Comprehensive discovery of endogenous Argonaute binding sites in Caenorhabditis elegans

Dimitrios G Zisoulis1,3, Michael T Lovci2,3, Melissa L Wilbert2, Kasey R Hutt2, Tiffany Y Liang2, Amy E Pasquinelli1 & Gene W Yeo2

MicroRNAs (miRNAs) regulate gene expression by guiding Argonaute proteins to specific target mRNA sequences. Identification of bona fide miRNA target sites in animals is challenging because of uncertainties regarding the base-pairing requirements between miRNA and target as well as the location of functional binding sites within mRNAs. Here we present the results of a comprehensive strategy aimed at isolating endogenous mRNA target sequences bound by the Argonaute protein ALG-1 in C. elegans. Using cross-linking and ALG-1 immunoprecipitation coupled with high-throughput sequencing (CLIP-seq), we identified extensive ALG-1 interactions with specific 3′ untranslated region (UTR) and coding exon sequences and discovered features that distinguish miRNA complex binding sites in 3′ UTRs from those in other genic regions. Furthermore, our analyses revealed a striking enrichment of Argonaute binding sites in genes important for miRNA function, suggesting an autoregulatory role that may confer robustness to the miRNA pathway.

RESULTS

ALG-1 CLIP-seq in C. elegans identifies known miRNA targets

As miRNAs guide Argonaute proteins to specific complementary sequences in mRNAs, we applied the CLIP-seq (also referred to as HITS-CLIP) method11–13 to capture and identify the miRNA and target-site sequences bound by the miRNA complex (miRISC) in developing worms. A recent application of this approach in mouse brain resulted in a map of Argonaute binding sites in this tissue14. C. elegans offers several advantages for applying the CLIP-seq procedure to detect global Argonaute protein–RNA interactions. A single Argonaute protein, ALG-1, is largely responsible for miRNA function, and viable alg-1 genetic mutants exist15. A short but well-established list of miRNA targets expected to be bound by ALG-1 at discrete positions is available16–31. Of these targets, extensive studies have confirmed that lin-41 is regulated by let-7 miRNA during the fourth larval (L4) stage via two clustered sequences, let-7 complementary sites 1 and 2 (LCS1 and LCS2)26–28. We used this example to optimize the CLIP-seq method to detect bona fide ALG-1 binding sites (Supplementary Fig. 1). Synchronized L4-stage wild-type (WT) worms and alg-1(gk214) mutants (hereafter referred to as alg-1(−)), which lack the anti–ALG-1 antibody epitope sequence, were treated with UV irradiation to stabilize in vivo protein–RNA interactions (Supplementary Fig. 1a). A custom antibody specific for the C. elegans ALG-1 protein (Supplementary Fig. 1b) was used to enrich for ALG-1 complexes expected to include miRNA and target RNA species. Immunoprecipitated complexes were processed for isolation of sequences protected by ALG-1 protein from nuclease digestion.

We obtained 3,864,848 and 5,127,241 reads from WT and alg-1(−) CLIP-seq libraries, respectively, out of which 1,651,523 (42.7%) and 695,895 (13.6%) mapped uniquely to the repeat-masked C. elegans genome (Supplementary Fig. 2a). Using MiResque, a microRNA prediction algorithm designed to analyze small-RNA reads obtained from high-throughput sequencing (S. Aigner and G.W.Y., unpublished data),
136 previously reported miRNAs, 37 of which represent the ‘star’ strand, and 1 novel miRNA gene were identified in the WT library (Supplementary Table 1). Heterogeneity in the terminal sequences, which primarily consisted of lost nucleotides from the 3′ ends, could be due to the cloning method, but in two cases we found base additions to the 5′ ends that altered the seed sequence. Identification of the distinct pool of miRNAs bound to ALG-1, which included a large number of star sequences, enabled selective analysis of pairing capacity between miRNAs and mRNA sequences associated with ALG-1 at the stage of sample collection.

