Deep Conservation of microRNA-target Relationships and 3′UTR Motifs in Vertebrates, Flies, and Nematodes

K. CHEN* AND N. RAJEWSKY*†

**Center for Comparative Functional Genomics, Department of Biology, New York University, New York, New York 10003; †Max Delbrück Centrum for Molecular Medicine, Berlin-Buch, 13092 Berlin, Germany*

microRNAs (miRNAs) are a class of small noncoding RNAs that posttranscriptionally regulate a large fraction of genes in animal genomes. We have previously published computational miRNA target predictions in five vertebrates, six flies, and three nematodes. Here, we report a comprehensive study of the "deep" conservation of miRNA targets and conserved 3′UTR (untranslated region) motifs in general across vertebrates, flies, and nematodes. Our data indicate that although many miRNA genes and 3′UTR motifs are well-conserved, miRNA-target relationships have diverged more rapidly, and we explicitly assign each gained or lost miRNA-target relationship to one of the three clades. However, we also identify a small but significant number of deeply conserved miRNA targets and show that these are enriched for essential processes related to development. Finally, we provide lists of 3′UTR motifs that are significantly conserved, and thus likely functional, classified by their distribution in the three clades. We find hundreds of such motifs specific to each clade, dozens specific to each pair of clades, and ten shared by vertebrates, flies, and nematodes. These findings suggest that posttranscriptional control has undergone extensive rewiring during metazoan evolution and that many deeply conserved miRNA-target relationships may be vital subunits of metazoan gene regulatory networks.

The main interest of our lab is to identify and characterize gene regulatory elements and to ultimately arrive at a better understanding of gene regulatory networks. A recent focus of our lab has been gene regulation mediated by miRNAs. miRNAs are a class of small noncoding RNAs that posttranscriptionally regulate a large fraction of genes in animal genomes. To understand the function of miRNAs, it is necessary to identify and characterize their targets. We have previously published computational miRNA target predictions in five vertebrates, six flies, and three nematodes. For example, we and other groups have shown that at least 30% of all human genes are likely to be regulated by about 60 conserved vertebrate miRNAs. Moreover, in collaboration with experimental groups, we helped to determine the biological function of a few miRNAs (for a general review, see Rajewsky 2006). However, since many miRNAs are seemingly specific to certain metazoan clades, whereas others are conserved in virtually all animals, miRNAs and their targets are also an excellent system in which to study the evolution of a whole layer of gene regulation. For example, it seems possible to predict miRNA targets that are specific to a certain lineage within a metazoan clade (e.g., the *Sophophora* and *Drosophila* lineages within flies). Here, we focus on "deep" conservation of miRNA-mediated gene regulation, i.e., targets of conserved miRNAs that are shared by vertebrates, flies, and nematodes.

Although it is well known that many miRNAs are well-conserved across large evolutionary distances (Lagos-Quintana et al. 2001), the conservation of miRNA targets has only been studied in a few anecdotal cases. Notably, Pasquinelli et al. (2000) showed that *let-7* and one of its targets, *lin-41*, are broadly conserved in animals, Moss and Tang (2003) showed that *lin-28* is a conserved target of *let-7* and *lin-4* in mammals and nematodes, whereas Floyd and Bowman (2004) and Axtell and Bartel (2005) showed that a number of miRNA-target relationships are conserved in plants.

Here, we undertake a systematic study of miRNA-target conservation among vertebrates, flies, and nematodes, using computational target predictions from our previously published PicTar algorithm (Grun et al. 2005; Krek et al. 2005; Lall et al. 2006). Recently, an independent, large-scale experimental study estimated that approximately 90% of PicTar predictions in *Drosophila* are correct at a sensitivity of 70% (Stark et al. 2005). We discovered 5 miRNA-target relationships conserved in all three clades and 264 more miRNA-target relationships conserved in two clades (for these numbers, families of paralogous miRNAs and target genes are collapsed to a single representative).

We refer to a gene that is predicted to be a target of the same miRNA in at least two clades as a "deeply conserved target." The set of such targets is significantly enriched for genes involved in essential biological processes related to development (P value <5.8e-3 in humans and flies) (see Materials and Methods). Among the most interesting cases, we recovered five different subunits of vacuolar ATPase as deeply conserved targets of the *miR-1* family, *lin-28* as a deeply conserved target of the *let-7* family, and *odd-skipped* as a conserved target of the *miR-8* family.

Despite the suggestive biological significance of the deeply conserved miRNA targets, since PicTar predicts a very large number of miRNA targets in each of the three genomes (9379 in humans, 3082 in *D. melanogaster*, and 2679 in *Caenorhabditis elegans*), our results imply that miRNA targets are poorly conserved overall. In principle, this result could be due to three scenarios. First, the target predictions could simply be erroneous. Second, the

ancestral miRNAs could have had only a few targets and these targets have indeed been conserved, but many more targets have been gained subsequent to speciation, leading to an apparent lack of conservation of the ancestral targets. Third, the ancestral miRNA could have had many targets, and the network of miRNA-target relationships has undergone extensive rewiring during metazoan evolution. Our analysis suggests that the third scenario is most likely to be correct, and we speculate that rewiring of miRNA networks, like rewiring of transcription factor networks, may be important in bilaterian evolution and diversification (Davidson 2001). We stress the importance of analyzing miRNA conservation in at least three clades for distinguishing between the latter two evolutionary scenarios, since pairwise comparisons cannot discriminate between a gain and loss of an miRNA target.

