Identification of *Drosophila* MicroRNA Targets

Alexander Stark☯, Julius Brennecke☯, Robert B. Russell, Stephen M. Cohen*

European Molecular Biology Laboratory, Heidelberg, Germany

MicroRNAs (miRNAs) are short RNA molecules that regulate gene expression by binding to target messenger RNAs and by controlling protein production or causing RNA cleavage. To date, functions have been assigned to only a few of the hundreds of identified miRNAs, in part because of the difficulty in identifying their targets. The short length of miRNAs and the fact that their complementarity to target sequences is imperfect mean that target identification in animal genomes is not possible by standard sequence comparison methods. Here we screen conserved 3′ UTR sequences from the *Drosophila melanogaster* genome for potential miRNA targets. The screening procedure combines a sequence search with an evaluation of the predicted miRNA–target heteroduplex structures and energies. We show that this approach successfully identifies the five previously validated *let-7*, *lin-4*, and *bantam* targets from a large database and predict new targets for *Drosophila* miRNAs. Our target predictions reveal striking clusters of functionally related targets among the top predictions for specific miRNAs. These include *Notch* target genes for *miR-7*, proapoptotic genes for the *miR-2* family, and enzymes from a metabolic pathway for *miR-277*. We experimentally verified three predicted targets each for *miR-7* and the *miR-2* family, doubling the number of validated targets for animal miRNAs. Statistical analysis indicates that the best single predicted target sites are at the border of significance; thus, target predictions should be considered as tentative until experimentally validated. We identify features shared by all validated targets that can be used to evaluate target predictions for animal miRNAs. Our initial evaluation and experimental validation of target predictions suggest functions for two miRNAs. For others, the screen suggests plausible functions, such as a role for *miR-277* as a metabolic switch controlling amino acid catabolism. Cross-genome comparison proved essential, as it allows reduction of the sequence search space. Improvements in genome annotation and increased availability of cDNA sequences from other genomes will allow more sensitive screens. An increase in the number of confirmed targets is expected to reveal general structural features that can be used to improve their detection. While the screen is likely to miss some targets, our study shows that valid targets can be identified from sequence alone.

Introduction

MicroRNAs (miRNAs) are small RNAs, typically of approximately 21–23 nt, that direct posttranscriptional regulation of gene expression by binding to messenger RNAs (mRNAs). Many endogenously encoded miRNAs have been cloned from plants and animals (Lagos-Quintana et al. 2001, 2002; Lau et al. 2001; Lee and Ambros 2001; Mourelatos et al. 2002; Reinhart et al. 2002; Ambros et al. 2003; Aravin et al. 2003; Lim et al. 2003). Combining these data with computational cross-genome comparison predicts 100–120 miRNA-encoding genes in *Caenorhabditis* and *Drosophila* and approximately 250 in mouse and human (Ambros et al. 2003; Grad et al. 2003; Lai et al. 2003; Lim et al. 2003a, 2003b). However, functions have been assigned to only four animal miRNAs (Reinhart et al. 2000; Brennecke et al. 2003; Lee et al. 1993; Wightman et al. 1993; Xu et al. 2003), in part owing to the difficulty in identifying mutations in such small genes. A method to identify the target genes that are regulated by miRNAs would greatly help the study of miRNA function in animals (Ambros 2001).

Two modes of miRNA-directed target inhibition have been demonstrated. The same small RNA can cause degradation of its target mRNA or block its translation depending on the degree of miRNA–target sequence complementary (Hutvagner and Zamore 2002; Doench et al. 2003). Target RNAs containing sequences with perfect complements of the miRNA (or exogenously supplied short interfering RNA [siRNA]) are cleaved by the RNA-induced silencing complex (RISC) ribonuclease (Hutvagner and Zamore 2002; Martinez et al. 2002; Zeng et al. 2002). Endogenous plant miRNAs have been shown to regulate target RNAs by RNA interference (RNAi) involving perfect or near-perfect target site complementarity (Llave et al. 2002b; Kasschau et al. 2003; Palatnik et al. 2003; Tang et al. 2003; Xie et al. 2003). Targets for plant miRNAs have been identified on a genome-wide scale by searches that require a high degree of sequence complementarity (Rhoades et al. 2002). However, this approach did not find targets for known animal miRNAs. The animal miRNAs tested until now pair imperfectly with their targets and act to control translation. Indeed, systematic analysis of the complete *C. elegans* miRNA complement has confirmed the absence of targets with perfect or near-perfect sequence complementarity (Ambros et al. 2003).

