
CAGUCUU	miR-14
CCCUGAG	miR-125
CGGUGGG	miR-278
CGUAUAC	miR-iab-4-5p
CUUUGGU	miR-9a;miR-9c;miR-9b
GAACACA	miR-317
GAAGUCA	miR-284
GACUAGA	miR-279;miR-286;miR-996
GAGAUCA	bantam
GAGGUAG	let-7
GAGUAUU	miR-12;miR-960
GAUUGUC	miR-219
GGAAGAC	miR-7
GGAAUGU	miR-1
GGACGGA	miR-184
GGCAAGA	miR-31b;miR-31a
GGCAGUG	miR-34
GGUAUAC	miR-iab-4-3p
GGUGCAU	miR-33
GUAUUUA	miR-280
GUCAUGG	miR-281
GUCUUUU	miR-316
GUGUUGA	miR-287
UAAAGCU	miR-4
UGAGCAA	miR-87
UGGUCCC	miR-133
UGUGCGU	miR-210
UUAAUGG	miR-263a
UUCAUGU	miR-288
UUGGCAC	miR-263b
UUGUACU	miR-305
UUUGAUU	miR-315
UUUGUGA	miR-274



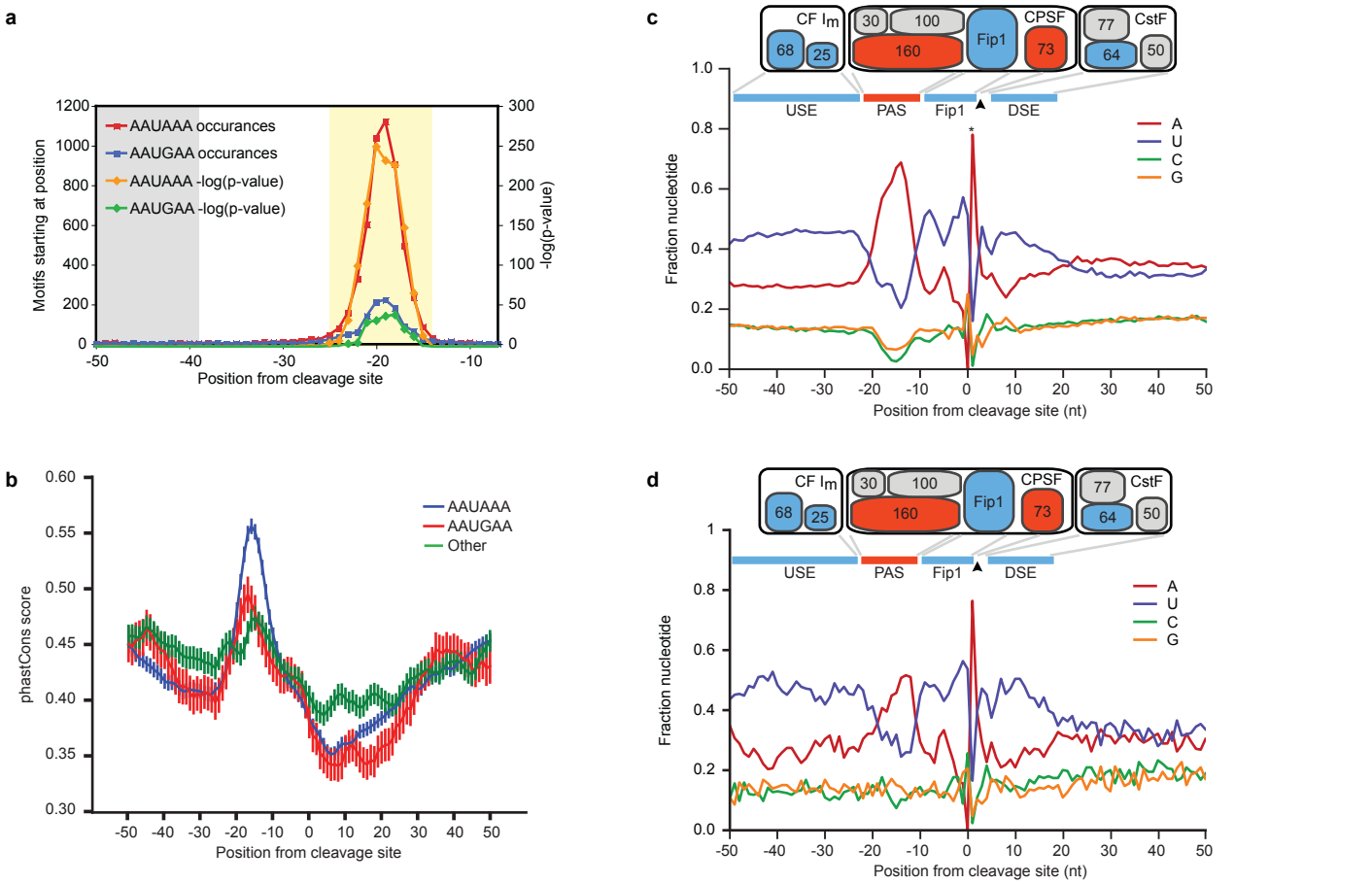
Supplementary Figure 1. Properties of genes with or without 3P-defined 3´UTRs. a, Bivariate histogram of UTR lengths derived from 3P-Seq compared to the previous RefSeq annotations. UTR lengths were binned in increments of 25 nt, considering only the longest UTR isoform for each gene with a RefSeq mRNA. **b,** Fraction of genes with and without 3P-Seq–defined 3´UTRs (+ UTR or – UTR, respectively), classified as "confirmed," "partially confirmed" or "predicted" in Wormbase 190. Confirmed genes have experimental support (typically ESTs) for the expression of all bases in the gene model. Partially confirmed genes have experimental support for only part of the gene model, and predicted genes have no experimental evidence of expression. The area of each pie chart corresponds to the number of genes. **c,** Cumulative distributions of reads per kilobase per million unique genome-matching reads (RPKM) for genes with or without a 3P-Seq–defined 3´UTR. RPKMs were calculated using published RNA-Seq data from L2, L3, L4 and young adult N2 worms[10]. **d,** Considering reads with only templated terminal adenylates would have added little sensitivity while greatly compromising specificity. Shown are the numbers of inferred cleavage sites with an AAUAAA motif beginning 25–14 nucleotides upstream. Clusters were built with reads possessing ≥2 templated terminal A's (Supplementary Table 1), as described for clusters built with 3P tags. The resulting 58,666 clusters mostly overlapped 3P clusters (Supplementary Table 1). Of the 10,199 clusters that did not overlap a 3P-Seq cluster, only 185 (the value plotted) had an upstream AAUAAA. For comparison, 10,199 3P-Seq–identified cleavage sites were randomly selected and the number preceded by AAUAAA (3959) is plotted (3P tags). To estimate the background expected by chance, 10,199 random positions from UTRs with single cleavage sites (excluding the last 25 bases) were selected and the number with an AAUAAA beginning 25–14 nucleotides upstream was determined. This procedure was repeated 1,000 times and the mean number (132) is plotted with error bars showing the 95% confidence interval (± 24). This interval agreed well with the 2.5th and 97.5th percentile of AAUAAA instances observed in the random samples of (108 and 155, respectively). To estimate the number of cleavage sites that were missing from clusters requiring untemplated A's, we subtracted the background from the observed and then accounted for the observation that AAUAAA is found upstream of 43% of 3P-Seq–identified cleavage sites, thereby yielding an estimated 124 ± 56 cleavage sites that were missed because we did not consider reads with only templated terminal adenylates. **e,** Maximum number of terminal adenylates for 3P tags that identified cleavage sites. For each cluster of 3P tags that identified a cleavage site (Fig. 2A), the maximum number of 3´ adenylates from tags within that cluster was tallied. This distribution shows that most sites were defined by at least one tag with many terminal adenylates. For example, 90% of clusters included at least one tag with ≥6 3´ adenylates.