We complement this analysis with a study of conserved 3′UTR motifs in vertebrates, flies, and nematodes, using techniques derived from Xie et al. (2005). The results of Xie et al. imply that conserved 3′UTR motifs in vertebrates are highly enriched for miRNA-binding sites, and we show that these results extend to flies and nematodes. In addition, we discovered a significant correlation between the patterns of motif conservation in these three clades, which implies that *cis*-regulatory motifs in 3′UTRs have remained well-conserved across very large evolutionary distances. We find that this correlation is strongest for human and flies and weakest for human and nematodes. We classify the hundreds of significantly conserved 3′UTR motifs that we have discovered according to their distribution in the three clades and hypothesize that many of these are likely to be functionally important *cis*-regulatory sequences.

In addition, our data are consistent with the interpretation that many of the most highly conserved miRNAs have already been discovered in these three clades, meaning that many of the remaining ones to be found are clade-specific. Of the three clades, the miRNA gene complement of *D. melanogaster* may be the most poorly sampled of the three.

Subsequent to obtaining these results, a similar computational study of miRNA target conservation in flies and nematodes appeared (Chan et al. 2005). Our work differs in several respects. First, we study three clades instead of two. This is interesting not only because we include many more species in our analysis and thus discover about three times as many potential deeply conserved regulatory relationships, but also because we can use the third clade as an outgroup in our evolutionary studies. In particular, we are able to differentiate between evolutionary scenarios two and three described above. Second, our target predictions rely on previously experimentally validated methods (Grun et al. 2005; Krek et al. 2005; Lall et al. 2006), whereas those of Chan et al. (2005) have not yet been subjected to experimental verification. For example, our fly 3′UTRs are defined based on full-length cDNAs and not artificially truncated to 500 nucleotides. As part of this work, we also attempted to predict binding sites in genes conserved across all three clades without requiring that they be aligned, exactly as in Chan et al. (2005). However, we failed to find any excess of predicted targets in real miRNAs versus randomized controls. This is in stark contrast to the strong signal-to-noise ratios observed in alignment-based target prediction methods (Lewis et al. 2003; Brennecke et al. 2005b; Grun et al. 2005; Krek et al. 2005; Lall et al. 2006). Third, our motif conservation score is defined differently and relies on methods previously used successfully for motif discovery in vertebrates (Xie et al. 2005).

MATERIALS AND METHODS

Construction of miRNA families. We clustered all miRNAs with PicTar target predictions by linking two miRNAs if they shared a nucleus (the first or second 7-mer) and applying single-linkage clustering. PicTar predictions are only generated for miRNAs conserved in all species under consideration. (For details, see Grun et al. 2005; Krek et al. 2005.)

Motif analysis. For the vertebrate alignment, we used repeat-masked UCSC alignments of the human, mouse, rat, and dog genomes downloaded from the UCSC genome browser (http://genome.ucsc.edu) as described previously by Krek et al. (2005). For the fly alignment, we used tandem-repeat-masked Mercator/MAVID alignments (Bray and Pachter 2003; http://hanuman.math.berkeley.edu/~cdewey/mercator) of *D. melanogaster*, *D. erectus*, *D. ananassae*, *D. yakuba*, and *D. pseudoobscura*, and for the nematode alignment, we used unmasked Mercator/MAVID alignments of *C. elegans*, *C. briggsae*, and *C. remanei* as described previously by Grun et al. (2005) and Lall et al. (2006). For all genes with multiple transcript variants, we kept only the transcript with the longest 3′UTR.

We experimented with different conservation scores, using as our metric the average score of all known miRNAs. We attempted to improve sensitivity by excluding very short UTRs (<200 nucleotides for vertebrates and <100 nucleotides for flies and worms), dividing by the average total count in all species, instead of the total count in just the reference species, and considering 5-, 6-, and 8-mers in the analysis, but we found that none of these performed as well as the simplest measure. When matching the motifs against miRNA sequences, we matched against the entire Rfam7.0 data set (Griffiths-Jones et al. 2003), not just the miRNAs that had PicTar predictions.

For the purposes of viewing the scatter plots (Fig. 1), we added the chicken genome to the vertebrate alignments, and for the fly alignments, we used repeat-masked UCSC alignments of *D. melanogaster*, *D. ananassae*, *D. yakuba*, *D. pseudoobscura*, *D. mojavensis*, and *D. virilis*. Visually, the additional species tend to accentuate the "arms" of the scatter plot, but otherwise, the overall patterns of conservation are not affected. In addition, to minimize noise, we eliminated all motifs that have a conserved count of zero. Intuitively, small counts induce high variance in the Z scores and tend to blur the "arms" of the scatter plots. Finally, we took the absolute value of the Z scores instead of the raw Z scores because we were concerned only with overrepresentation of conserved occurrences.

Conservation of targets. For each miRNA family, we considered the overlap between each pair of clades separately. For each clade, we took all the miRNAs in the family from that clade, as well as the union of their target sets, removing all but one copy of each set of paralogs to get a unique set of target genes for the clade. We then computed the overlap rate as the intersection of the two unique sets divided by the cardinality of the smaller of the two sets. This approach was designed to maximize sensitivity for conserved targets and does not penalize for target genes with no ortholog in the other clade.