To correctly assign CLIP-seq reads to authentic transcribed regions, we reannotated the 5′ and 3′ untranslated regions (UTRs) of gene loci using publicly available 36-bp reads obtained from high-throughput sequencing of poly(A)-selected cDNA libraries from the L3 and L4 stages of C. elegans larval development. Reads that mapped upstream and downstream of currently annotated genes were used to redefine the 5′ and 3′ UTRs. In total, the 5′ and/or 3′ ends of 8,231 genes (40% of genes in the genome) were reannotated by our analysis. The median (average ± s.d.) lengths of bases extended for 5′ and 3′ UTRs are 56 (391 ± 621) and 215 (543 ± 700) nt, respectively. This substantial change in the landscape of C. elegans gene predictions was important for defining the genic location of ALG-1 binding sites and for choosing control sequences for computational analyses (see below).

To distinguish authentic and specific ALG-1 binding sites, we developed a new version of our CLIP-cluster identification algorithm. Briefly, for each of the three biological replicates of the ALG-1 CLIP-seq experiments (WT or alg-1(−)), we first defined ‘regions’ in each gene by extending the sequencing reads to account for the length of the RNA fragments in our CLIP libraries (Supplementary Fig. 2a). To retain biologically reproducible regions while accounting for the different number of sequenced reads in each replicate library, we weighted regions that overlapped across replicate experiments by the fraction of reads in the region relative to all the reads in that experiment mapping within the gene. Regions that passed our stringent threshold corresponded to being reproducible in at least two of three replicate experiments (see Online Methods). Reads within accepted regions were further integrated from replicates to form a ‘cluster’, and clusters containing more reads than statistically expected were kept for further analyses. Finally, clusters that overlapped by at least 25% between WT and alg-1(−) were removed as potential sources of false positives, such as reads from highly abundant rRNA and protein-coding genes (see act-5 gene in Supplementary Fig. 2b).

In total, 5,310 WT and 826 alg-1(−) clusters were identified, 4,806 of which were unique to WT (Supplementary Fig. 2a), representing 3,093 genes, approximately one-fifth of the annotated C. elegans protein-coding genes expressed at this stage in development. Over half of these genes contained a single cluster (Supplementary Fig. 2c). The CLIP-seq results provided a significantly refined and biologically based dataset for identifying miRNA target sites and studying ALG-1 binding properties. Compared to the entire transcriptome, 3′ UTRs only, or 3′ UTRs of miRNAs from miRISC immunoprecipitates, we greatly reduced the search space for functional regions, by a factor of 47, 20 or 5, respectively (see Online Methods for calculation). The tracks for the reannotated gene regions, WT and alg-1(−) reads and clusters are available at the UCSC genome browser (http://genome.ucsc.edu) under ‘ALG1 CLIP-seq’ within the ‘Regulation’ section in the ce6 genome.
Isolation of sequences containing well-established miRNA target sites demonstrates the sensitivity of the ALG-1 CLIP-seq method. Extensive genetic and reporter gene experiments have pointed to LCS1 and LCS2 in the lin-41 3′ UTR as critical sequences for miRNA regulation of this gene. Our ALG-1 CLIP-seq results identified a series of reads forming a significant cluster that maps directly on top of the closely spaced LCS1-LCS2 region (Fig. 1a). Notably, regulation of lin-41 by let-7 miRNA results in substantial mRNA degradation. Thus, the detection of lin-41 by ALG-1 CLIP-seq demonstrates the sensitivity of this method for detecting miRNA targets regardless of regulatory mechanism. The first discovered miRNA target, lin-14, is regulated by lin-4 miRNA via multiple 3′ UTR complementary elements (LCEs). We identified three significant clusters that encompass the proposed LCEs 1–5, and 6–7, respectively (Fig. 1b). Another cluster, toward the end of the lin-14 3′ UTR, is consistent with evidence that this gene is also regulated by other miRNAs. Multiple let-7 binding sites have been predicted to mediate regulation of hbl-1 and daf-12 (refs. 16, 18, 24), and clusters cover a select few of the LCSs in the 3′ UTRs of these genes (Fig. 1c, d). Thus, ALG-1 CLIP-seq provides direct biochemical evidence for predicted miRNA target sites and reveals regions of greater relative occupancy by miRISC within a regulated 3′ UTR.