To date targets have been experimentally validated for just three animal miRNAs: the *lin-4* targets *lin-14* and *lin-28* (Wightman et al. 1993; Ha et al. 1996; Moss et al. 1997; Olsen

Received August 15, 2003; Accepted September 24, 2003; Published October 13, 2003
DOI: 10.1371/journal.pbio.0000060

Abbreviations: bHLH, basic helix–loop–helix; CDS, coding sequence; domT, domain bit score threshold; EGFP, enhanced green fluorescent protein; *E(spl)*, *Enhancer of split* gene; GFP, green fluorescent protein; kb, kilobase; miRNA, microRNA; mRNA, messenger RNA; nt, nucleotides; ORF, open reading frame; ptc–Gal4, yeast GAL4 transcriptional activator expressed under control of the *Drosophila patched* gene promoter; RISC, RNA-induced silencing complex; RNAi, RNA interference; SD, standard deviation; siRNA, short interfering RNA; 3′ UTR, untranslated region of a messenger RNA following the coding sequence

Academic Editor: Ronald H. A. Plasterk, Utrecht University

*To whom correspondence should be addressed. E-mail: cohen@embl.de

☯These authors contributed equally to this work

and Ambros 1999; Seggerson et al. 2002), the *let-7* targets *lin-41* and *lin-57/hbl-1* (Reinhart et al. 2000; Slack et al. 2000; Abrahante et al. 2003; Lin et al. 2003), and the *bantam* target *hid* (Brennecke et al. 2003). These miRNA–target duplexes contain mismatches, gaps, and G:U basepairs at different positions. Even allowing for G:U basepairs, the longest contiguous alignments in these examples range from 8 to 10 nt. Such limited information content makes it difficult to identify targets within whole-genome or transcriptome databases, since standard alignment methods produce many false positives with such short variable sequences. Furthermore, the small number of validated examples makes the development and benchmarking of a generally applicable computational method problematic at present. Here we present a screen for miRNA targets in *Drosophila* that identifies all of the previously known miRNA targets and we demonstrate that it successfully predicts new targets that we validate experimentally.

Results

Database Design

For each of the validated miRNA–target pairs, functional target sites are located in the 3′ untranslated region (UTR) of the mRNA and are conserved in the 3′ UTRs of the homologous genes from related species (Wightman et al. 1993; Moss et al. 1997; Pasquinelli et al. 2000; Brennecke et al. 2003). We used pairwise comparison of the 3′ UTRs of orthologous genes in related genomes to identify conserved 3′ UTR sequences. Figure 1A shows the resulting pattern of 3′ UTR conservation for the known targets in worms and flies. The vast majority of miRNA target sites (red bars in Figure 1A) are located in blocks of conserved sequence (white blocks in Figure 1A). Figure 1B shows cross-genome conservation of these miRNA target sites. A striking pattern of uninterrupted conservation emerges at the end of the target sequences that pair with the 5′ end of the miRNAs.

To permit genome-wide searches for targets of *Drosophila* miRNAs, a conserved 3′ UTR database was prepared by comparison of *Drosophila melanogaster* and *Drosophila pseudoobscura* 3′ UTRs. As very few 3′ UTRs are defined by cDNA sequence data in *D. pseudoobscura*, we used genomic sequence following the last exon of the *D. pseudoobscura* gene as the orthologous UTR (see Materials and Methods). Last exons were reliably detected in *D. pseudoobscura* for approximately two-thirds of *D. melanogaster* genes. On average, 22% of the *D. melanogaster* 3′ UTR sequence is conserved in the predicted *D. pseudoobscura* 3′ UTR. Much of this reflects isolated blocks of very high conservation interspersed among less-conserved sequence. Use of conserved 3′ UTRs reduces the expected number of sequence matches that would occur at random by 4- to 5-fold in relation to full-length 3′ UTRs, and severalfold further compared to the full transcriptome. We considered using the *Anopheles gambiae* genome to extend the cross-species' comparison. Although genome annotation identifies orthologs for two-thirds of *D. melanogaster* genes (Zdobnov et al. 2002), we were unable to identify the last *D. melanogaster* exon for approximately half of these. We therefore chose not to require conservation in *Anopheles*, but use it as an additional level of validation for predicted *Drosophila* targets, where possible.

Screening Strategy

We have adopted a two-step approach to target identification that combines a sensitive sequence database search with an RNA folding algorithm to evaluate the quality of the RNA duplex formed between the miRNA and its predicted targets. We examined the known target sites for *lin-4*, *let-7*, and *bantam* for common features. All of these sites showed better complementarity to the 5′ end of the miRNA, with no obvious common features elsewhere (Figure 2A and 2B). There were few sequence mismatches or G:U basepairs in the alignment of the first eight residues at the 5′ end of the miRNA. We used the alignment tool HMMer (Eddy 1996) to search for sequences complementary to the first eight residues of the miRNA, allowing for G:U mismatches. Where possible, the corresponding sites were also identified in the *D. pseudoobscura* 3′ UTR, and the sites from both genomes were considered, since the regions outside of the sequence match can vary between the two organisms, leading to differences in subsequent steps (see below).

The identified sequences were extended to the length of the miRNA plus five residues to allow for bulges and were evaluated for their ability to form energetically favorable RNA–RNA duplexes with the miRNA using Mfold, which combines knowledge of known RNA structures with thermodynamic parameters, such as those involved in basepairing to evaluate the free energy of folding (ΔG) (Mathews et al. 1999; Zuker et al. 1999). Mfold requires a single linear sequence as input, so each predicted target was linked to the miRNA using a standard hairpin-forming linker sequence (GCGGGGAC-GC). An example of the Mfold output is shown in Figure 2C for the top-scoring *bantam* miRNA target site that we had previously identified in the 3′ UTR of *hid* (Brennecke et al. 2003).