The null model that we considered is that of sampling uniformly at random and independently without replacement from the set of all genes with orthologs in the other clade, keeping the number of unique genes (i.e., no paralogs) in the two sets constant. In this model, we assumed that 3′UTRs of homologous genes in different clades were sufficiently diverged as to be statistically independent. To compute the expected number of targets in the null model, we assumed that there were N clusters of homologs between two clades and the target sets were of size $j \leq k$. Fixing the number of targets in one clade to be j, by linearity of expectation, the number of overlaps is $k * j / N$ (recall that independence is not required for the linearity of expectation), which implies that the overlap rate is k / N. For three-way comparisons with target sets of size i, j, k, and N homologs between the three clades, i * j * k / N^2 so the rate is j*k/N^2. Another way to derive this is by taking the expectation of an appropriate hypergeometric distribution.

To compute the overall average rate, we took the unweighted average of the per-family overlap rate. An alternative is to weight the rates by the size of the smaller of the two reduced sets, thus reducing the effect of noise from small sets. When we compute the overall average rates this way, the numbers are almost exactly the same.

To identify families with significantly high conservation of targets, we computed P values for Fisher's exact test using code from the GeneMerge program (Castillo-Davis and Hartl 2003) and the multiple testing correction using the Multtest package for the R programming language from the Bioconductor project. To identify additional putative triply conserved target relationships that were missed by PicTar, we took all doubly conserved target relationships for which an orthologous miRNA and orthologous target gene existed in the third clade. We did not collapse paralogous target genes since not all of these are expected to contain the binding site. There were 265 genes that fit this category, many of which were targeted by multiple 6-mers; 73 genes did not appear in our alignments at all, leaving a total of 241 gene–6-mer pairs to check. For each of these 241 pairs, we checked each 3′UTR from the relevant species (human, mouse, rat, dog, *D. melanogaster, D. ananassae, D. pseudoobscura, D. yakuba, C. elegans, C. briggsae, C. remanei*) for the presence of the 6-mer without requiring it to be aligned, and we considered a gene to be a putative target gene if at least three of the species contained the 6-mer, one of which had to be the reference species. To compute the expected number of hits, we concatenated the lengths of the 3′UTRs (without removing repeat-masked sequence) and divided by 4^6.

In our GO term analysis, we used as our background set the set of all PicTar targets that have orthologs in one of the other two clades, and which are themselves targeted by miRNAs from multiclade families. In the evolutionary analysis, we assumed a star phylogeny and used the simple parsimony criteria to assign each target relationship conserved in either one or two clades as a gain or loss, respectively.

RESULTS

Conservation of miRNA Families

We took all human, *D. melanogaster*, and *C. elegans* miRNAs for which PicTar target predictions are available and clustered them into families of homologous miRNAs (see Materials and Methods). Our clustering method relies on the notion of a critical region for target recognition, the "nucleus," defined here to be the first or second 7-mer in the mature miRNA sequence as described by Krek et al. (2005). Recent research has established that the nucleus is the most important element of miRNA-target-binding specificity (Lewis et al. 2003; Doench and Sharp 2004; Brennecke et al. 2005b). PicTar uses this model to make most of its predictions, but it also uses the binding free energy to predict imperfect (3′ compensatory) sites. Careful examination of the families indicated that nearly all of them have very good alignments across the entire length of the mature sequence and that they are generally concordant with the human–nematode families defined by Lim et al. (2003).

The large number of families conserved in all three clades (15) is particularly striking since it is larger than the number of families conserved in any pair of clades (5 human–fly, 5 human–nematode, and 9 fly–nematode), which suggests that there is significant conservation of miRNA genes across the clades. A closer examination of the families themselves revealed that many are very well conserved at the sequence level beyond matching at the nucleus. For example, four of the five families specific to humans and flies (*miR-219, miR-7, miR-210, miR-184*) are perfectly conserved between these two clades, and in the case of *miR-7*, the genomic location of the gene inside the last intron of the *D. melanogaster* bancal gene is conserved.

As a negative control for our clustering method, we repeated our procedure with all miRNAs from humans, *D. melanogaster*, and *C. elegans*, as well as three plant species *Arabidopsis thaliana*, *Oryza sativa,* and *Zea mays*. We found that all animal and plant miRNAs clustered separately, with the exception of one cluster made up of a human miRNA and two *O. sativa* miRNAs, which is likely to be a coincidence. This suggests that the families as defined represent true evolutionary homologs.

CONSERVATION OF 3′UTR MOTIFS

To further investigate the evolution of miRNAs, we examined the conservation of 3′UTR motifs in these three clades, using techniques adapted from Xie et al. (2005).

We used the same 3′UTR alignments that were used for PicTar target predictions to identify 7-mers whose rate of conservation was significantly higher than that of random 7-mers. Formally, for each clade, we computed a conservation score for each 7-mer, defined as the number of conserved instances in the 3′UTR alignment divided by the total number of instances in the reference species (*Homo sapiens*, *D. melanogaster*, or *C. elegans*, respectively). We note that our analysis differs from that of Xie et al. (2005) in that they consider motifs of length 6–18 and perform clustering of similar motifs, whereas we consider only 7-mers (i.e., potential miRNA-binding sites).