Of 13 well-established miRNA target genes in C. elegans, all but 3 were found to contain at least one significant 3′-UTR cluster (Table 1). Moreover, the clusters include the cognate miRNA target site for 9 of these 10 genes. The majority of these miRNA target genes were also found to be enriched in ALG-1 interacting proteins 1 and 2 (AIN-1 and AIN-2, members of the GW182 family of proteins) immunoprecipitation experiments. Beyond showing miRISC association with specific endogenous mRNAs, the ALG-1 CLIP-seq dataset contributes nucleotide-level resolution of the actual target region (Table 1 and Supplementary Table 2).

### Genomic and sequence properties of ALG-1 binding sites

Although most genetic and computational studies support a bias for the location of miRNA target sites in 3′ UTRs, functional interaction of miRISC at other genomic positions has also been demonstrated. To study the global distribution of ALG-1 binding in C. elegans protein-coding genes, we mapped the positions of clusters relative to the length of targeted mRNAs. We observed a distinct profile of CLIP-derived cluster (CDC) occupancy proximal to the 3′ ends of spliced mRNAs from WT but not alg-1(−) worms (Fig. 2a). Notably, the frequency of clusters throughout the composite gene model was higher in WT than in alg-1(−) worms, showing that ALG-1 binding extends to other genomic regions (Fig. 2a). Furthermore, the CDC distribution, as a percentage of 3′-UTR length, was not enriched proximal to the stop codon or poly(A) sites, in contrast to the bias for predicted miRNA target sites residing near either end of mammalian 3′ UTRs (Fig. 2b). In fact, the fraction of clusters that mapped a given distance from the stop codon largely mirrored the distribution of 3′-UTR lengths in C. elegans (Supplementary Fig. 3). In total, 1,656 (34.5%) of CDCs were located in 3′ UTRs, 2,473 (51.5%) in coding exons, 602 (12.5%) in introns and 75 (1.6%) in 5′ UTRs.

To characterize the sequence properties of ALG-1 binding sites, we subjected CDCs and a control set of random derived clusters (RDCs) to a battery of computational analyses. In order to perform an equitable comparison as possible, we minimized biases due to GC content, evolutionary conservation, genic region and length of the bound region when selecting RDCs (Supplementary Fig. 4a, b). Furthermore, RDCs were selected from genes depleted of ALG-1 binding sites. Caveats of this approach are that a chosen RDC may actually be bound by ALG-1 at a different developmental stage or that the target mRNA may be present at such low abundance that it is not detected. Our ability to detect the lin-41 LCS1-LCS2 region (Fig. 1a and Supplementary Fig. 1c) despite strong downregulation of this mRNA at the L4 stage suggests that this second point is a minor issue. In spite of the potential limitations for assigning RDCs, our results corroborate expected properties and identify new features associated with miRISC binding to endogenous sequences (CDCs) on a global scale.

Preferential evolutionary conservation is a common feature used to predict miRNA target sites. Indeed, we observed substantially
higher conservation levels within CDCs compared to RDCs in 3′ UTRs (Supplementary Fig. 3a) and a similar trend for coding exon and intron regions (Supplementary Fig. 5a). Also, consistent with the observation that functional miRNA target sites are frequently located in RNA sequences of higher accessibility (in other words, less secondary structure)\textsuperscript{13-47}, the ALG-1–bound regions (CDCs), as well as the 100-nucleotide upstream and downstream flanking sequences, were significantly more accessible than RDCs in the 3′ UTRs (P < 10\textsuperscript{-10}) (Fig. 3b).