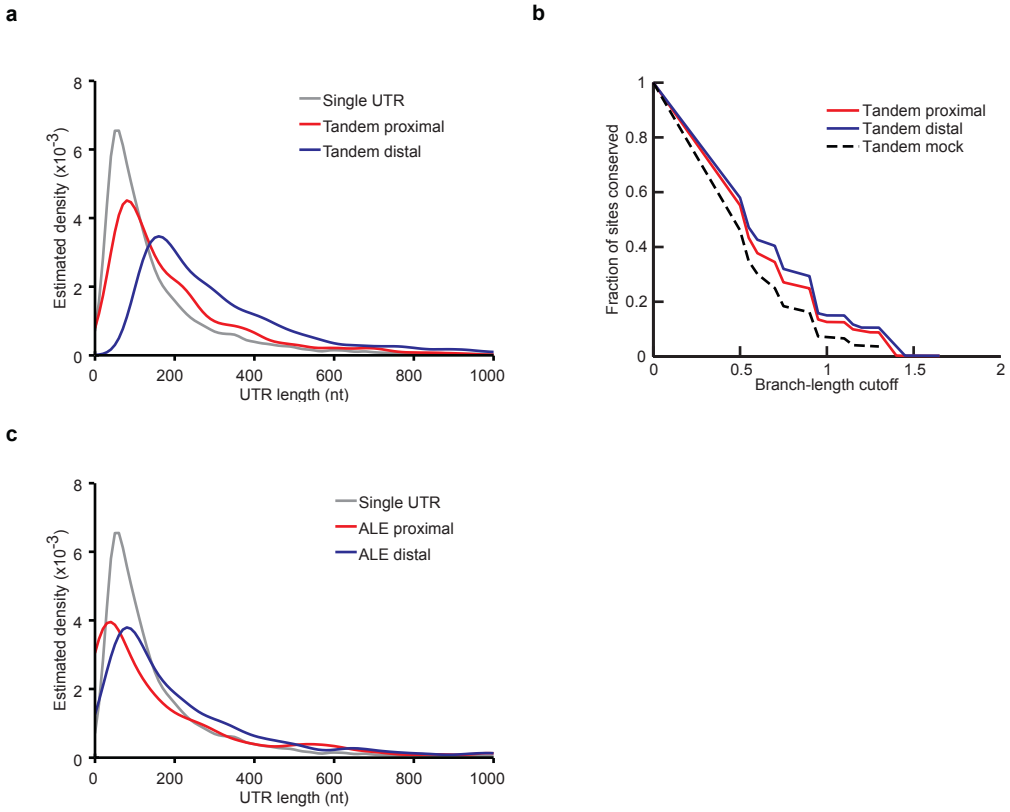
Supplementary Figure 2. Polyadenylation status of histone mRNAs. Shown are RNA blots after probing for 3´ fragments of mRNAs from the indicated core histone genes (*his-31* and *his-59*) or a control gene (*rpl-10*). The mRNA 3´ fragments were generated by oligonucleotide-directed RNase H cleavage – GTCTTTCTCTTGGCGTGCTC (*his-31*), CACAAGTTGGTGTCCTCGAA (*his-59*), TGATTTGACGTCCTGGGAACTTG (*rpl-10*) – with or without oligo(dT), which was added to direct digestion of the poly(A) tail. Above each blot are schematics indicating the distance between the RNase H cleavage site (black bar) and either the 3P-mapped cleavage site (above the mRNA representation) or the conserved stem-loop found at the ends of core-histone mRNAs (below the UTR representation). Also shown in red are the positions of the probes: GACGCTTCAGAGCATAGACGACGTCC, CAGAAGATAAATGAATTATCCTCCG (*his-31*); CGGTGAGTTTTGAGTTGAAGCTTAC, CTACCATAAGGTATTTAAGCGCGC (*his-59*); TAGTCTTCGC-GATCCCACTTGG, GAGTTGAACTCCAACTCCGTCG (*rpl-10*). For each of the histone mRNAs the size of the fragment and absence of a change in mobility after RNase H digestion with oligo(dT) indicated that the major steady-state product lacked a poly(A) tail and terminated immediately after the stem-loop.

The replication-dependent histone mRNAs differ from other mRNAs in their 3´-end formation. In most metazoans, the core histones H2A, H2B, H3, and H4 as well as the linker histone H1 mRNAs are unspliced, capped and non-polyadenylated mRNAs with a highly conserved stem-loop at their 3´ ends[63]. This stem-loop structure is recognized by SLBP, which is indispensable for proper expression and cell-cycle regulation[63]. Histone pre-mRNAs are recognized through characteristic histone downstream elements (HDEs), purine-rich sequences that base pair to the pyrimidine-rich 5´ end of U7 snRNA to recruit shared components of the canonical cleavage and polyadenylation machinery (CPSF-73, CPSF-100, Symplekin)[63].

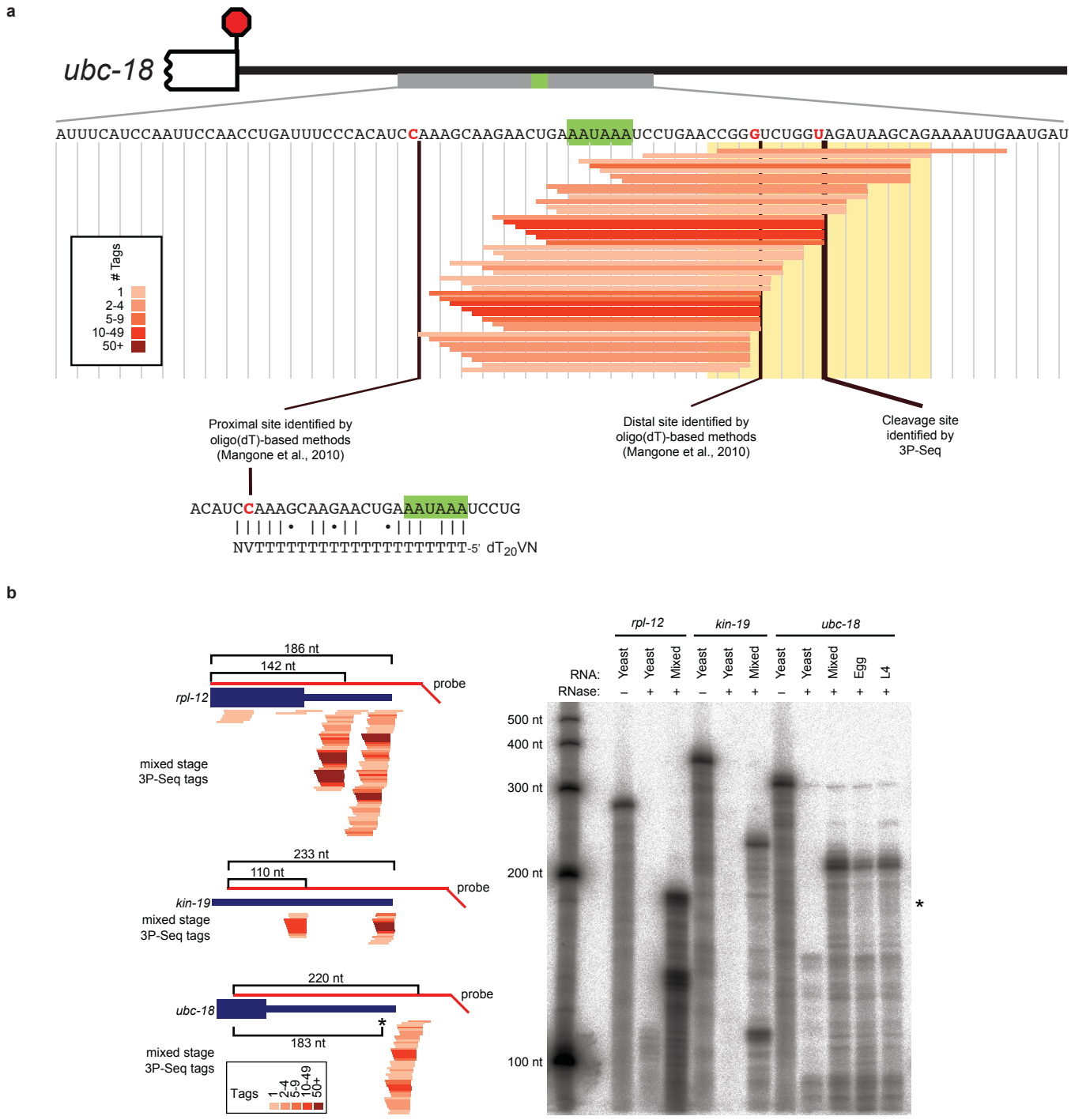
In *C. elegans*, the histone H1 orthologs appear to be spliced, polyadenylated and lack the well conserved stem-loop. This is perhaps due to the compact genome size and high gene density, which lessens the need for H1-dependent higher-order chromatin structure. In contrast, H2A, H2B, H3, and H4 are intronless and contain conserved stem-loop structures in their 3´UTRs. Given the significant structural changes in H1 mRNA, we examined whether any of the 71 genes encoding core histone proteins had evidence of polyadenylation. Of these, 31 genes had 3P-Seq tags, providing direct experimental evidence of their polyadenylation. Given the extensive duplication of these genes, we searched for reads that mapped to multiple genomic loci and found that the remaining 40 histone genes each also had tags mapping within the 200 nt downstream of their respective stop codons. Mangone et al. also observed polyadenylation of histone mRNAs[8]. Evidence for some polyadenylation of each of these mRNAs, combined with the absence of an identified U7 homolog in *C. elegans*, raised the possibility that there might be an alternative mechanism for 3´-end formation in nematodes[64]. We hypothesized that *C. elegans* histone mRNAs have evolved to retain regulation through the SLBP, but terminate through canonical cleavage and polyadenylation pathways. Alternatively, the 3P tags we observed at histone mRNAs could originate from products of a back-up pathway, as suggested in *Drosophila*[65]. Our results showing that all the detected mRNA is not polyadenylated and instead ends just after the stem loop support the second scenario, although we cannot rule out the possibility that termination is through canonical cleavage and polyadenylation followed by very rapid trimming back to the stem-loop.