For each clade, we compute a Z score for each motif based on the distribution of conservation scores, and we call a motif with a Z score >3 a highly conserved motif (HCM). In each clade, HCMs are highly enriched for miRNA-binding sites. In vertebrates, 64 of 206 HCMs correspond to binding sites; in flies, 50 of 206 HCMs correspond to binding sites; and in worms, 42 of 223 HCMs correspond to binding sites. This analysis extends the results of Xie et al. (2005) to flies and nematodes.

Recently, Bentwich et al. (2005) identified 53 candidate primate-specific miRNAs. We extracted these and computed the distribution of Z scores of their nuclei. Surprisingly, at first, the distribution of the 69 unique nuclei was not significantly different from that of all human miRNAs in a two-tailed Wilcoxon test, as would be expected if these nuclei were indeed primate-specific. However, upon closer investigation, we found that 10 of these nuclei matched to other (nonprimate-specific) human miRNAs or miRNAs in *D. melanogaster* or *C. elegans*. Excluding these k-mers from the analysis, we found that the average Z score of the remaining 59 primate-specific nuclei was 0.71 versus 1.34 for other human miRNAs and that this difference is statistically significant in a one-tailed Wilcoxon test (*P* value 0.013). This suggests that some of the primate-specific genes are paralogous to miRNAs conserved more broadly in mammals, or even in flies or nematodes.

When we considered HCMs that have a Z score >3 in more than one clade, we found that these HCMs are very highly enriched for conserved miRNAs. Of 36 HCMs conserved in both vertebrates and flies, 19 match to known miRNAs, 16 of which are conserved in both humans and *D. melanogaster*. Comparable numbers were seen in the other two-way clade comparisons (21 of 34 HCMs conserved in flies and worms matched miRNAs, of which 16 were conserved in both, and 13 of 20 HCMs conserved in humans and worms matched miRNAs, of which 10 were conserved in both). We present these results for the human–worm comparison as a scatter plot in Figure 1, along with a random control in which the motif-to-Z score assignments have been randomly permuted in all three clades (see Materials and Methods). The scatter plots for the other two pairwise comparisons and a list of all HCMs are available upon request from the authors.

The scatter plots indicate a high degree of correlation between the conservation of 3′UTR motifs in these three clades, as well as an extremely high enrichment for conserved miRNA-binding sites in the HCMs conserved in both clades. This can be visualized by comparison with randomized controls in which the Z scores for the k-mers have been randomly permuted. These control plots show very few outliers outside of the two "arms" corresponding to the clade-specific motifs, as might be expected from a scatter plot of two independent Poisson distributions, and virtually all miRNAs lie in the insignificant regime. A comparison of the human–nematode and human–fly scatter plots shows that the two "arms" are more prominent in the human–nematode plot, indicating that the motifs in these two clades are less correlated. Quantitatively, the Pearson correlation coefficients for humans versus flies, humans versus worms, and flies versus worms are 0.39, 0.12, and 0.29, respectively.

Extending this analysis to a comparison of all three clades, we found that of the ten HCMs conserved in all three clades, eight matched some miRNA and seven matched in all three clades. Two of the most well-known conserved miRNA families, *let-7* and *miR-1*, appear as HCMs conserved in all three clades. The two motifs that remain unaccounted for are GTAAATA and AGTGCCT.

GTAAATA matches to a predicted miRNA precursor sequence in *C. elegans* with an miRScanII score of 9.23 bits (Ohler et al. 2004). Only precursors of score >12.7 bits were experimentally screened in Ohler et al. (2004). It also matches the first seven bases of the 32nd highest scoring 8-mer in humans (Xie et al. 2005), which in turn matches one novel predicted miRNA precursor. The highest score of any predicted precursor sequence in *C. elegans* containing AGTGCCT is only 0.39 bits, and the motif is not found in any predicted mature miRNA sequence in humans from Xie et al. (2005).

When we examined the number of HCMs in each clade that match an miRNA from one of the other two clades, we found only 3 and 9 such HCMs for humans and *C. elegans*, respectively, but 19 for *D. melanogaster*. This suggests that many of the most highly conserved miRNAs have already been discovered in humans and *C. elegans*, whereas the *D. melanogaster* miRNA complement may be the most poorly sampled of the three. This inference is consistent with the small total number of *D. melanogaster* miRNA families as compared to worms and humans (Fig. 1), as well as the relatively low amount of effort expended on miRNA gene finding in *D. melanogaster* thus far. It is also consistent with the work of Bentwich et al. (2005), which suggests that many of the remaining human miRNAs are primate-specific. At the practical level, our results imply that miRNA gene-finding techniques relying on comparisons over moderate phylogenetic distances should continue to have success in *D. melanogaster*, whereas phylogenetic shadowing techniques using comparisons over much shorter phylogenetic distances (Berezikov et al. 2005) may be required in human and *C. elegans*.