However, this was not true for CDCs in the other genomic regions (Supplementary Fig. 5b). It has also been suggested that a high local AU content is responsible for the more accessible 3′ UTR sites targeted by miRNAs\textsuperscript{48,49}. Thus, we analyzed the nucleotide composition within and 100 nucleotides upstream and downstream of 3′ UTR CDCs to search for motifs statistically enriched relative to RDCs. Unexpectedly, the ten most enriched 5- to 7-mers in 3′ UTR CDCs are almost exclusively composed of CU nucleotides (P < 10\textsuperscript{-6}), revealing alternative sequence elements that may mediate miRNA–ALG-1 target recognition and regulation in \textit{C. elegans} (Fig. 3c). Moreover, this striking pattern was not associated with CDCs from 5′ UTR, coding exon or intron regions (Supplementary Fig. 5c) or with clusters from \textit{alg-1(−)} animals, indicating that the CU-rich motifs are a specific characteristic of ALG-1–bound regions in 3′ UTRs. Multiple computational prediction methods and extensive reporter validation assays point to the miRNA ‘seed’ (defined as perfect pairing between miRNA bases 2–7 and the target site) as a primary determinant of target regulation. Given that the overwhelming majority of clusters reside in the 3′ UTR and coding exons, we sought to investigate whether the location of clusters affects mRNA levels. To test whether genes bound by ALG-1 at the 3′ UTR and coding exons were subject to regulation at the mRNA level, we performed microarray experiments comparing WT to \textit{alg-1(−)} L4-stage worms. Consistent with previous reports that miRNA regulation can result in substantial target-mRNA degradation in \textit{C. elegans}\textsuperscript{53,54}, \textit{lin-41}, \textit{lin-14}, \textit{lin-28} and many other established miRNA targets were upregulated in the \textit{alg-1(−)} mutant worms (Table 1 and Supplementary Table 2). Notably, genes containing 3′ UTR clusters were strongly upregulated in \textit{alg-1(−)} mutants compared to genes that had no ALG-1–bound sites (Fig. 4a). In contrast, no relationship was detected between mRNA expression levels and genes with clusters in coding exons (Fig. 4a). These findings suggest that the mechanism of target regulation may be different for genes with ALG-1–binding sites in 3′ UTRs versus coding exons.

We next asked if genes with ALG-1–binding sites or expression changes in \textit{alg-1(−)} compared to WT worms were enriched (P < 0.05) in particular functional classes based on the “Topomap” categories,...
which group co-regulated genes from extensive microarray datasets\textsuperscript{52}. Notably, several functional categories were distinctly associated with genes that contained CDCs in 3' UTRs versus coding exons and were up- or downregulated in \textit{alg-1(–)} mutants (Fig. 4b). For example, genes belonging to the functional classes "Protein kinases" and "Cell biology" are enriched for containing 3' UTR CDCs and being upregulated in \textit{alg-1(–)} worms. Genes in the "Histone" category are also associated with upregulation but tend to have CDCs in their coding exons. This difference in location of miRNA binding may be related to the typically short and nonpolyadenylated status of histone mRNAs\textsuperscript{35}. Some functional categories included genes with CDCs in the 3' UTR and coding exons and/or up- and downregulated genes. The overlap in categories is not surprising given the large fraction of genes with ALG-1–bound regions and the likely widespread direct and indirect effects on mRNA expression by the miRNA pathway. Our results reveal biological pathways targeted \textit{in vivo} by ALG-1 in developing worms and indicate that some gene categories tend to be differentially bound and regulated by ALG-1.

miRNA pathway genes are enriched in ALG-1 targets

During our analyses of categories of genes bound by ALG-1, we discovered a strong enrichment for genes implicated in the miRNA pathway. CDCs in the 3' UTR in the \textit{alg-1} gene indicate autoregulation of this core miRNA factor (Fig. 4c). Additionally, significant clusters were identified in the 3' UTRs of \textit{ain-1} and \textit{ain-2}, and miRNA levels of these genes and of the \textit{alg-1} homolog \textit{alg}-2 were found to be upregulated in \textit{alg-1(–)} worms (Supplementary Table 3). The potential cross-regulation of these miRNA effector genes may explain the nonlethal phenotype associated with loss of any single one of these genes\textsuperscript{6,15}. To investigate the extent of ALG-1 regulation of miRNA pathway genes, we analyzed two published lists of genes specifically connected to miRNA function by proteomic and genetic evidence\textsuperscript{6,54}. We observed that this network of miRNA pathway genes showed statistically significant enrichment in ALG-1 CDCs (30 out of 39 genes with CDCs and flanking sequences (100 nt upstream or downstream) of CDCs, compared to RDCs, are shown along with the range of Z-scores for the specific categories. (d–g). The number of conserved hexamers within CDCs (solid line) and RDCs (dashed line) that base-pair to miRNA or scrambled miRNA regions (dotted line), allowing for zero (orange) and only one G•U base pair (black). Error bars in dashed and dotted lines represent the s.d. among ten independent sets of RDCs and scrambled miRNAs, respectively. Hexamers within 3' UTR CDCs and RDCs (d) or coding-exon CDCs and RDCs (e) that base-pair to cloned miRNAs or shuffled versions of cloned miRNAs. Hexamers within 3'-UTR CDCs and RDCs that base-pair to the \textit{let-7} or shuffled \textit{let-7} miRNA (f) and \textit{lin-4} or shuffled \textit{lin-4} miRNA (g). Regions of the miRNA(s) that have statistically enriched numbers of complementary hexamers within CDCs when compared to RDCs or shuffled miRNAs are denoted by * (P < 0.01) and ** (P < 10\textsuperscript{-5}) as measured by a Z-test.