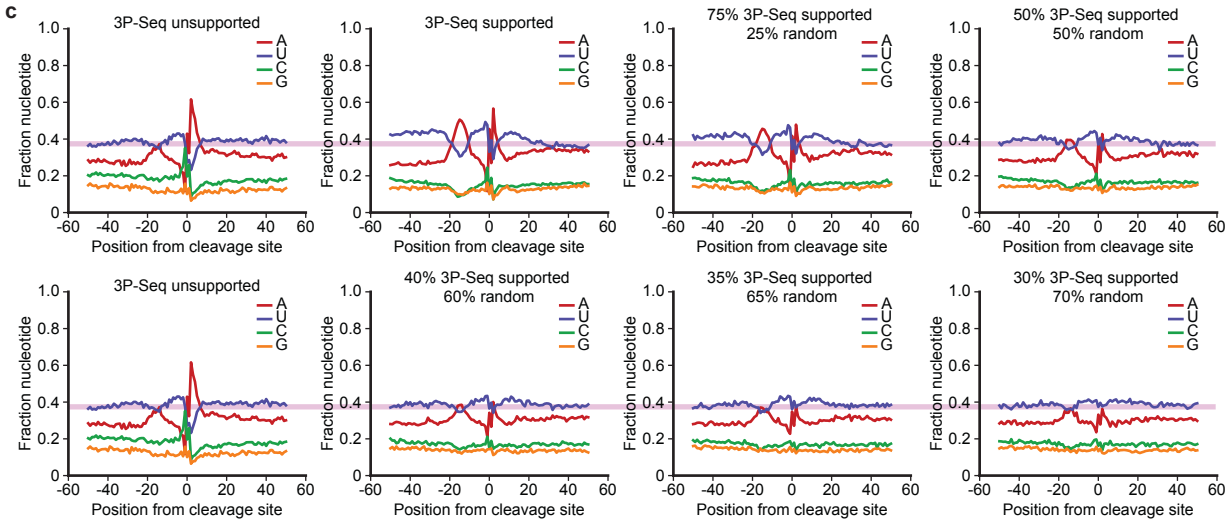
Supplementary Figure 3. Cleavage and polyadenylation signals of *C. elegans* transcripts. **a,** Position of the first nucleotide of the two major polyadenylation signals (AAUAAA, AAUGAA) relative to the cleavage site. We searched for position-specific enrichment of hexamers in the 50 nts upstream of the cleavage site of genes with a single cleavage site, considering first the canonical poly(A) signal (PAS), AAUAAA. The enrichment of AAUAAA centered at position –19 relative to the cleavage site and remained significant compared to a position-specific dinucleotide control ($p < 0.05$, binomial test after Bonferroni correction) at each position extending to –25 and –9. Having calibrated the window in which functional PASs occur, we searched for additional hexamers that were over-represented in this region relative to either an upstream control region or the first-order Markov expectation in the same window. Both control methods recovered similar motifs, which agreed well with previous studies of smaller EST-based datasets[66] and indicated the relative importance of A2, U3, A/G4, A5 and A6 within the hexamer (Supplementary Table 4). **b,** Conservation of poly(A) sites. Plotted are average phastCons scores (± s.e.m.) for poly(A) sites, AAUAAA (red), AAUGAA (blue), or the remaining significantly enriched motifs (Supplementary Table 4) (green). Nucleotide-level conservation was assessed using phastCons scores obtained from the UCSC genome browser[36]. **c,** The nucleotide sequence composition near end regions, depicting complexes implicated in cleavage and polyadenylation. Otherwise, as in Figure 1e. Factors are colored based on whether they recognize A-rich (red) or U-rich (blue) elements. The peak in U enrichment ~10 nt downstream of the cleavage site presumably represented the Downstream Element (DSE), which in other species binds the CstF complex[30]. In other species the DSEs can be either U or G/U rich[30,67], but in *C. elegans* it appeared to be enriched only in U. The U enrichment upstream of the PAS presumably reflected the presence of Upstream Elements (USEs), which in other species act at variable distances upstream of the cleavage site to bind the CFI(m) heterodimer, which stabilizes CPSF binding[30,68]. The region between the PAS and cleavage site might serve as a binding site for Fip-1, which in humans binds U-rich sequences and interacts with poly(A) polymerase to stimulate polyadenylation[69]. In human, cleavage, catalyzed by CPSF73, preferentially occurs at an adenosine[70]. The *C. elegans* enzyme appeared to have similar preferences. **d,** Nucleotide sequence composition near cleavage sites lacking a common PAS. Plotted are the nucleotide frequencies relative to cleavage sites for single UTRs lacking a PAS variant (Supplementary Table 4). Enrichment for adenosine is observed at the same location as PASs, but is reduced compared to single-UTR genes, consistent with the lack of a common PAS variant. U-rich regions containing presumptive USE, DSE and Fip1 elements appear slightly exaggerated compared to those of single-UTR cleavage sites with common PASs (Fig. 1e).

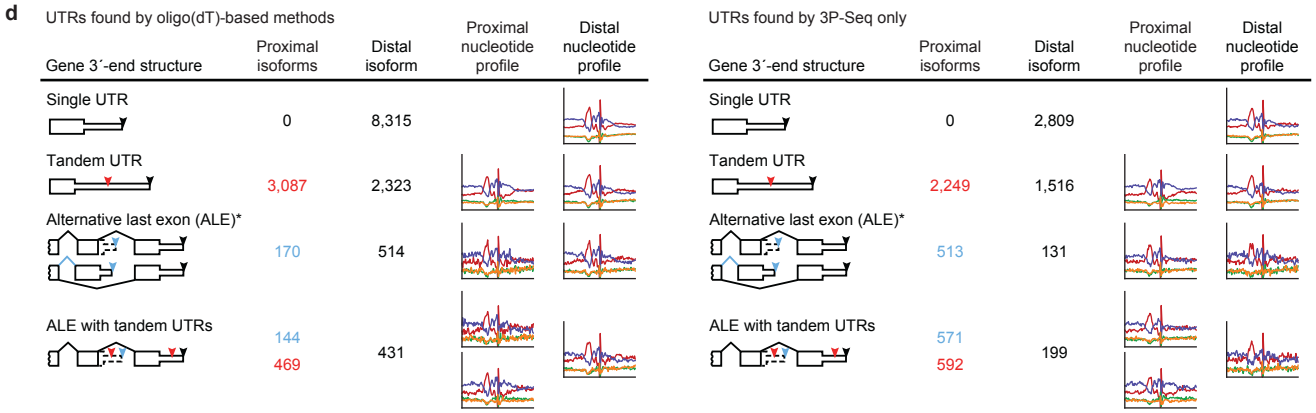


Supplementary Figure 4. Properties of UTRs from alternatively polyadenylated genes. **a,** Length distribution of single UTRs and of proximal and distal tandem UTRs. Distal tandem UTRs tended to be longer than single UTRs ($P < 10^{-300}$, Wilcoxon rank-sum test). **b,** PAS conservation in proximal and distal tandem UTRs as a function of conservation stringency (branch-length cutoff). As a control, conservation of nonfunctional PAS motifs, defined as those falling within UTRs but ≥ 30 nt from 3P tags, is also plotted (control). PASs were considered conserved if a PAS variant was found within 10 nucleotides of the C. elegans PAS in the multiple alignment. Mock PASs were drawn from the top 10 PAS hexamers associated with cleavage sites and matched for the frequency of usage for each hexamer. Both proximal and distal tandem PASs were more conserved than controls ($P < .005$, two-sided Kolmogorov-Smirnov test) and distal more conserved than proximal ($P = 2.5 \times 10^{-5}$, two-sided Kolmogorov-Smirnov test) **c,** Length distribution of single UTRs and of UTRs from proximal and distal ALEs. Proximal ALEs possess UTRs shorter than either single UTRs or distale ALEs ($P < 10^{-5}$ and $< 10^{-14}$, respectively, Wilcoxon rank-sum test).



Supplementary Figure 5. Potential artifacts resulting from internal oligo(dT) priming. a, 3P-Seq data supporting the distal cleavage site but not the proximal cleavage site identified by oligo(dT)-based methods[8]. Locations of mixed-stage 3P tags, colored as in Figure 1d, are shown below a schematic of the *ubc-18* locus. The stop codon (red octagon) indicates the end of the ORF. The best candidate PAS is boxed in green. Cleavage sites annotated using oligo(dT)-based methods of Mangone et al., (2010) or 3P-Seq are indicated in red and labeled below. The region containing the cluster of 3P-defined cleavage sites that identified the end of the *ubc-18* 3´UTR is shaded in yellow. Shown below is a segment of the *ubc-18* 3´UTR immediately downstream of the cleavage site of the short isoform reported by Mangone et al. (2010). The potential pairing shown between this segment and oligo(dT20)VN primers used in Mangone et al. (2010) would lead to misannotation of a proximal cleavage site. **b,** Ribonuclease-protection assay confirming 3P-Seq UTR annotations. The positions of probes and 3P tags are shown to the left, with the expected sizes of protected fragment indicated for each isoform. Protected fragments with the predicted lengths supported each isoform identified by 3P Seq, whereas protected fragments were not observed at the length predicted for the reported proximal isoform of ubc-18 (*), even in stages for which this isoform is reported to be 65% or 40% of the total *ubc-18* mRNA (egg and L4, respectively)[8].