Our motif analysis is consistent with the currently accepted view of miRNA-target-binding specificity, in that in all three clades, the second 7-mer was more conserved than the first 7-mer. In contrast, using a different measure of conservation, Chan et al. (2005) found that the first 7-mer is more conserved in flies but the second 7-mer is more conserved in worms. TargetScanS (Lewis et al. 2003), a different target prediction program for ver-

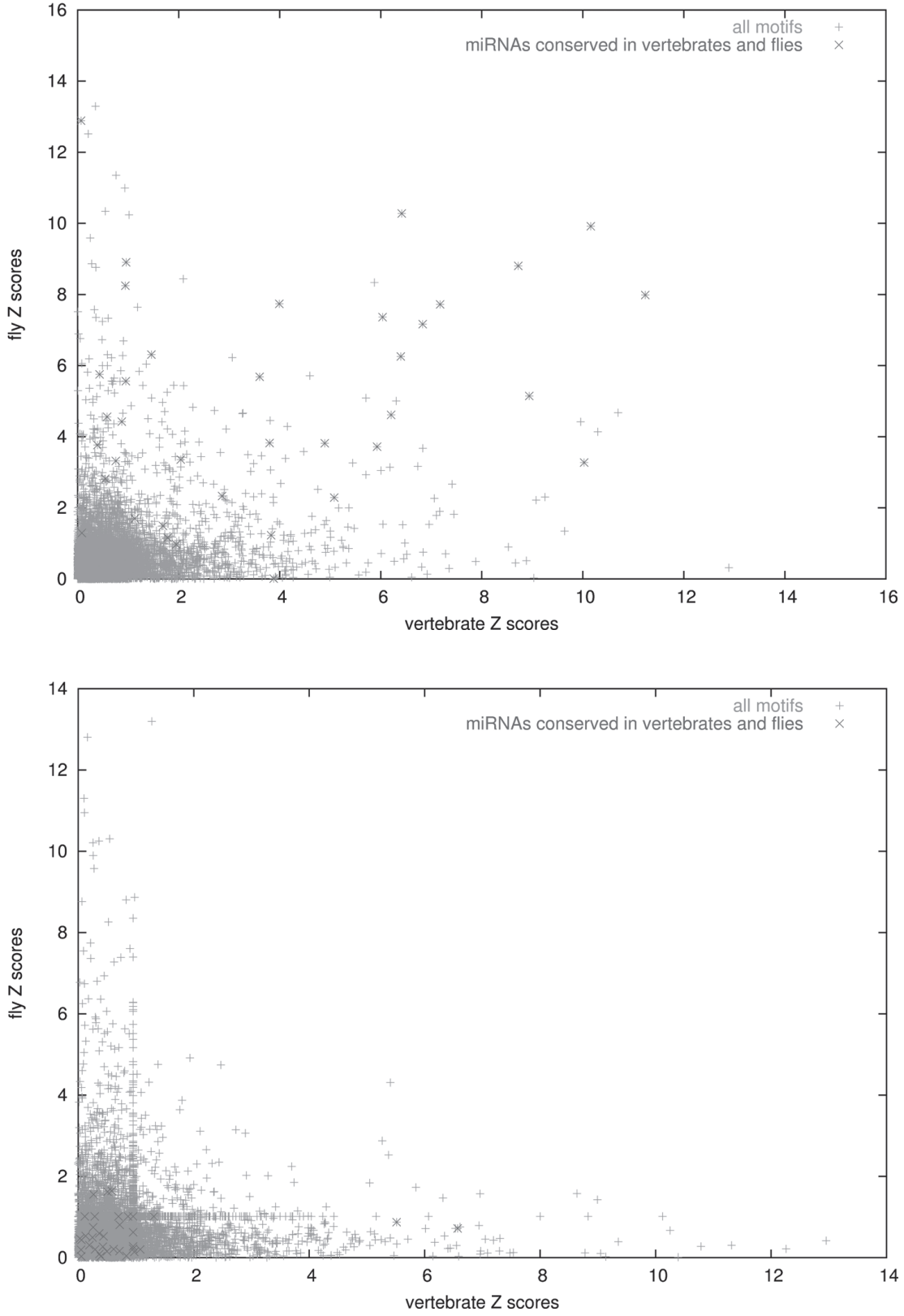

Figure 1. Scatter plot of Z scores of (*top*) 3′UTR motifs in vertebrates and flies and (*bottom*) 3′UTR motifs in vertebrates and flies with Z score-to-motif assignments randomly permuted. miRNAs found in both vertebrates and flies are marked with an X.

tebrates, uses a conserved adenosine in the first position regardless of the nucleotide at the first position in the miRNA. Although we find that an adenosine in the first position is indeed more conserved than the complementary nucleotide in humans and flies, this does not extend to nematodes. Although it is possible that the mechanism of miRNA binding is different in these clades, we have instead chosen not to use the conserved adenosine in any of our predictions.

Conservation of miRNA-target Relationships

Given the high degree of miRNA conservation observed so far, we turned next to the conservation of miRNA-target relationships. For this analysis, we used target predictions from the PicTar algorithm. Briefly, PicTar is a probabilistic algorithm that takes as input 3′UTR alignments from several related species and a set of miRNAs and computes a maximum likelihood parse of the 3′UTR sequence into binding sites and background sequence (Grun et al. 2005; Krek et al. 2005; Lall et al. 2006; http://pictar.bio.nyu.edu).