The Mfold free energy of folding (ΔG) was determined for each predicted target, which allows predicted sites to be ranked according to ΔG. However, ΔG depends on miRNA length and GC content, so it is not possible to distinguish systematically real targets from random matches using the raw ΔG score or to compare different miRNAs. Instead, we calculated *Z*-values as a measure of nonrandomness, with an average random site scoring $Z = 0$ (Z = standard deviations [SD] above the mean of background matches). Figure 2D shows the distribution of folding energies for predicted targets of the *bantam* miRNA compared to 10,000 randomly selected target sequences.

Most of the previously validated targets have more than one predicted miRNA-binding site in their 3′ UTRs. Use of the *Z*-value allows us to add the scores of several sites within one UTR by selecting only those scores that are different from background matches. This is not possible with ΔG alone because even average random matches have favorable energy values (Figure 2D) and the sum of several average random matches in a UTR could score better than a single true site. We have selected $Z \geq 3$ as a cutoff value, as folding energies of more than 3 SD above the mean ($Z \geq 3$) are expected to occur for only 0.3% of random matches. Use of a higher *Z*-value increases the likelihood that predictions are correct, but also increases the risk of missing out contributions from real sites of lower folding energy. For example, only three of the five conserved *bantam* sites previously identified in the *hid* 3′ UTR score $Z \geq 3$ (with the best site at $Z = 7.4$). We evaluated our