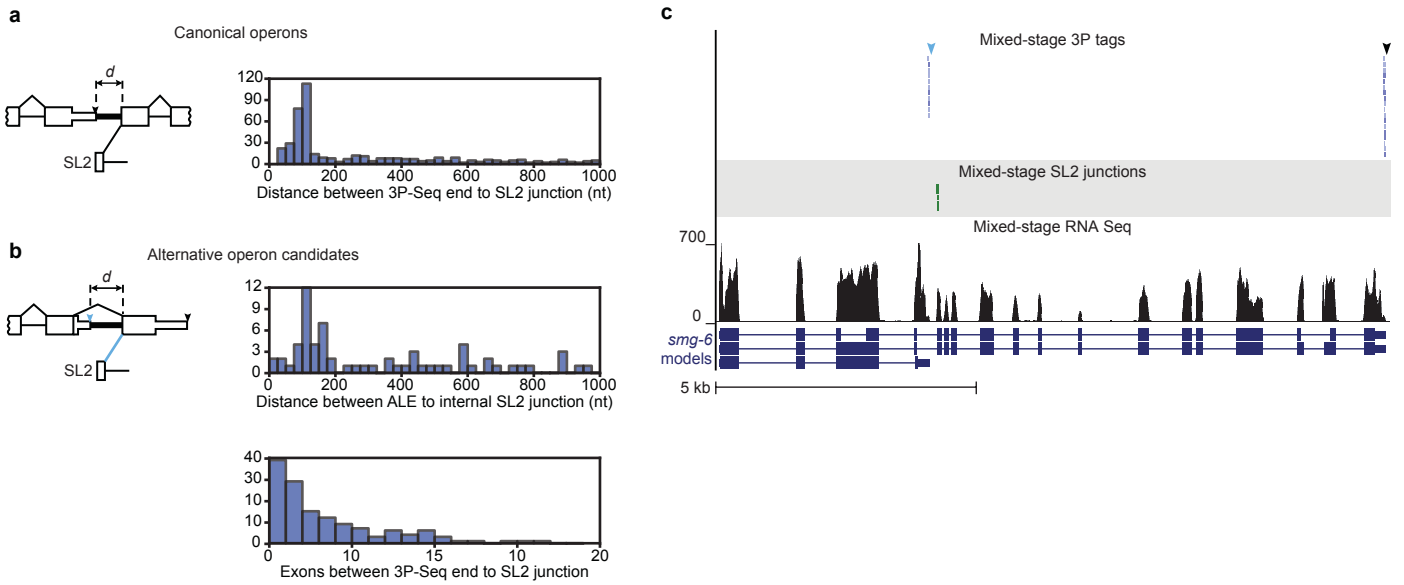




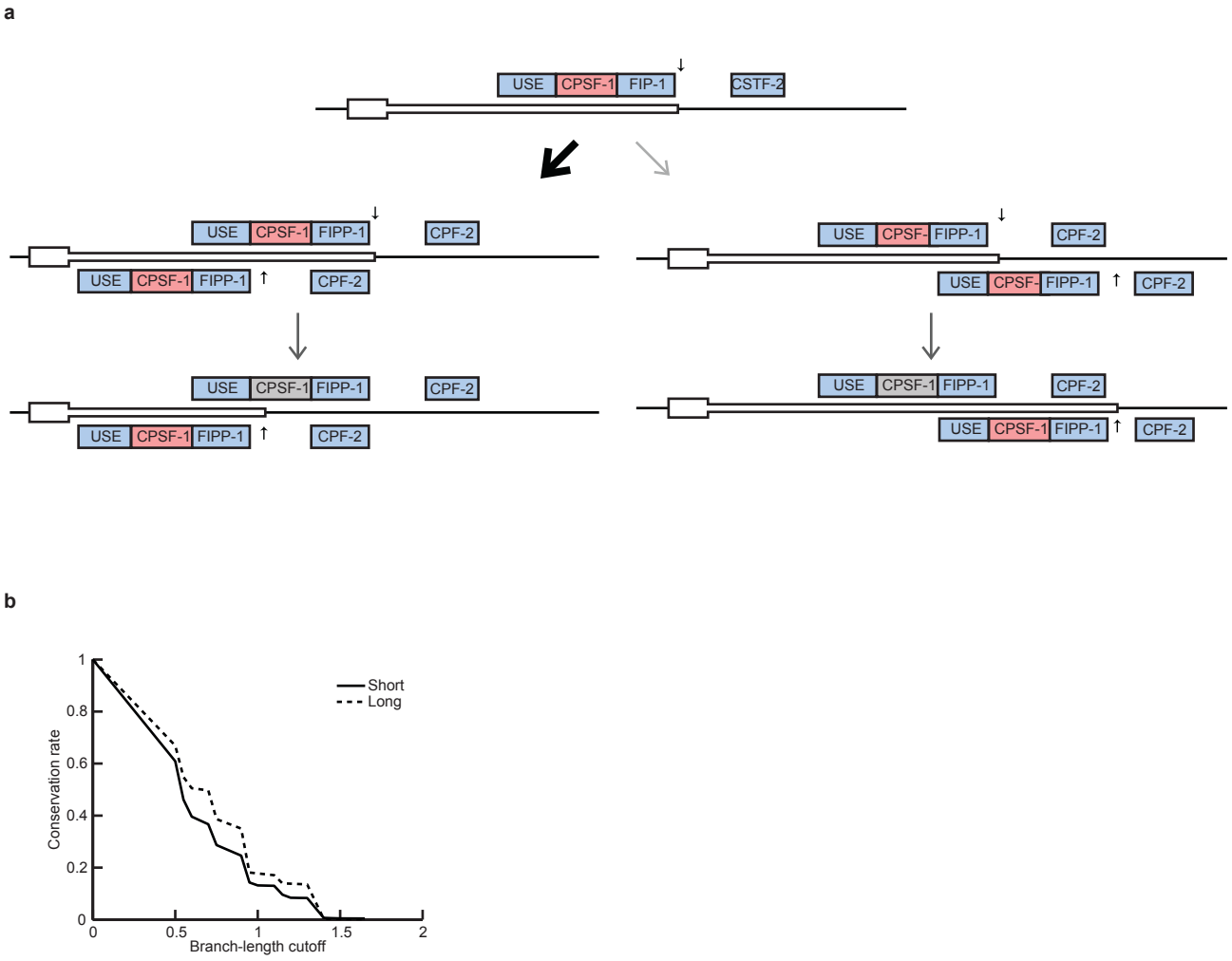
Supplementary Figure 5 (continued) c, Nucleotide composition surrounding cleavage sites of proximal isoforms of tandem UTRs reported by Mangone et al. (2010), comparing nucleotide composition near sites that were also identified by 3P-Seq to that of sites that were not identified by 3P-seq. Composition near the subset of UTRs that were not 3P-Seq supported is shown in the left-most plot of each row. Composition near the subset of UTRs that were 3P-Seq supported is shown in the second plot of the first row. To estimate the fraction of false positives, UTRs with 3P-Seq support were mixed in varying proportions with random sequence from single UTRs tallied in Figure 2a. A reference line (pink) is shown across all profiles to facilitate comparison. The nucleotide profile of proximal tandem isoforms without 3P-Seq support best matches that of the mixture containing 70% random sequence, indicating that 70% of these ends are likely false-positives. In this analysis we considered only UTRs for which the coordinate of the first and last base were provided in the dataset of Mangone et al. (2010). Because multiple UTRs could be assigned the same cleavage site, UTRs were first consolidated, for each cleavage site selecting the UTR with the nearest 5´ end. Tandem UTRs were then identified as UTRs that had the same 5´ end but different 3´ ends. This procedure yielded a set of unambiguously reported tandem UTRs, which contained 4,640 distal isoforms and 7,315 proximal isoforms. The fraction also identified by 3P-Seq was greater for the distal isoforms (4,107/4,640) than for the proximal isoforms (4,805/7,315). This panel compares the nucleotide composition of the 4,805 3P-supported proximal sites to that of the remaining 2,510 proximal sites. When the modENCODE consortium determined mRNA 3´ends, they reported more proximal isoforms identified using the Mangone et al. dataset (Supplementary Table 3)[11]. One reason they found more is because they mapped cleavage sites onto their new exon models, which enabled them to use sites with missing stop-codon annotations. Another reason is that they also used the 3P-Seq data, which would have enabled them to identify distal isoforms for some UTRs that Mangone et al. annotated as single UTRs. These reasons also help explain why the consortium reported more distal isoforms identified using the Mangone et al. dataset, although the main reason for more distal isoforms is that single UTRs are include in the distal isoform tallies (Supplementary Table 3).

Our interpretation of the results of this supplemental figure is that ~70% of the proximal sites reported uniquely by Mangone et al. are false positives. An alternative interpretation is that the class of proposed proximal sites that lack a conventional PAS and instead have an A-rich motif immediately following the site was not found by 3P-Seq because sites immediately followed by an A-rich motif would not be represented by reads with an untemplated A. This alternative interpretation might seem feasible if most of the sites were followed by a homopolymeric run of more than six adenylates (Fig. 1b). However, out of concern for internal-priming artifacts, Mangone et al. appear to have filtered such sites from their dataset[8]. Perhaps as a result of this filtering, the class of putative proximal sites is not enriched for AAAAAA immediately following the sites, despite the A-rich composition of this region[8]. Moreover, 90% of the cleavage sites identified by 3P-Seq had support from at least one tag with more than six adenylates (Supplementary Fig. 1e). Thus, for nearly all the sites uniquely reported by Mangone et al., our procedure would have generated multiple 3P tags that met the criterion of possessing at least one untemplated terminal A. Another explanation for why 3P-Seq might have missed this class of proximal sites would be if isoforms with these sites were not present in the stages we sampled. However, this explanation is ruled out because most of these isoforms are reported to be expressed in multiple stages examined in our study[8]. The notion that this class of isoforms is comprised of false-positives might seem incongruent with the report that the ubc-18 proximal isoform appeared differentially regulated[8]. However, internal-priming artifacts might be particularly sensitive to sample-to-sample variation in reaction conditions, such as temperature or the amount of primer in excess over template, and RPA analysis showed that this proximal isoform was absent even in a stage where it is reported to be the most abundant isoform (**b**).

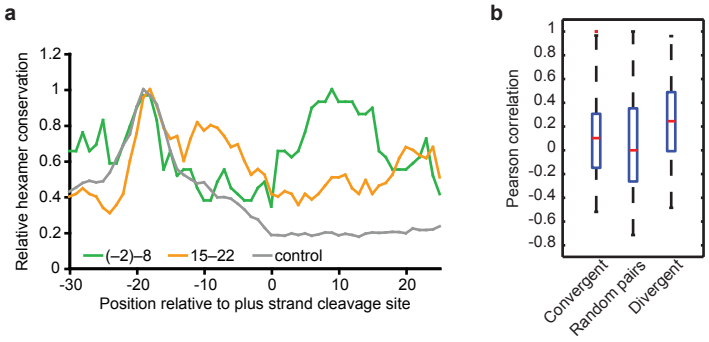
d, The distribution of UTRs among different types of alternative isoforms with or without support from oligo(dT)-based methods. Left, the distribution of the 15,455 3P-Seq–supported UTRs that were also found by oligo(dT)-based methods[8]. For genes with ALEs that have tandem isoforms (bottom), the ALE tally indicates the number of distal isoforms of proximal ALEs (blue) and the tandem tally indicates the proximal tandem isoforms of all ALEs (red). In all cases, the distal isoform is the 3´-most cleavage site for each gene (black arrowhead). The nucleotide composition at the end regions of proximal and distal UTRs is shown for the 50 nucleotides on either side of the cleavage site. Right, the distribution of the 8,581 UTR isoforms found only by 3P-Seq but not by oligo(dT)-based methods.