ALG-1–bound regions as a resource for miRNA target predictions

A number of different algorithms (mirWIP, rna22, PicTar, TargetScan, PITA and miRanda) are available for predicting miRNA target sites in \textit{C. elegans} genes\textsuperscript{22,35,44,45,53,55}. Most of these prediction methods use a common set of criteria (seed, conservation and energy requirements), except for PITA, which does not require conservation, and rna22, which uses a different set of parameters. Because predictions are typically available for 3'-UTR sequences, we assessed the ability of these methods to detect predicted miRNA target sites within the ALG-1–bound 3'-UTR CDCs (Supplementary Fig. 8, tracks for the predicted sites from these algorithms are available under 'ALG1 CLIP-seq' within the 'Regulation' section in the ce6 genome). Although 93% of the 3'-UTR CDCs contained a miRNA target site predicted by at least one of the algorithms (1,539 CDCs), only 3% of the CDCs had at least one site predicted by all 6 programs (52 CDCs). As an example, five of the six target prediction programs list potential miRNA target sites, largely disparate in both location and number, in the...
Figure 4 Relationship between ALG-1 binding and mRNA expression levels. (a) Effects of ALG-1 binding on mRNA levels. Box plots representing the differential expression (as a t-statistic) of genes from biological replicate microarray experiments comparing alg-1(–) to WT L4-stage worms. Genes are divided into those that contained no CDCs and those that contained CDCs only within 3′ UTRs or coding exons. Compared to genes with no CDCs or coding-exon CDCs, genes with 3′-UTR CDCs are significantly more upregulated in alg-1(–) relative to WT as assayed by the Wilcoxon rank-sum test (P < 10⁻⁴). (b) Functional enrichment of genes that have CDCs only within 3′ UTR or coding exons that are up- or downregulated in alg-1(–) worms using significantly enriched (P < 0.05 in at least one row; Holm-Bonferroni corrected) functional categories defined by the C. elegans Topomap algorithm. The intensity on the heat-map denotes -log₁₀(P-value). Genes represented by these functional categories can be divided in a matrix (right) depending on the location of the CDCs (3′ UTRs or coding exons), and whether the genes are up- or downregulated in the alg-1(–) mutants relative to WT worms. Several categories occupy multiple cells in the matrix, for example “Cell structure,” “Collagen,” “Cell adhesion,” “Protein expression,” “RNA binding” and “Germ line–enriched.” (c) UCSC Genome Browser view depicting clusters in the 3′ UTR of the alg-1 gene (blue, WT clusters; red, alg-1(–) clusters, none present) and the predicted miRNA binding sites by the various algorithms.