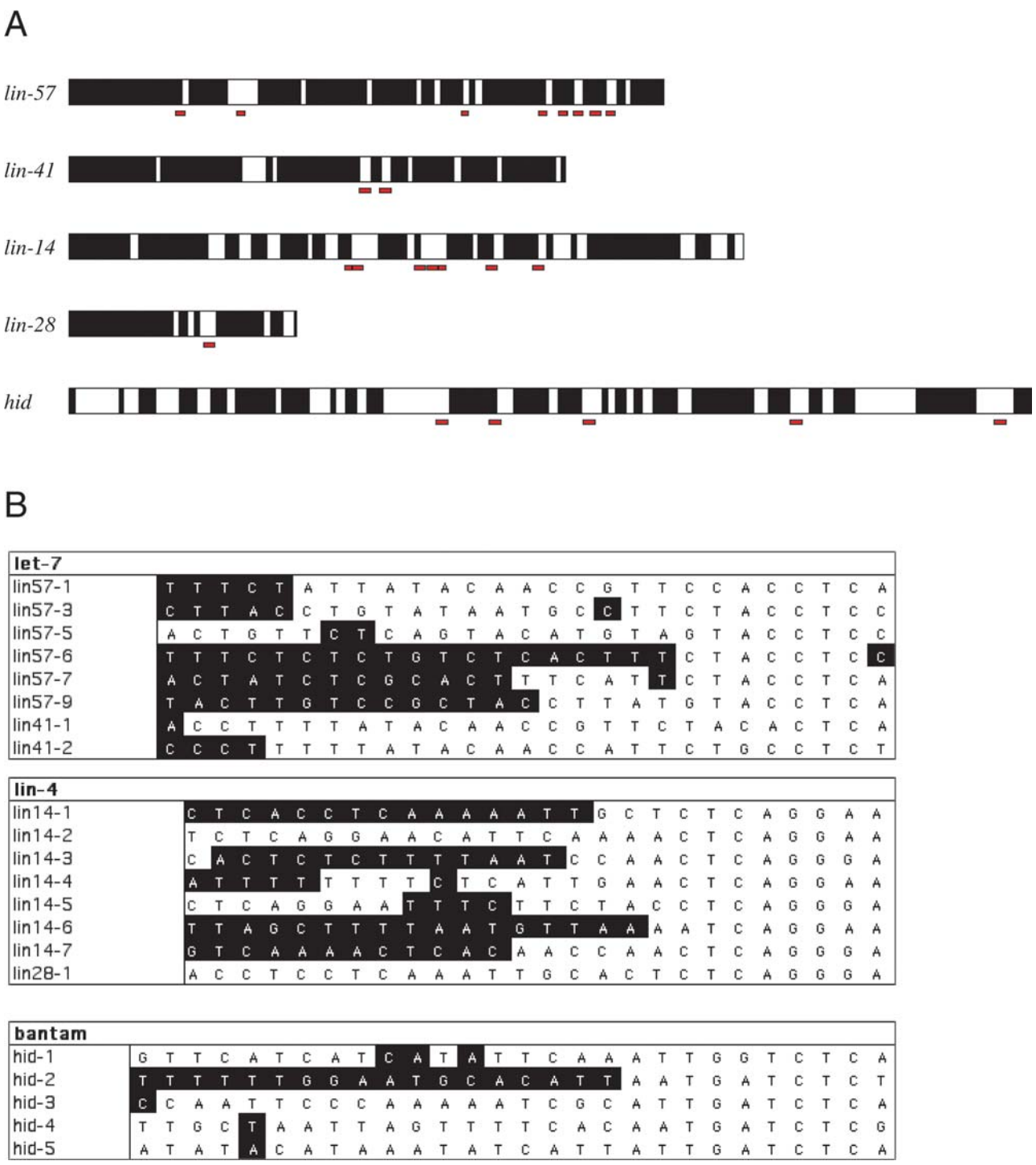

Figure 1. Features of Known miRNA Target Sequences

(A) Comparison of sequence conservation in the 3′ UTRs of miRNA target genes. For *lin-14*, *lin-28*, *lin-41*, and *lin-57*, comparison was between *C. elegans* and *C. briggsae*. For *hid*, comparison was between *D. melanogaster* and *D. pseudoobscura*. White regions indicate conservation, and black regions are not conserved under the conditions used for producing the 3′ UTR database (see Materials and Methods for details). The positions of predicted miRNA target sites from the literature are shown in red. Most of these UTRs contain multiple predicted target sequences, and while regulation of the UTR has been experimentally validated in each case, most individual sites have not been tested for function.

(B) Detailed comparison of the pattern of sequence conservation in the conserved sites. Target site length is miRNA length plus 5 nt. For *lin-57* we excluded three of the eight previously predicted sites that were not located in conserved sequence blocks and included a newly identified ninth site that is conserved. White type on black indicates residues that are not identical in the target sites in the two genomes. Black type on white indicates identity. All residues basepairing with positions 2–8 of the miRNA are identical in the conserved sites in both genomes.

DOI: 10.1371/journal.pbio.0000060.g001

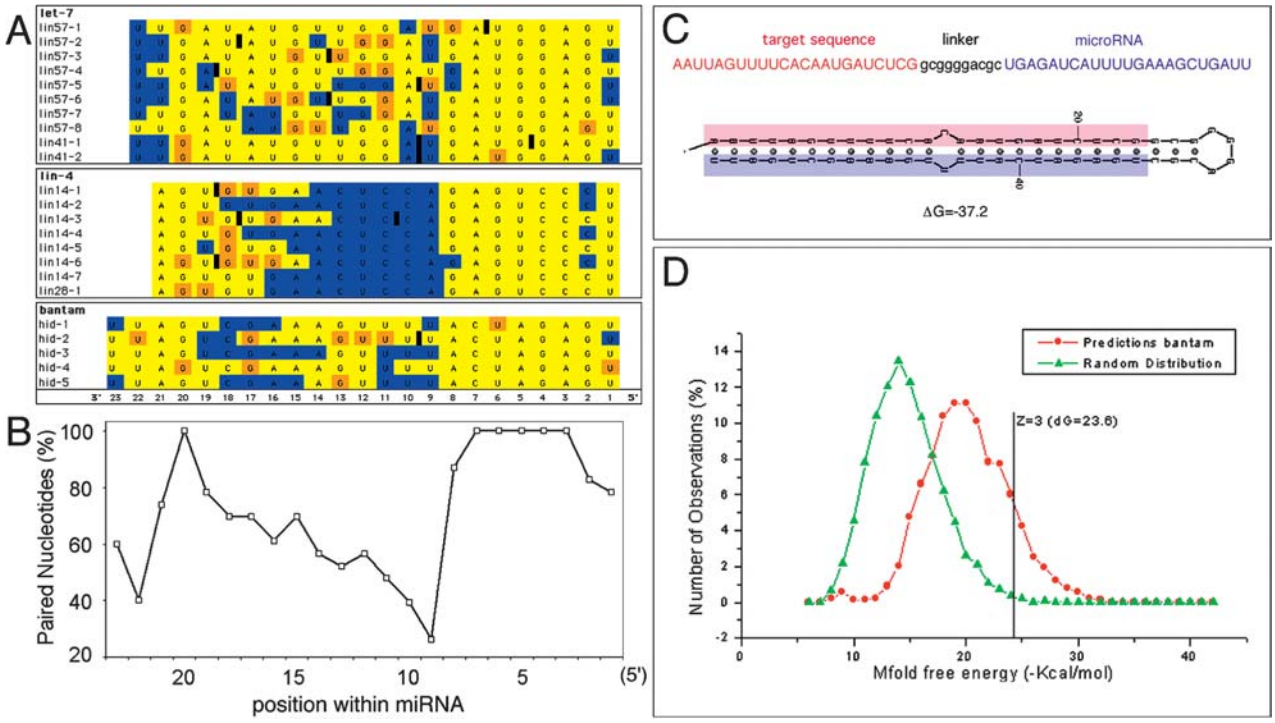

Figure 2. miRNA Target Prediction Strategy

(A) *let-7*, *lin-4*, and *bantam* miRNA sequences showing the pattern of basepairing to their known targets. Yellow indicates a conventional basepair. Orange indicates a G:U basepair. Blue indicates a mismatch. The black bars indicate the positions of loops in the target sequence. Note that the extra bases that form the loops in the target sequences are not shown. Sequences are shown at the length of the miRNA.