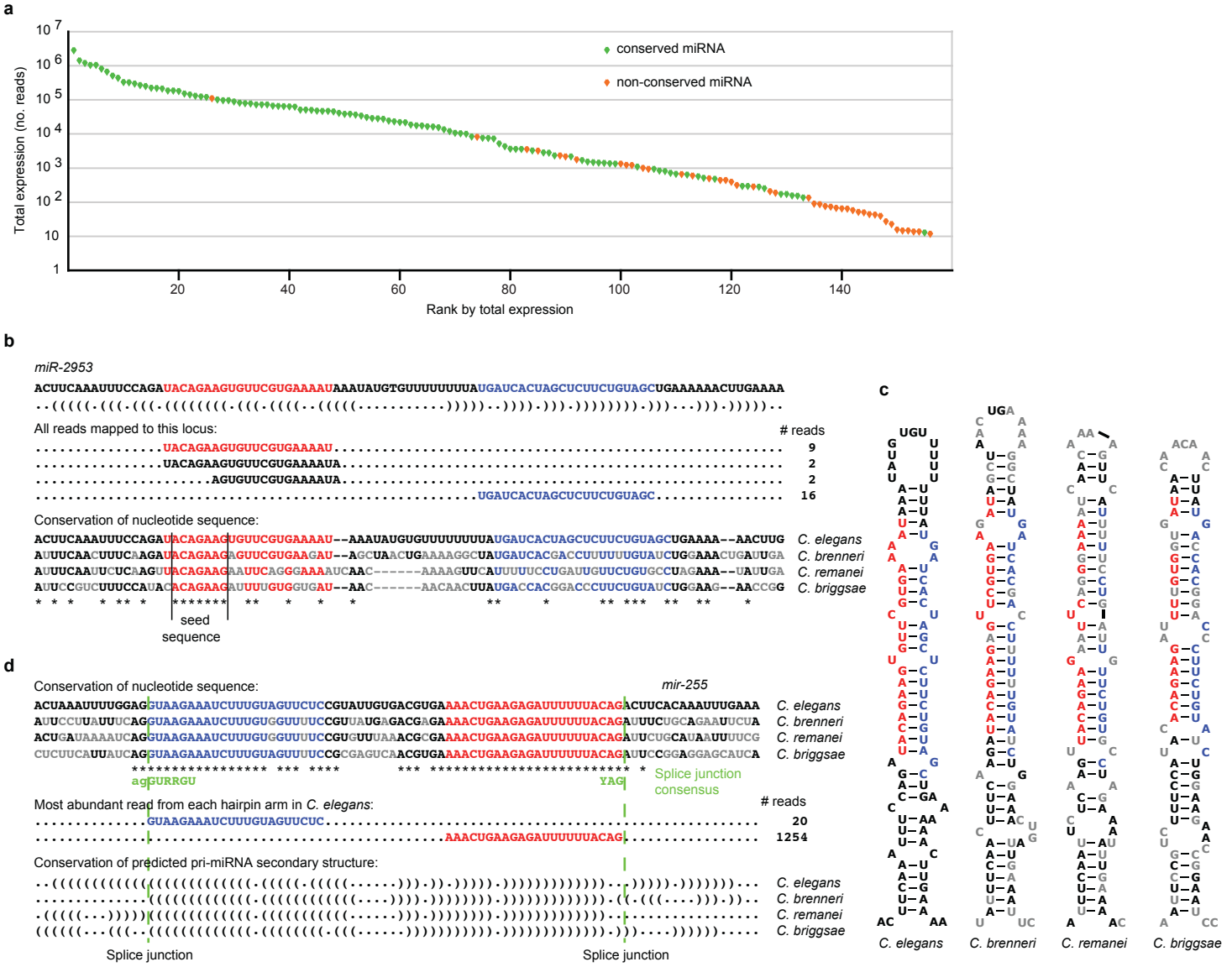
Supplementary Figure 6. Candidate alternative operons share some properties with canonical operons. **a,** Distribution of distances (*d*) between cleavage sites and SL2-trans splice sites for canonical SL2 operons. **b,** Distribution of distances between ALE cleavage sites and SL2-trans splice sites for candidate alternative operons (top) and distribution of the number of annotated exons that are contained between ALE cleavage sites and the nearest downstream SL2 trans-splice site in the same gene. These candidate alternative operons were hand-curated to identify those most likely produce trans-spliced isoforms encoding functional proteins (Supplementary Table 5). **c,** An example of an alternative operon. 3P tags were mapped relative to the RefSeq *smg-6* gene models and to RNA-Seq tags corresponding to SL2 junctions (green bars) and the genome (black histogram). Distal and proximal cleavage sites are indicated (black and red arrowheads, respectively). Smg-6 is involved in the nonsense-mediated decay (NMD) pathway, but unlike other *smg* genes, *smg-6* is essential, suggesting functions beyond NMD[71]. Usage of the *smg-6* ALE produces an mRNA encoding a C-terminal–truncated protein that lacks the PIN and EST domains. This ALE has a conserved PAS and alternative stop codon, suggesting conserved production and function of the short isoform. The 3´-splice site immediately downstream of the ALE is spliced to SL2, and a conserved AUG in this exon restores the original reading frame of the full-length transcript, which codes for an N-terminal truncation that includes the PIN and EST domains.



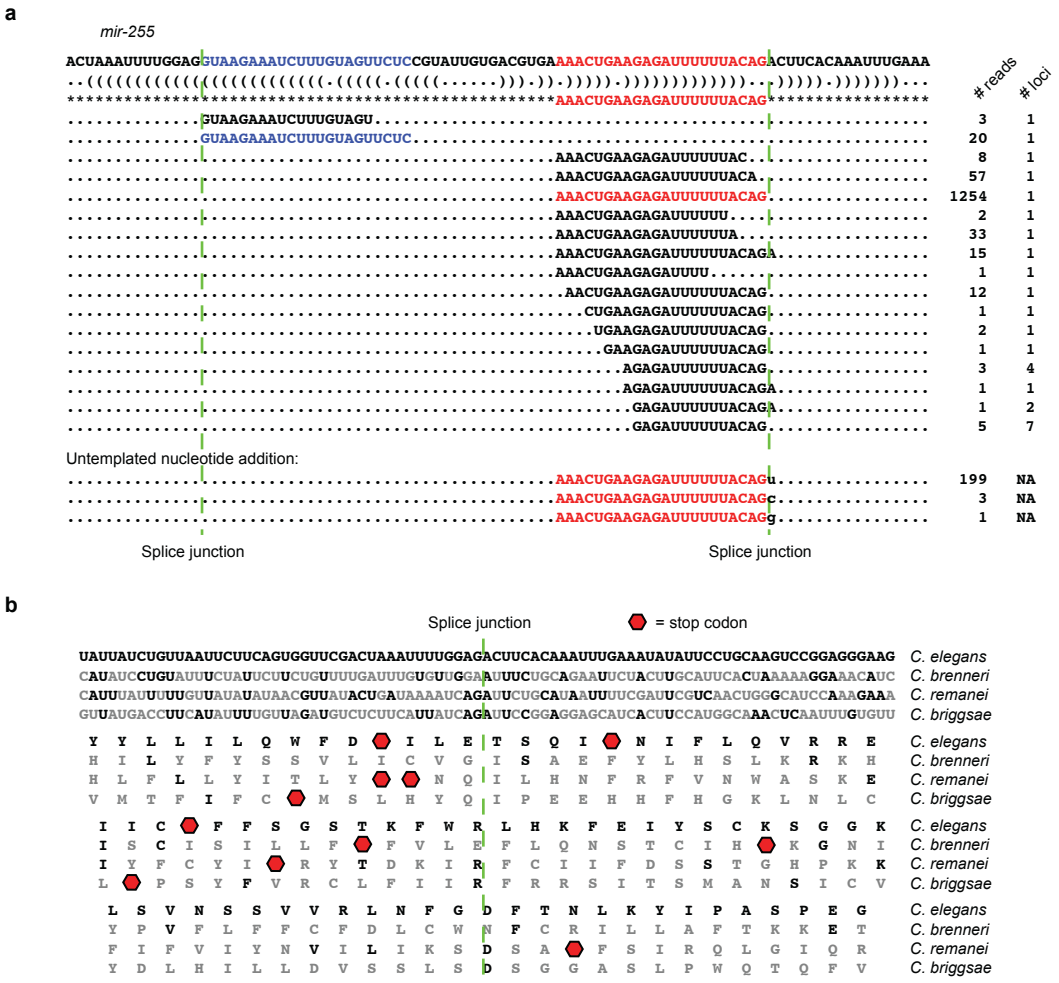
Supplementary Figure 7. Overlapping end regions as evolutionary intermediates of progressive UTR shortening. **a,** Model for changes in UTR length through emergence and loss of OERs. Alternative cleavage sites emerge preferentially where U-rich elements (blue) can bind either the original factors (indicated above the UTR) or the alternative factors (indicated below the UTR), thereby generating OERs. Mutations (grey) that disrupt function of the original site result in a single UTRs that is shorter (left pathway) or longer (right pathway) than the original. Random mutations are more likely to create a function PAS within long U-rich region upstream of cleavage sites (Fig. 1e, Supplementary Fig. 3c), than in other regions that have a lower overall U-richness, thereby favoring the left pathway. **b,** PAS conservation as in Supplementary Figure 4b for genes with single UTRs that are either short (10–80 nt) or long (>300 nt) at different conservation stringencies. PASs from long UTRs are more conserved than those from short UTRs (P = 1.4×10^{-16}, two-sided Kolmogorov-Smirnov test)