Our overall strategy was to maximize sensitivity for conserved miRNA-target relationships at the cost of specificity. To this end, we first chose the most sensitive PicTar settings available, which involved using nonrepeat-masked alignments of fewer species (human, mouse, rat, and dog for vertebrate predictions; *D. melanogaster*, *D. ananassae*, *D. pseudoobscura*, and *D. yakuba* for the fly predictions; and *C. elegans*, *C. briggsae*, and *C. remanei* for the nematode predictions). For the species used in this study, the estimated signal-to-noise ratios for the PicTar algorithm were 2.3 in humans, 2.5 in *D. melanogaster*, and 2.7 in *C. elegans*.

We downloaded pairwise sets of homologous genes for the human, *D. melanogaster*, and *C. elegans* genomes from the Inparanoid database (O'Brien et al. 2005) and augmented these with additional homologous genes from the Ensembl database (http;//www.ensembl.org). We clustered all of these together into three-way homology genes using single-linkage clustering. A comparison with a previously published analysis of human and *Drosophila* targets (Grun et al. 2005) gave a very good overlap, although the homologous gene sets used were different, indicating that the quality of homology detection, even across clades as distant as these, is not a major factor in the analysis.

For each family of miRNAs, we took the union of the target sets of each of the miRNAs in a clade as the representative set. As noted previously, most of the miRNA target sets are expected to be very similar within each family since the miRNAs typically share the same nucleus, so this procedure does not inflate the number of targets much. We computed the overlap for each pair or triple of clades, counting paralogs of miRNAs or target genes only once and computing the percentage overlap as the ratio of the number of overlapping genes to the size of the smaller of the two sets of genes. We only counted genes in the representative sets that had orthologs in the other clade for the purposes of this calculation, so that miRNAs were not penalized for having many clade-specific genes.

Despite all efforts to increase sensitivity, we reached the surprising conclusion that the percentage of overlap was very low in all cases. Averaging over approximately 20 families for each pair of clades, the percentage of overlapping targets is 10% for humans and *D. melanogaster*, 11% for human and *C. elegans*, 4% for *D. melanogaster* and *C. elegans*, and 0.7% for all three clades. In a random model where the target genes are sampled randomly from the set of all genes with orthologs, we expect to see the percentage of overlapping targets to be 4% for humans and *D. melanogaster*, 5% for humans and *C. elegans*, 2% for *D. melanogaster* and *C. elegans*, and 0.09% for all three clades. Thus, we conclude that the number of conserved targets is slightly higher than random, but still small. A list of all conserved miRNA-target relationships is available by request from the authors.

For each family and pair of clades, we tested the significance of the overlap by computing a *P* value for Fisher's exact test and correcting for multiple hypothesis testing with the Bonferroni correction. Four families had conservation statistically greater than that expected at random (*P* value <0.04). Using their *D. melanogaster* representatives, these are the *miR-263b*, *miR-8*, *miR-7*, and *miR-92* families (see Materials and Methods for the other members of the families). Among the families showing significant overlap, we single out the *miR-92* family, which contains by far the largest number of conserved targets of any family (32 conserved targets between vertebrates and flies). Many of the conserved targets are important developmental transcription factors, such as *engrailed, mef2, crooked legs, erect wing, held out wings, spalt major, grain, E2f2, antennapedia, longitudinals lacking,* and *jun-related antigen.* GO term analysis of both the human and *D. melanogaster* target sets shows significant enrichment for transcription factors (*P* value <0.0001; see Materials and Methods). We point out that this family has been suggested to be important in *Drosophila* development (Leaman et al. 2005) but also note that this result has recently been put to question (Brennecke et al. 2005a).

An analysis of the conserved targets shows that they are significantly enriched for genes involved in development in flies, as compared to the set of all target genes with orthologs and which are targeted by orthologous miRNAs. These include morphogenesis (*P* value 2.4e-3), organ development (*P* value 2.0e-3), development (*P* value 1.8e-4), and histogenesis (*P* value 7.2e-3). A similar result for flies was seen previously by Grun et al. (2005). In humans, conserved targets are enriched for development (*P* value 5.8e-3) and amino acid phosphorylation (*P* value 4.3e-3). A few specific conserved miRNA-target relationships are of particular interest. The muscle-specific *miR-1* targets a subunit of vacuolar ATPase in all three clades, and four other subunits of vacuolar ATPase are conserved targets in two clades. *let-7* targets *lin-28* in both humans and worms, whereas *odd-skipped* is a conserved target of the *miR-8* family in all three clades.

We investigated the possibility that binding sites are present in 3′UTRs but not detected by PicTar, since the binding sites may have undergone rearrangements and

therefore are not aligned correctly (Krutzfeld et al. 2005). We identified all miRNA-target relationships conserved in exactly two clades for which a homologous miRNA and a homologous target gene existed in the third clade. In each of these cases, we searched all the 3′UTRs of the homologous target genes in the third clade for the presence of a 6-mer-binding site (positions 2–7 from the 5′ end), without requiring that the binding site be conserved in the alignment. These requirements are significantly less strict than those implemented in the PicTar algorithm, which requires aligned 7-mer-binding sites, and are closer to the target prediction requirements of Chan et al. (2005).