(B) Quantitation of the data from (A). This comparison shows that the 5′ ends of the miRNA are always involved in good pairing with target sequences and suggests that searches for complementarity to the first eight residues of the miRNA would select all known targets.

(C) Graphic representation of the Mfold output for the *bantam* miRNA and a target site from the 3′ UTR of the *hid* gene. To use Mfold, it is necessary to join the predicted target site (red) and the miRNA (blue) into a single sequence using a hairpin-forming linker sequence. In this example, the target sequence and the miRNA are the same length, so the additional 5 nt in the tail of the predicted target sequence are not shown.

(D) Plot of the Mfold free energy distribution for 10,000 random sequences (green) and for predicted targets of the *bantam* miRNA (red). X-axis: ΔG calculated for each site by Mfold.

DOI: 10.1371/journal.pbio.0000060.g002

predictions by the best single site in the 3′ UTR (Z_{max}) and by the sum of sites with $Z \geq 3$ (Z_{UTR}).

Tests with Previously Validated Targets

Table 1 summarizes the performance of the method in predicting the known targets of the *C. elegans* miRNAs *lin-4* and *let-7* and the *Drosophila* miRNA *bantam*. The *Drosophila hid* gene ranked first of all predicted *bantam* targets sorted by the single best site (Z_{max}) or by the sum of sites (Z_{UTR}). All of the known targets of *lin-4* and *let-7* were found when their 3′ UTRs were added to the *Drosophila* 3′ UTR database. Like *hid*, the *let-7* target *lin-57* ranked near the top of the list by both measures. With several sites predicted of $Z \geq 3$, *lin-57* ranked first by Z_{UTR}. Its best single site ranked in position 2 ($Z = 6.8$). *C. elegans lin-14* was predicted to contain multiple *lin-4* sites (Wightman et al. 1993). Three of these scored $Z \geq 3$. *lin-14* ranked first when the list of predicted *lin-4* targets was sorted for Z_{UTR},although the best single site in *lin-14* placed it in position 20 ($Z = 4.3$). The *lin-4* target *lin-28* and the *let-7* target *lin-41* ranked highly when the lists were sorted for the best single site, but ranked lower when multiple sites were summed because they had few high-scoring sites. The *Drosophila* homolog of *lin-41*, *dpld*, also ranked high among *let-7* targets ($Z = 5.6$; see below). We compared our results with previous target predictions from the literature that have not been experimentally validated (Table 1). Our screen supports some of them (e.g., *let-7* regulating *lin-14*), but we consider others unlikely because they rank very low on their lists or have no sites of $Z \geq 3$ (e.g., *let-7* and *lin-28* or *miR-4* and *m4*). The predicted *miR-14* target *Drice* (Xu et al. 2003) is unlikely to be valid because the site is not conserved in the predicted *Drice* 3′ UTR from *D. pseudoobscura*.

This analysis shows that all known targets were detected and ranked among the top-scoring predictions in genome-wide searches. This suggests that other valid targets should rank among the small number of best predictions that can be tested experimentally. Of particular interest were three miRNAs for which we predicted clusters of functionally related targets: *miR-7*, the *miR-2* family, and *miR-277* (Table 2; Table 3). Clustering of top-scoring sites in a group of related genes is likely to be significant when it arises from an unbiased genome-wide analysis. *miR-7* and *miR-2* were selected for target validation.

miR-7 Regulates Notch Targets

Among the top 10 predictions for *miR-7*, we found *Enhancer of split* (*E(spl)*) and *Bearded* (*Brd*) complex genes (Figure 3A). *HLHm3* encodes a basic helix–loop–helix (bHLH) transcrip-

Table 1. Assessment of Predictions for Known and Predicted miRNA Targets

miRNA–Target Pair	ΔG	Z_{Max}	Rank Z_{Max}	# $Z \geq 3$	Rank Z_{UTR}	References
Confirmed Pairs						
lin-4/lin-14	−29.9	4.3	20	3	1	Wightman et al. (1993); Ha et al. (1996)
lin-4/lin-28	−30.9	4.6	8	1	15	Moss et al. (1997)
let-7/lin-41	−32.3	6.4	3	2	20	Reinhart et al. (2000); Slack et al. (2000)
let-7/lin-57 (hbl-1)	−33.4	6.8	2	14	1	Abrahante et al. (2003); Lin et al. (2003)
bantam/hid	−37.4	7.4	1	3	1	Brennecke et al. (2003)
Predicted Pairs						
lin-4/lin-41	−28.9	4.0	32	1	36	Slack et al. (2000)
lin-4/lin-57	−21.6	1.7	361	0	—	Abrahante et al. (2003); Lin et al. (2003)
let-7/lin-14	−35.1	7.2	1	13	2	Reinhart et al. (2000)
let-7/lin-28	−20.6	2.8	861	0	—	Moss and Tang (2003)
miR-13a/hb	—	—	—	0	—	Abrahante et al. (2003)
miR-4/hb	—	—	—	0	—	Abrahante et al. (2003)
miR-3/hb	—	—	—	0	—	Abrahante et al. (2003)
miR-11/HLHm8	−29.4	4.7	27	1	46 (predicted UTR)	Lai (2002)
miR-4/m4	−21.5	2.1	272	0	—	Lai (2002)
miR-7/HLHm3	−37.3	7.0	2	1	16	Lai (2002)
miR-7/Tom	−34.5	6.1	5	2	1	Lai (2002)
miR-14/Drice	—	—	—	0	(Site not conserved)	Xu et al. (2003)