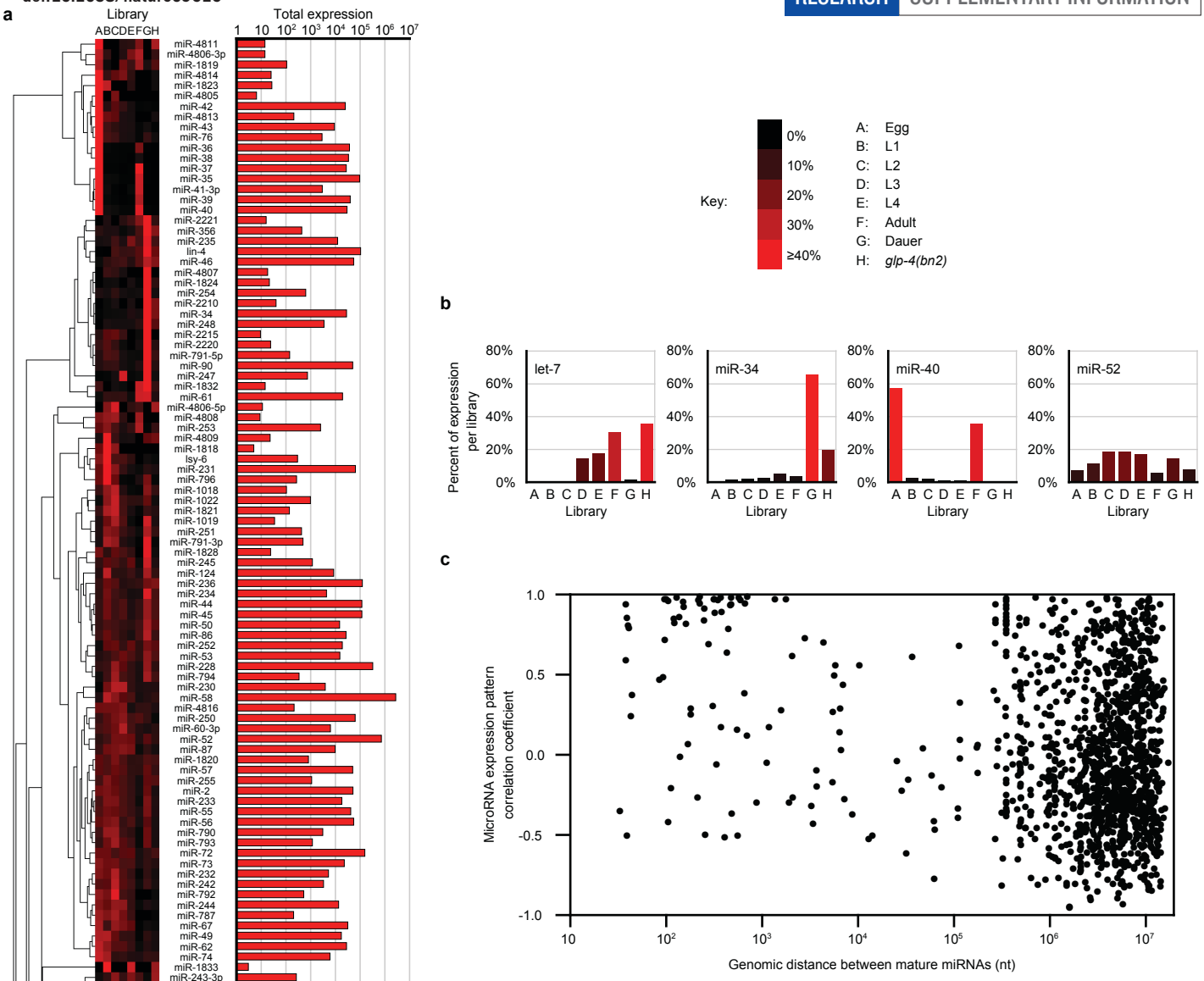
Supplementary Figure 8. Palindromic arrangement of cleavage and polyadenylation elements. a, Conservation of hexamer UTR segments for convergent UTRs with different amounts of overlap. The UTR segments start at the indicated positions relative to the (+)-strand cleavage site, for genes with 15–22 nts or (–2)–8 nts of overlap (Fig. 3e) or no overlap, as indicated. Hexamers were considered conserved if at least two other nematodes had the same hexamer at the same position in 6-way multiple alignments (Supplementary Fig. 12a)[36]. Conservation is plotted relative to that of highest observed for each UTR set. **b,** Correlation coefficients for mRNAs expressed from gene pairs with the indicated genomic arrangements. SAGE data were obtained from the Genome BC *C. elegans* Gene Expression Consortium http://elegans.bcgsc.bc.ca/.



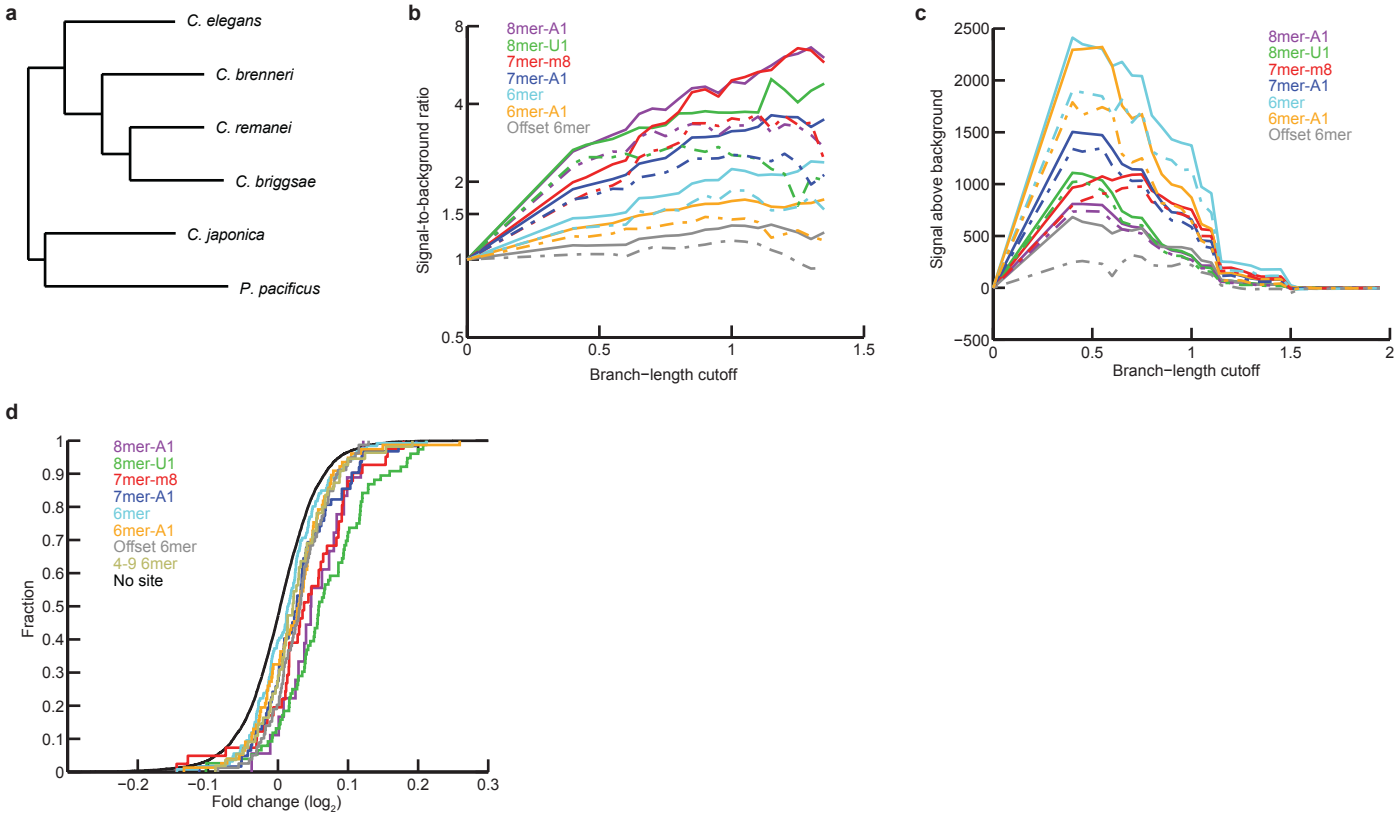
Supplementary Figure 9. MicroRNA sequencing and conservation in C. elegans. a, Illumina sequencing reads for confidently annotated miRNAs ranked by read count (conserved miRNAs, green; miRNAs not found in other sequenced genomes, orange). **b,** Sequenced small RNAs aligned to the *mir-2953* pri-miRNA sequence, accompanied by its bracket-notation secondary structure. The miRNA is shown in red and the miRNA* in blue. The number of reads observed for each sequence is indicated. Below: An alignment of the *Caenorhabditis elegans mir-2953* gene sequence with orthologous sequences from *Caenorhabditis brenneri*, *Caenorhabditis remanei*, and *Caenorhabditis briggsae*[36]. Nucleotides or gaps that did not match the *C. elegans* sequence are in grey. The conserved segment containing the miRNA seed is indicated. **c,** The predicted hairpin structures of *mir-2953* orthologs, colored as in (**b**). **d,** Conserved mirtronic properties of the *mir-255* locus. At the top is an ungapped alignment of orthologous genome fragments from C. elegans and other nematodes, with the miRNA and miRNA* colored in red and blue, respectively, and nucleotides not matching the *C. elegans* sequence in grey. In the middle are the sequences with the most abundant reads from each arm of the pre-miRNA hairpin. at the bottom are predicted secondary structures for each of the orthologous loci shown in bracket notation as an ungapped alignment. Inferred intron/exon boundaries are indicated with green lines. Intron/exon junction consensus sequences[72] are in green text.