After excluding those genes that do not appear in our alignments, we were left with 241 pairs of genes and k-mers to test (for this analysis, we kept all paralogous target genes, since in many cases, only a subset of paralogous genes contain a binding site). Among these, we discovered only 14 other potential triply conserved target relationships. In only two cases was a 7-mer nucleus present in all species but not conserved in the alignment. The small number of such cases suggests (1) that targets are indeed lost, and not simply missed by the algorithm, and (2) that our target prediction methods seem to be robust to rearrangements in the positions of the binding sites.

Interestingly, three of the five new *D. melanogaster* potential targets are targets of *miR-4/miR-79* and the other two are targets of *miR-8*, whereas all four of the new *C. elegans* potential targets are targets of *miR-235*. This suggests that the binding mechanism of some miRNAs may be different from that of other miRNAs (i.e., some miRNAs may require only a 6-mer nucleus or their target recognition sequence could depend on their nucleotide composition). This is potentially a valuable observation to incorporate into future target prediction algorithms.

One case of particular interest is the vacuolar-ATPase complex in which five subunits are conserved in at least two of the three clades. We failed to find binding sites in the third clade even with our relaxed requirements, or even when requiring only an unaligned 5-mer-binding site (positions 2–6). One possible reason for this is that down-regulation of a few subunits is sufficient to down-regulate the entire protein complex.

Taken together, these results demonstrate that many miRNA target sites are indeed lost over large evolutionary timescales. As previously discussed, a pairwise comparison of clades cannot distinguish between a loss or gain of an miRNA target, whereas the three-way comparison we perform here makes this distinction clear and implies that miRNA regulatory networks have undergone extensive rewiring. The results also imply that the low conservation of regulatory relationships is not due to overly stringent target prediction criteria.

CONCLUSIONS

We have performed a systematic study of the conservation of miRNAs, 3′UTR motifs, and miRNA targets in vertebrates, flies, and nematodes and have shown that although miRNAs and 3′UTR motifs are well-conserved, miRNA targets have diverged far more rapidly. We have also argued that miRNA targets are indeed lost and gained over evolutionary timescales (as opposed to being just gained), implying a certain amount of flexibility in the network of miRNA regulation.

In this section, we argue that the lack of target conservation we observe is unlikely to be due to erroneous target predictions, but instead is likely to reflect biological reality. First, to address the possibility that our miRNA families do not represent true evolutionary homologs, we repeated our analysis on only a subset of 13 human–*Drosophila* miRNA families that contain essentially perfect alignments across the entire mature sequence and for which homology is clear and obtained exactly the same result. Second, to address the problem of false positives in the PicTar predictions, we observe that since the signal-to-noise ratios are all approximately 2, false positives affect the overlap rate by at most an expected factor of 2, which is still a small overlap.

A third and more serious problem is that of false negatives. Indeed, we estimate that 25–30% of targets are either not conserved in all the species studied or are missed due to details of the algorithm (Stark et al. 2005; N. Rajewsky, unpubl.). However, we argue that these as yet undiscovered targets would not change the overlap rate in expectation, assuming their rate of deep conservation is similar to the current set of predicted targets. In fact, because the PicTar algorithm depends in large part on strict conservation over quite a broad range of species, it is much more probable that the current set of targets is heavily biased toward the most conserved genes, and hence our overlap rate is in fact an overestimate in this regard.

The idea that the miRNAs themselves are well-conserved, whereas their targets have changed rapidly, is plausible since miRNAs typically target hundreds of genes; thus, a small change in an miRNA would therefore affect many genes, whereas the change in a single target would have a much smaller effect. We note that it is relatively easy to lose an miRNA target site (typically, it would take just one point mutation), whereas by comparison, it is more difficult to destroy a transcription-factor-binding site because these sites tend to accommodate more degeneracy. Future quantitative comparison of the evolution of miRNA-target sites, as compared to transcription-factor-binding sites, could reveal whether post-transcriptional regulatory networks mediated by miRNAs evolve faster or slower than transcriptional regulatory networks. Similarly, it will be interesting to see how the evolution of these networks relates to other types of biological networks, such as protein–protein interaction networks (Sharan et al. 2005).

The evolution of gene regulatory networks over the very large evolutionary distances we consider here has rarely been studied at the global level, largely because of the difficulties in predicting binding sites for transcription factors and posttranscriptional regulators. Because of the comparatively simple nature of miRNA-target prediction, we are able to suggest a high-level view of the evolution of gene regulatory networks in metazoans. Our results indicate that the regulators themselves (the miRNAs) appear to be well-conserved, whereas regulatory relationships between miRNAs and target genes have changed

more rapidly. Nonetheless, a small core of developmentally important regulatory relationships appears to be conserved and thus may be a crucial component of the metazoan regulatory network. In addition, we identify clade-specific conserved 3′UTR motifs that may be functionally important for posttranscriptional regulation.

ACKNOWLEDGMENTS

We thank Uwe Ohler for providing us with the predicted *C. elegans* miRNA precursor sequences from the MiRscanII pipeline. This research was supported in part by the Howard Hughes Medical Institute grant through the Undergraduate Biological Sciences Education Program to New York University. In addition, K.C. thanks Lior Pachter for generous support (National Institutes of Health grant R01-HG02362-03).

REFERENCES

Axtell M.J. and Bartel D.P. 2005. Antiquity of microRNAs and their targets in land plants. *Plant Cell* **17:** 1658.