Confirmed pairs indicates experimentally validated target 3′ UTRs. ΔG, Z_{Max}, and Z_{UTR} are defined in the text. Predicted pairs indicates examples predicted in the literature for which there was no experimental validation. The *let-7/lin-14* pair ranks very high on the list of *let-7* predictions and is likely to be a functional target. The *lin-4/lin-41* pair requires experimental validation. The other *C. elegans* predictions cannot be distinguished from random matches. The 5′ end of the K box shows sequence complementarity to the *miR-2/miR-13* family and to *miR-6* and *miR-11* (Lai 2002). The prediction of *HLHm8* as a target for *miR-11* seems plausible (using predicted UTR), as do the two *miR-7* GY box-based predictions. None of the conserved sites predicted for *Drosophila hunchback* (Abrahante et al. 2003) were on our lists because of interrupted 5′ alignments.
DOI: 10.1371/journal.pbio.0000060.t001

tional repressor; *Tom* and *m4* encode Brd family proteins. The bHLH repressor *hairy* was also among the top 10. These sites were conserved in the orthologous genes from *Anopheles*, when those could be identified. This prompted us to examine all the genes in *E(spl)* and *Brd* complexes for *miR-7* sites. We found possible target sites in many of them. Alignment of these sites showed a pattern of 5′ end conservation quite similar to that for validated targets, with no mismatches and few G:U basepairs for about half of these genes (Figure 3B).

To assess the validity of some of the predicted targets, transgenic flies expressing the *miR-7* miRNA and several sensor transgenes were prepared. A genomic fragment containing the *miR-7* hairpin was cloned into the 3′ UTR of a UAS–DSRed2 plasmid to allow GAL4-dependent expression of *miR-7*. The 3′ UTRs of *HLHm3*, *m4*, and *hairy* were cloned into a tubulin promoter–EGFP (enhanced green fluorescent protein) reporter plasmid. As a control, a specific *miR-7* sensor transgene was produced by cloning two copies of a perfect complement of the *miR-7* miRNA sequence into the 3′ UTR of the tubulin promoter–EGFP reporter. The *miR-7* sensor was expressed uniformly in the wing imaginal disc. GAL4-dependent expression of *miR-7* miRNA reduced expression of *miR-7* green fluorescent protein (GFP) sensor transgene (Figure 3C). As the target sites in the sensor construct are perfect complements of the *miR-7* miRNA, this is expected to be due to RNAi. GAL4-dependent expression of *miR-7* also reduced expression of the *m4* 3′ UTR sensor transgene (Figure 3D). The *miR-7* site in *m4* is identical in *D. pseudoobscura* and conserved in *Anopheles*.

Expression of *miR-7* also caused a clear downregulation of the *hairy* 3′ UTR sensor transgene, although its overall level of expression was lower in the wing disc (Figure 3E). The *hairy* gene has been cloned and cDNAs sequenced from two additional insect genomes: the flour beetle *Tribolium castaneum* and *Drosophila simulans*. The predicted *miR-7*-binding site is conserved in these genomes, as well as in *Anopheles*, and shows striking conservation of alignment at the 5′ and 3′ ends of the predicted miRNA-binding site (Figure 3F). The level of expression of the *HLHm3* 3′ UTR sensor was too low to be reliably studied, but also showed regulation by *miR-7*. Again, the *miR-7* site in *HLHm3* is identical in *D. pseudoobscura*. These observations validate the utility of the screen in predicting new miRNA targets.

To assess *miR-7* function in vivo, we examined wings in which *miR-7* was overexpressed under ptc–Gal4 control. Notching of the wing margin was observed (Figure 3G), which is characteristic of reduced *Notch* signaling (de Celis and Garcia Bellido 1994; Diaz-Benjumea and Cohen 1995; Rulifson and Blair 1995; Micchelli et al. 1997). The *Notch* target *cut* was expressed at reduced levels in the *miR-7*-