Discussion

We present a global snapshot of an endogenous miRISC RNA binding profile in whole animals. We demonstrate that binding of the core miRNA effector protein Argonaute is strongly enriched at the 3′ ends of transcripts, although substantial numbers of CDCs also reside within the 5′ UTR, coding exonic and intronic regions of genes as well. A striking signature of the ALG-1–bound 3′-UTR CDCs emerged: the regions showed greater sequence conservation and accessibility, they contained and were flanked by CU-rich motifs, they were enriched for sequences complementary to the 5′-end seed regions of miRNAs and they were associated with upregulation of miRNA expression in the alg-1(–) mutant background. Although some of these characteristics were shared with clusters in other genomic regions, the marked overall differences in 3′ UTR versus other regions suggests that separate rules may regulate ALG-1 binding to distinct positions within an mRNA. The importance of context could underlie the conflicting conclusions that have been drawn about the ability of miRNAs to target different regions in mRNAs and, in some cases, the failure of reporter assays to demonstrate miRNA regulation of genes bound and regulated by alg-1 (see Supplementary Table 2). In addition to providing a map of ALG-1 interaction sites for the C. elegans protein-coding genes potentially under miRNA regulation in late larval development (see Supplementary Figs 9 and Supplementary Data), compared to previously available methods our strategy substantially reduced the search space by factors of 5, 20 and 47 for identifying direct miRNA target sites. Although we detected a strong signal for pairing to the miRNA seed region in 3′-UTR ALG–1–bound sites, ~40% of the ALG-1 clusters lacked conserved seed pairing capacity, indicating that more flexible base-pairing rules may guide a large fraction of miRNA target recognition in vivo. Furthermore, the observation of different patterns for let-7 or lin-4 miRNA paired to sites within ALG–1–bound sequences raises the possibility of individual miRNA pairing rules. The discovery of miRNA pathway genes as an exceptional class of genes bound and regulated by endogenous alg-1 suggests that cross-regulation of miRNA cofactors contributes substantially to this essential posttranscriptional control mechanism. In conclusion, our analyses and data provide a framework and a rich resource for understanding in vivo miRNA–mRNA interactions in a context-specific manner.

Methods

Methods and any associated references are available in the online version of the paper at http://www.nature.com/nsmb/.

Accession codes. Microarray CEL files have been deposited at the Gene Expression Omnibus database repository under accession number GSE19138.

Note: Supplementary information is available on the Nature Structural & Molecular Biology website.

Acknowledgments

The authors thank G. Ruvkun, X. Fu, W. McGinnis and members of our laboratories for critical reading of the manuscript. We thank B. Hehler, S. Hunter and S. Bagga for technical assistance, V. Ambros and M. Hammell for providing the list of mirWIP predictions and D. Bartel for helpful advice. MLW is supported by the Genetics Training Program at the University of California, San Diego and a graduate fellowship from Genentech. This work was supported by grants from the US National Institutes of Health (GM071654-01 to A.E.P. and HG004659 and GM084317 to G.W.Y.), the Keck, Searle, V. Emerald and Peter Gruber Foundations (A.E.P.) and the Stem Cell Program at the University of California, San Diego (G.W.Y.).


The temporal patterning of microRNA let-7 regulates several transcription factors at the larval to adult transition in C. elegans (2007).


M.L.W., K.R.H, T.Y.L. and G.W.Y. performed the bioinformatics analyses.

M.T.L. wrote the paper; D.G.Z. and T.Y.L. performed the experiments; M.T.L., M.L.W., K.R.H, T.Y.L. and G.W.Y. performed the bioinformatics analyses.


ONLINE METHODS
Cross-linking and immunoprecipitation coupled to sequencing. For three independent experiments using WT and alg-1(−) worms, we harvested approximately 40,000 L4-stage worms and irradiated them with UV-B (3 kJ m⁻²). We lysed the irradiated worms by sonication. For the immunoprecipitation, we used a custom antibody and performed the rest of the CLIP-seq method essentially as described. For further details see Supplementary Methods.

Microarray analysis. We prepared three independent sets of RNA from WT and alg-1(−) worms at the L4 stage, labeled them as per manufacturer's instructions (Affymetrix) and hybridized them to Affymetrix C. elegans gene microarrays. To assign a value for differential gene expression between the two groups, we computed a t-statistic as described before. For further details see Supplementary Methods.

Improved UTR annotation. We obtained publicly available stage-specific C. elegans RNA-sequence libraries from published work, aligned them to the genome of C. elegans and then assigned them to the composite gene loci. For further details see Supplementary Methods. To control for RNA transcripts from noncoding and unannotated transcribed regions, we applied a 3-kb cutoff such that only reads within 3 kb of an annotated gene end could be assigned to the gene.