Bentwich I., Avniel A., Karov Y., Aharonov R., Gilad S., Barad O., Barzilai A., Einat P., Einav U., Meiri E., et al. 2005. Identification of hundreds of conserved and nonconserved human microRNAs. *Nat. Genet.* **37:** 766.

Berezikov E., Guryev V., van de Belt J., Wienholds E., Plasterk R.H., and Cuppen E. 2005. Phylogenetic shadowing and computational identification of human microRNA genes. *Cell* **120:** 21.

Bray N. and Pachter L. 2003. MAVID: Constrained ancestral alignment of multiple sequences. *Genome Res.* **14:** 693.

Brennecke J., Stark A., and Cohen S.M. 2005a. Not miR-ly muscular: microRNAs and muscle development. *Genes Dev.* **19:** 2261.

Brennecke J., Stark A., Russell R.B., and Cohen S.M. 2005b. Principles of microRNA-target recognition. *PloS Biol.* **3:** e85.

Castillo-Davis C.I. and Hartl D.L. 2003. GeneMerge—Post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics* **19:** 891.

Chan C.S., Elemento O., and Tavazoie S. 2005. Revealing post-transcriptional regulatory elements through network-level conservation. *PLoS Comput. Biol.* **1:** e69.

Davidson E.H. 2001. *Genomic regulatory systems: Development and evolution*. Academic Press, New York.

Doench J.G. and Sharp P.A. 2004. Specificity of microRNA target selection in translational repression. *Genes Dev.* **18:** 504.

Floyd S.K. and Bowman J.L. 2004. Gene regulation: Ancient microRNA target sequences in plants. *Nature* **428:** 485.

Griffiths-Jones S., Bateman A., Marshall M., Khanna A., and Eddy S.R. 2003. Rfam: An RNA family database. *Nucleic Acids Res.* **31:** 439.

Grun D., Wang Y., Langenberger D., Gunsalus K.C., and Rajewsky N. 2005. microRNA target predictions across seven *Drosophila* species and comparison to mammalian targets. *PLoS Comput. Biol.* **1:** e13.

Krek A., Grun D., Poy M.N., Wolf R., Rosenberg L., Epstein E.J., MacMenamin P., de Piedade I., Gunsalus K.C., Stoffel M., and Rajewsky N. 2005. Combinatorial microRNA target predictions. *Nat. Genet.* **37:** 495.

Krutzfeld J., Rajewsky N., Braich R., Rajeev J.G., Tuschl T., Manoharan M., and Stoffel M. 2005. Silencing of microRNAs in vivo with "antagomirs." *Nature* **438:** 685.

Lagos-Quintana M., Rauhut R., Lendeckel W., and Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* **294:** 853.

Lall S., Grun D., Krek A., Chen K., Wang Y.L., Dewey C.N., Sood P., Colombo T., Bray N., MacMenamin P., et al. 2006. A genome-wide map of conserved microRNA targets in *C. elegans*. *Curr. Biol.* **16:** 460.

Leaman D., Chen P.Y., Fak J., Yalcin A., Pearce M., Unnerstall U., Marks D.S., Sander C., Tuschl T., and Gaul U. 2005. Antisense-mediated depletion reveals essential and specific functions of microRNAs in *Drosophila* development. *Cell* **121:** 1097.

Lewis B.P., Shih I.H., Jones-Rhoades M.W., Bartel D.P., and Burge C.B. 2003. Prediction of mammalian microRNA targets. *Cell* **115:** 787.

Lim L.P., Lau N.C., Weinstein E.G., Abdelhakim A., Yekta S., Rhoades M.W., Burge C.B., and Bartel D.P. 2003. The microRNAs of *Caenorhabditis elegans*. *Genes Dev.* **17:** 977.

Moss E.G. and Tang L. 2003. Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. *Dev. Biol.* **258:** 432.

O'Brien K.P., Remm M., and Sonnhammer E.L. 2005. Inparanoid: A comprehensive database of eukaryotic orthologs. *Nucleic Acids Res.* **33:** D476.

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

Pasquinelli A.E., Reinhart B.J., Slack F., Martindale M.Q., Kuroda M.I., Maller B., Hayward D.C., Ball E.E., Degnan B., Muller P., et al. 2000. Conservation of the sequence and temporal expression of let-7 heterochronic regulatory RNA. *Nature* **408:** 86.

Rajewsky N. 2006. microRNA target predictions in animals. *Nat. Genet.* **38:** S8.

Sharan R., Suthram S., Kelley R., Kuhn T., McCuine S., Uetz P., Sittler T., Karp R., and Ideker T. 2005. Conserved patterns of protein interaction in multiple species. *Proc. Natl. Acad. Sci.* **102:** 1972.

Stark A., Brennecke J., Bushati N., Russell R.B., and Cohen S.M. 2005. Animal microRNAs confer robustness to gene expression and have a significant impact on 3′UTR evolution. *Cell* **123:** 1133.

Xie X., Lu J., Kulbokas E.J., Golub T., Mootha V., Lindblad-Toh K., Lander E.S., and Kellis M. 2005. Systematic discovery of regulatory motifs in human promoters and 3′UTRs by comparison of several mammals. *Nature* **434:** 338.