Table 2. Top Ten Predictions for *miR-7* and *miR-2a*

miR-7	ΔG	Z_{max}	# $Z \geq 3$	Z_{UTR}	Gene	Alignment	Flags	*Ag* $Z \geq 3$
1	−38.7	7.47	1	7.47	*CG14989-RB*	ACAGCAGAAUCACGC-AGGG-CUUCCA UGUUGUUUUAGUG--AUC--AGAAGGU ***+**+******---**---******	–	+
2	−37.3	7.03	1	7.03	*HLHm3*	GCAACAAGAUCCGUU----GUCUUCCA UGUUGUUUUAG----UGAUCAGAAGGU +******+***--------********	–	NF
3	−35.3	6.39	1	6.39	*CG17657-RA*	ACAACGGUUAAG---CGCUGCGUCUUCCA UGUUGUU-----UUAGUGAU-CAGAAGGU *****++--------*+**+-********	–	NF
4	−35.0	6.29	1	6.29	*hairy*	ACAGCAAAUCAG--CAAA--AGUCUUCCA UGUUGUUU----UAGU--GAUCAGAAGGU ***+****-----**----********	–	+
5	−34.5	6.13	2	11.78	*Tom*	-UAGCC-GAAUCAUU-GUCUUCCA UGUUG-UUUUAGUGAUCAGAAGGU -+**+--+*****+**-*******	–	+
6	−33.9	5.94	1	5.94	*hep*	GCAGCAACAGUCGC-AGUUUUUCA UGUUGUU-UUAGUGAUCAGAAGGU +**+***-*+**+**-***+**+***	5′ cons CDS+	NF
7	−33.8	5.91	1	5.91	*CG8944-RA*	ACGACAAGAUCAAGCGCUACGUCUG-CCA UGUUGUUUUAG----UGAU-CAGA-AGGU **+****+***----+***-****--***	5′ helix CDS+	NF
8	−33.1	5.68	1	5.68	*CG10540-RA*	-CGACAAAGCG--GCCCAAUAGUCUUCCA UGUUGUUUU--AGUG----AUCAGAAGGU -*+*****+----+*----**********	–	–
9	−31.8	5.27	1	5.27	*CG10444-RA*	GCGACC-AAAA-CAG--AGUCUUCCA UGUUG-UUUU-AGU-GAUCAGAAGGU +*+**--***--**---*********	–	NF
10	−31.5	5.17	1	5.17	*m4*	-CAGCUUU--AAUCAAC---GUCUUCCG UGUUG---UUUUAGU--GAUCAGAAGGU -***+-----*****-----*******+	–	+

miR-2a	ΔG	Z_{max}	# $Z \geq 3$	Z_{UTR}	Gene	Alignment	Flags	*Ag* $Z \geq 3$
1	−39.0	6.78	2	11.85	*CG1969-RB*	----------GCUGGCUGGC-GGUG CGAGUAGUUUCGACCGAC--ACUAU ----------********---*+*+	5′ cons 5′ helix mispairing	NF
2	−38.6	6.66	1	6.66	*CG4269-RA*	GCUCCUG--CAU-GGAUUGGCUGUGAUA CGAG---UAGU-UUC-GACCGACACUAU ****-----**--+*-+***********	–	–
3	−38.0	6.49	1	6.49	*reaper*	-CUCAUCAAAGCGA---UUGUGAUA CGAGUAGUUUCG--ACCGACACUAU -***********-----+*******	–	NF
4	−34.3	5.42	1	5.42	*GLaz*	GCUUUGAU----GAGC--GCUGUGAUA CGAG----UAGUUUCGACCGACACUAU ***+--------+***--*********	mispairing	–
5	−33.5	5.19	1	5.19	*BG:DS05899.3*	GUUCAUCCCUU--GGCGUUG-GGCUGUGU-UA CGAGUAG----UUUCG----ACCGACAC-UAU *+*****------+**-----*******---**	5′ helix	NF
6	−33.2	5.1	1	5.1	*Scr*	GCUCGGUG-GGAGUG-GGUG-GUGGUG CGAGU-A-GUUUCG-ACCG-ACACUAU ****+-*---+***+---**+---***+*+	–	–
7	−33.0	5.04	1	5.04	*hbs*	---CAUGCGGC-GCUCGAAGGCUGUGAUA CGAGUA-GUU-UCGA----CCGACACUAU ---***-*++---***----**********	–	–
8	−32.8	4.99	1	4.99	*amon*	GUUCAA-UAAAAGUGCUGGCUGUG--- CGAGU-AGUUU---CGACCGACACUAU *+****--+***---**********---	5′ helix	–
9	−32.5	4.9	1	4.9	*grim*	GCUCAAUCAAAGCGCA---UUGUGAU- CGAGU-AGUUUCG---ACCGACACUAU *****-*******------+******-	–	NF
10	−32.1	4.78	2	8.01	*CG7187-RA*	GCUUUGAU----GAGC--GCUGUGGUG CGAG----UAGUUUCGACCGACACUAU ***+--------+***--******+*+	5′ cons mispairing	NF

ΔG, Z_{max}, and Z_{UTR} are explained in the text. *Alignment*: The target site is shown on top. Conventional basepairs are indicated by asterisks and G:U basepairs by plus signs; mismatches and gaps are indicated by dashes. *Flags*: The "5′ conservation" (5′ cons) flag identifies sites that differ in the two genomes at any residue complementary to positions 2–7 of the miRNA. The "5′ helix" flag identifies sites that do not have at least six contiguous basepairs in positions 1–8. The "CDS+" flag indicates that the predicted site overlaps coding sequence on the same strand; "CDS−" indicates that the overlap is on the opposite strand. In some cases, Mfold structures include basepairs that are not between the miRNA and its target. "Mispairing" flags sites with artificially high folding energies. *Ag* ($Z \geq 3$)*: Anopheles* genes that cannot be reliably identified by our criteria are indicated as "NF." For the cases in which the orthologous *Anopheles* gene was found, the presence of a target site with a *Z* score ≥ 3 is indicated by a plus sign. Absence of a site or a *Z* score < 3 is indicated by a minus sign. Heavy outlining indicates those loci that would pass stringent filtering of the lists using the flags and requiring lack of a conserved target in an *Anopheles* ortholog.