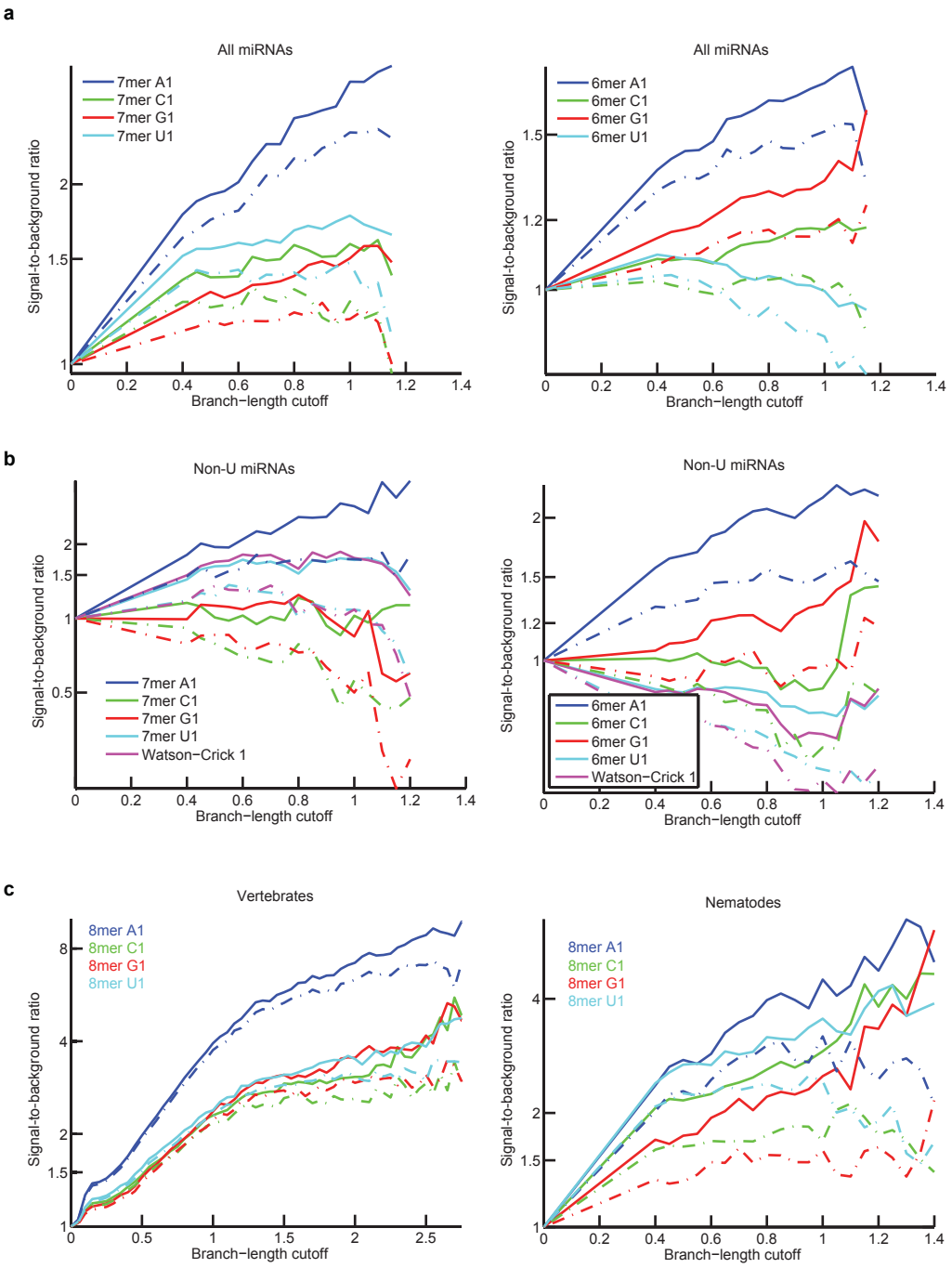
Supplementary Figure 10. Evidence that *mir-255* is a mirtron not hosted by a protein-coding mRNA. a, Sequenced small RNAs aligned to the *mir-255* precursor sequence, accompanied by its bracket-notation secondary structure. The miRNA is shown in red and the miRNA* in blue. The number of reads observed for each sequence, as well as the number of loci to which each sequence perfectly matched in the *C. elegans* genome, are indicated. At the bottom, sequenced miRNA variants with 3´-untemplated nucleotides are shown, with their read numbers indicated. Inferred intron/exon boundaries are indicated with green lines. **b,** Alignment and translation of the spliced exons flanking the mirtron. The inferred exon/exon boundary is indicated by the green line. Three amino-acid alignments are shown, each with the RNA translated in a different frame. Nucleotides and amino acids differing from those of *C. elegans* are in grey.



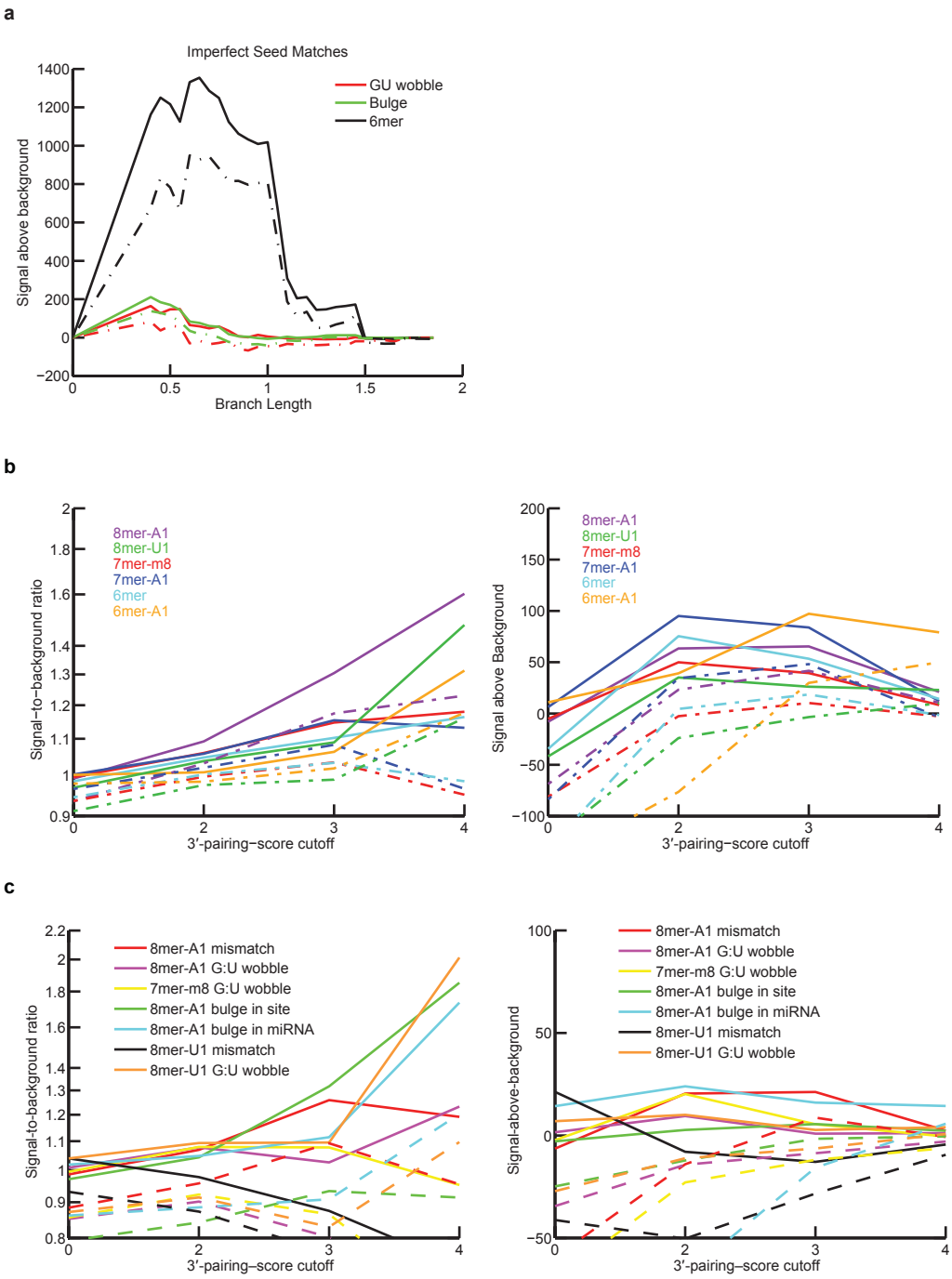
Supplementary Figure 11. Expression profiles of *C. elegans* miRNAs. a, The expression profiles and read numbers for all miRNAs across the eight libraries. For each library, miRNA reads were normalized to the total number of *C. elegans* genome-matching reads from that library, and the fraction of normalized read counts for each miRNA in each of the eight indicated libraries are indicated by red color intensity (key). The histogram on the right indicates the total number of non-library-normalized reads contributing to each expression profile. Reads from the mixed-stage library were not included, nor were miRNAs that were only observed in the mixed-stage library. Expression profiles were clustered using Cluster[73] and displayed with MapleTree (L. Simirenko). **b,** The expression profiles, normalized as in (**a**), for four representative miRNAs with developmental expression pattern previously determined across all but the dauer stage by RNA blots[29]. **c,** A plot of the distance between miRNA genomic loci versus the correlation of their expression profiles as defined in (**a**). Each point refers to a pair of miRNAs from (**a**) that derive from the same strand of the same chromosome. The x-axis indicates the distance between the centers of the mature miRNAs when mapped to the genome. The y-axis indicates the Pearson correlation coefficient between the normalized expression patterns of the two miRNAs.