Defining experimentally reproducible regions. We aligned reads from ALG-1 CLIP-seq to the repeat-masked C. elegans genome (ce6) and extended 50 bases in the 3’ direction to account for the size of the gel-extracted PCR product. Reads that overlapped within and across experiments formed contiguous ‘regions’. We assigned a score to each nucleotide within a region, and we gave more weight to the region from the experiment with the most reads based on the assumption that there was a higher likelihood of detecting real interactions, even for weakly abundant RNAs. After assigning weights to each nucleotide, we considered a region to be replicated across biological experiments only if at least one of the nucleotides had a score greater than a user-defined cutoff. For further details see Supplementary Methods.

Computational identification of ALG-1 binding sites. For finding peaks, we considered only the regions that had at least one nucleotide satisfying the user-defined cutoff within WT or alg-1(−) samples. We calculated significant peaks by first determining read-number cutoffs using the Poisson distribution. The Poisson distribution assumes all intervals are independent and have equal probability of an occurrence happening. We determined a global and local cutoff by assigning the cutoff value using the whole transcriptome frequency as the global cutoff and using a gene-specific frequency for the local cutoff. The gene-specific frequency was simply the number of reads overlapping a gene divided by the pre-mRNA length of that gene. After finding these cutoffs, we used a sliding window the size of the interval to determine where the actual read numbers exceed both the global and local cutoffs. At each significant interval, we attempted to extend the region by adding in the next read and recalculating the significance of this new interval. If the probability was still significant, and the distance between this extension and the previous interval was sufficiently small, this read was included and the peak width was updated. This extension was empirically limited to two times the size of the minimal interval. We identified clusters independently for CLIP-seq performed on WT and alg-1(−) strains. Next, we considered WT clusters that overlapped with alg-1(−) by 25% either as abundant unbound RNA cloned independently of ALG-1 interaction or as PCR artifacts and removed these clusters. We considered WT clusters that did not overlap with alg-1(−) clusters for more than 25% of their length as bona fide ALG-1–interacting loci and termed them CLIP-derived clusters (CDCs). For further details see Supplementary Methods.

Generation of randomly derived clusters (RDCs). We divided genes into quartiles based on the number of reads aligned uniquely to each gene locus. We compared binding sites within genes to ‘background regions’ in genes expressed within the same quartile. We further divided the transcriptome into functional regions: 3’ UTR, coding exon, intron and 3’ UTR. For each region, we determined the average evolutionary conservation level by the algorithm PhastCons and GC content. We then divided each region into quartiles based on their conservation level and GC content. For each bona fide binding site of length $I$ that is contained within an unambiguously assigned functional region (see above) of conservation level $C$ and GC content $G$, we picked a background binding site at random of length $I$ from the transcriptome that fell in the same functional region of conservation level $C’$ and GC content $G’$, where $C’$ and $C$ as well as $G’$ and $G$ are in the same conservation and GC quartile, respectively. We implemented controls for $C’$ and $G’$ on the level of the whole genomic region. For the determination of motifs and conservation levels for CDCs vs RDCs, we did not control GC-content or conservation levels, respectively. For further details see Supplementary Methods.

Assessing miRNA-target base-pairing rules. We used all the miRNAs that were sequenced in the WT library in order to analyze base-pairing to target regions (CDCs) as compared to background regions (RDCs). We assessed the number of CDC sites complementary to every adjacent position of each miRNA for two definitions of conservation: (i) ‘exact’ conservation and (ii) ‘semiconservation’ and for two different definitions of ‘binding capacity’: (i) Watson-Crick base-pairing and (ii) G•U mismatches, allowing a single G•U base pair in a 6-mer site (see inset in Supplementary Fig. 7).

Conservation levels. We determined conservation by parsing with perl scripts the multialignments of the C. elegans (May 2008, ce6) to the Caenorhabditis brevetti (Feb 2008, caePb1) genomes. For further details see Supplementary Methods.

Statistical analysis. We used a number of statistical tests in this study and identified them at the appropriate places throughout the text and figure legends. For further details see Supplementary Methods.