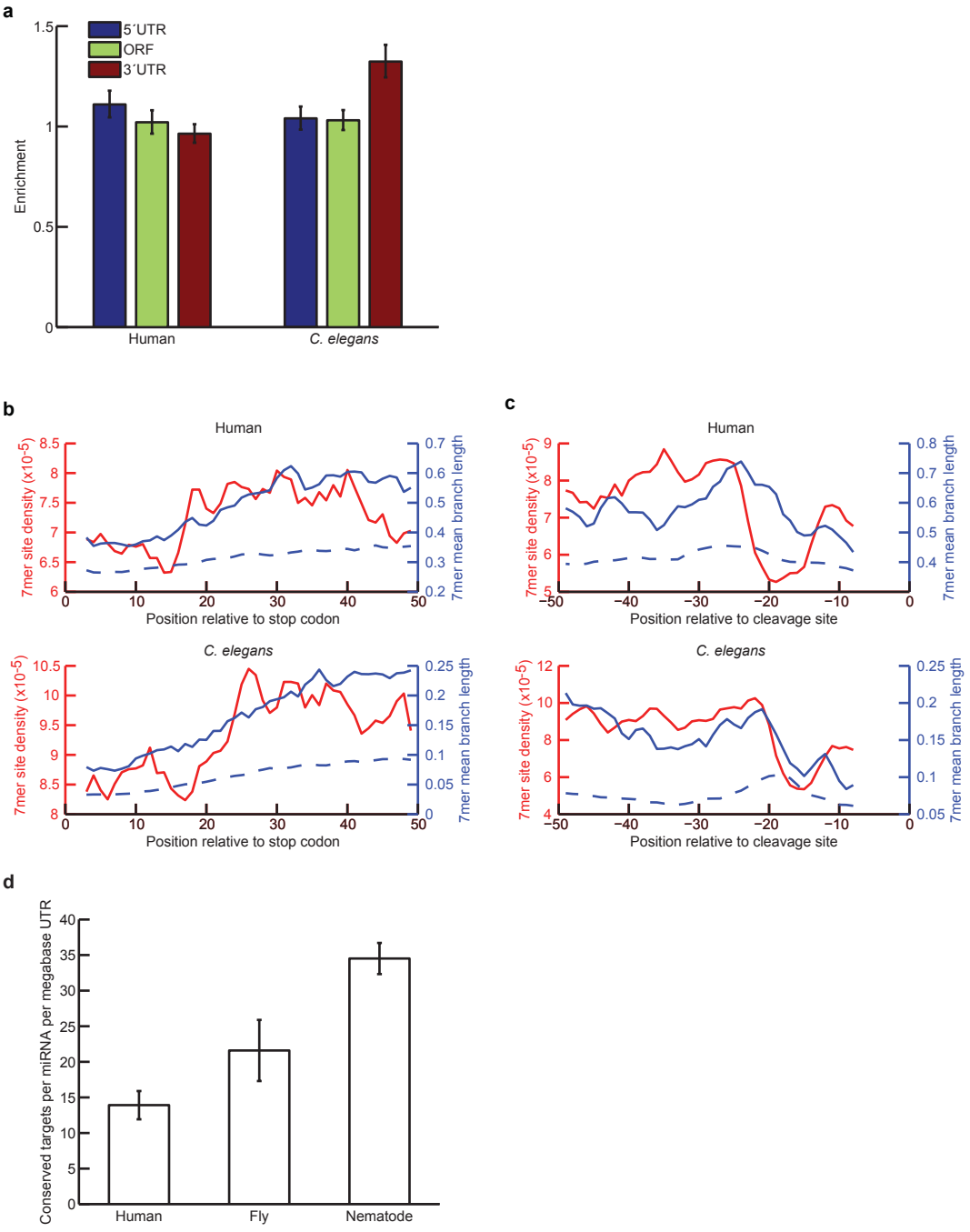
Supplementary Figure 12. Conservation and function of an expanded repertoire of miRNA sites in nematodes. a, Phylogeny of 3′UTR regions of nematodes with sequenced genomes. The phylogeny was based on regions that align to *C. elegans* 3′UTRs in multi-Z alignments extracted from the UCSC genome browser. Branch lengths were fit using the dnaml component of the PHYLIP software package. **b,** Preferential conservation of various types of miRNA sites, plotted as a function of conservation stringency (branch-length cutoff). Broken lines indicate 5% lower confidence limit estimated using cohorts of suitable control sequences (*z* test). Site types are colored as in Figure 4a. **c,** Number of miRNA sites conserved above background, plotted as a function of conservation stringency, otherwise as in (**b**). The six major types each had >600 sites confidently conserved above background. Combining the signals and backgrounds from each type at a sensitive branch-length cutoff of 0.5 yielded 9,092 ± 145 sites conserved above background, which represented our lower bound for the number of selectively maintained miRNA sites. This corresponded to an average of 0.56 ± 0.01 sites per *C. elegans* UTR. To estimate the number of conserved targets, we randomly selected conserved sites totaling the signal above background for each site type at each UTR conservation level and asked how many genes were represented by the target sites. This procedure yielded 4,488 ± 781 genes with conserved seed matches. **d,** Changed mRNA levels after knocking out *mir-124* in *C. elegans*. Shown are analyses of mRNA array data[20], plotting the cumulative distribution of mRNA changes. Each mRNA in a set had exactly one 3′UTR miR-124 site of the indicated type and no other miR-124 sites of any type. Each of the distributions significantly differed from the no-site distribution ($P < 0.03$, two-sided Kolmogorov-Smirnov test). Site types are colored as in Figure 4a. Microarrays were analyzed as described[56].



Supplementary Figure 13. Effect of position 1 identity on miRNA sites in nematodes. a, Effect of adenosine at position 1 for 6mer and 7mer sites. Signal-to-background ratio for conservation of matches to the miRNA flanked by the indicated nucleotide opposite position 1, plotted as a function of conservation stringency. 7mer matches (left) were Watson-Crick matches to miRNA nucleotides 2–7 (but not 8), and 6mer matches (right) were Watson-Crick matches to miRNA nucleotides 2–6 (but not 7). Broken lines indicate 5% lower confidence bounds (*z* test) and show that for both site types matches flanked by an adenosine were substantially more conserved. **b,** As in (**a**) but for only the seven conserved nematode miRNA families that do not start with a U. Ratios for sites with Watson–Crick matches to position 1 were also plotted. **c,** Effect of adenosine at position 1 in vertebrates and nematodes. Signal-to-background ratio for conservation of matches to nucleotides 2-8 flanked by the indicated nucleotides opposite position 1, plotted as a function of conservation stringency. Broken lines indicate 5% lower confidence limit (*z* test).



Supplementary Figure 14. Conservation of imperfect seed matches, supplemental and compensatory 3′ pairing. a, Conservation of 8mer sites with a single-nucleotide bulge in the target strand or a single G:U wobble pair. Signal above background is plotted as a function of conservation stringency, as in (Supplementary Fig. 12c). For comparison, signal above background is also plotted for canonical 6mer sites, which have perfect matches to miRNA positions 2–7. **b,** Conservation of 3′-supplemental pairing. Signal-to-background ratio (left) and signal above background (right) are plotted for conservation of the 3′ pairing supplementing conserved sites of the indicated types. The branch-length cutoff was 0.5, which yielded the signal above background with greatest statistical significance. Sites with a 3′-pairing score ≥5 were too few to be plotted. Broken lines indicate 5% confidence lower bounds (*z* test, using cohorts of chimeric miRNA controls). **c,** Conservation of 3′-compensatory pairing. As in (**b**), but examining conservation of 3′ pairing supplementing sites with the indicated single-nucleotide mismatch, wobble, or bulge. The branch-length cutoff was 1.05, which yielded the signal above background with greatest statistical significance. Sites with a 3′-pairing score ≥5 were to few to be plotted.



Supplementary Figure 15. Density of miRNA targeting in nematodes and humans. **a,** Enrichment of 8mer-A1 sites in human and *C. elegans* mRNA regions. Error bars represent one standard deviation for the cohorts. The only sample with statistically significant site enrichment was *C. elegans* 3′UTRs ($P = 3.5 \times 10^{-4}$, *z* test). **b,** Density and conservation of miRNA sites near the 3′UTR termini. 7mer-m8 site occurrence and conservation near the stop codon. Sites per base of sequence per miRNA family are plotted in red (left axis). Mean conservation branch length of sites is plotted in blue, with mean values for dinucleotide-matched control *k*-mers plotted as a blue dashed line (right axis). **c,** As in (**b**), except that position is relative to the cleavage site. The increased density of *C. elegans* miRNA target sites cannot be explained by increased utilization of the stop-codon proximal or cleavage-site proximal regions. **d,** Density of conserved miRNA sites in different clades. As in Figure 4b, except analysis focused on pairs of species with similar divergence of 3′UTR sequence (*H. sapiens* and *M. domestica*, *D. melanogaster* and *D. willistoni*, and *C. elegans* and *C. briggsae*). Limiting analysis to these pairs facilitated comparison between clades because it prevented differential detection sensitivity from clades with different divergence, phylogenetic topology, and numbers of species. Fly UTRs contain a higher density of conserved sites than human UTRs ($P = 0.0072$, *z* test) and nematode UTRs contain a higher density of conserved sites than fly UTRs ($P = 3.4 \times 10^{-5}$, *z* test).