DOI: 10.1371/journal.pbio.0000060.t002

Table 3. *miR-277* Targets

Rank	Gene	Function	Enzyme	# $Z \geq 3$	Z_{UTR}
1	*CG31651*	Protein GalNAc transferase	EC:2.4.1.41	2	7.31
2	*CG5599*	Val Leu Ile degradation	EC:2.3.1−	2	6.53
3	*CG1673*	Val Leu Ile degradation	EC:2.6.1.42	1	5.75
4	*fz*	Cell polarity		1	4.89
5	*CG8199*	Val Leu Ile degradation	EC:1.2.4.4	1	4.53
6	*CG18549*	—		1	4.23
7	*CG1140*	Val Leu Ile degradation	EC:2.8.3.5	1	3.9
8	*scu*	Val Leu Ile degradation	EC:1.1.1.35	1	3.81
9	*CG15093*	Val Leu Ile degradation	EC:1.1.1.31	1	3.79
10	*CG7740*	Membrane protein		1	3.64
11	*CG17896*	Val Leu Ile degradation	EC:1.2.1.27	1	3.61

Required: $Z \geq 3$ for *D. melanogaster, D. pseudoobscura, Anopheles*.
DOI: 10.1371/journal.pbio.0000060.t003

expressing cells at the wing margin (Figure 3E); *wingless* expression was also reduced (data not shown). Although bHLH transcription factors and *Brd*-like genes of the *E(spl)* complex are not strictly required for all aspects of Notch activity at the wing margin, they are required for *cut* expression (Ligoxygakis et al. 1999). *miR-7* expression might provide a means to simultaneously downregulate these and other proteins, which might otherwise function redundantly to mediate Notch activity in the wing margin. We also noted reduced spacing of veins 3 and 4 in these wings, which may also reflect reduced Notch activity in controlling tissue growth (Baonza and Garcia-Bellido 1999). *E(spl)* genes are also expressed in proneural clusters, where they are required for sense organ determination and bristle patterning. Clones of cells lacking multiple genes of the *E(spl)* complex form extra bristles (de Celis et al. 1996). Consistent with this, we observed that expression of *miR-7* under ptc–Gal4 control causes ectopic bristles and bristle duplication in the scutellum (data not shown). Taken together, these findings support the prediction that *miR-7* miRNA regulates expression of bHLH and Brd-like proteins encoded by *hairy* and the *E(spl)* and *Brd* complex genes and implicates *miR-7* as a possible regulator of *Notch* target gene expression. A more detailed analysis of *miR-7* function will require isolation of lack of function mutations in the *miR-7* gene.

Lai (2002) has reported complementarity between some miRNAs and sequence elements known as K boxes, Brd boxes, and GY boxes in the 3′ UTRs of *E(spl)* and *Brd* complex genes. K boxes and Brd boxes have been implicated in posttranscriptional regulation, though no function was assigned to the *miR-7* complementary GY boxes (Lai and Posakony 1997; Lai et al. 1998). The presence of GY boxes in several *E(spl)* and *Brd* complex genes, as well as in *hairy* and *extramachrochatae* has been reported (Lai and Posokony 1998). Based on the presence of these boxes, Lai (2002) predicted *miR-7* target sites in *HLHm3* and in *Tom*. We extend these predictions to a much larger gene family and provide experimental validation for some of them, indicating that GY boxes participate in the regulation of *Notch* target genes.

miR-2 Regulates Proapoptotic Genes

The proapoptotic genes *reaper* and *grim* were among the top predictions for *miR-2a* and *miR-2b* (see Table 2). *reaper*, *grim*, and the third proapoptotic gene *sickle* are clustered in the genome and show blocks of high conservation in their 3′ UTRs, which include the *miR-2* family target sites (underlined in Figure 4A). Alignment of these sites shows a very similar pattern of predicted miRNA binding (Figure 4B). The corresponding sites are highly similar in *D. pseudoobscura*, but the orthologous genes cannot be identified in *Anopheles*. To evaluate these predictions, we made 3′ UTR sensor transgenes for *reaper*, *grim*, and *sickle*. The expression level of the *reaper* 3′ UTR sensor transgene was too low to be reliably studied in transgenic flies, so we used an in vitro assay to assess its function. *Drosophila* Schneider S2 cells express the *miR-2* family of miRNAs (Lagos-Quintana et al. 2001). S2 cells were transfected with the *reaper* 3′ UTR construct or with a version of the construct in which the *miR-2*-binding site was mutated (the residues shown in Figure 4B were replaced by a *Not*I site). A low level of GFP expression was detected in immunoblots of cells transfected with the *reaper* 3′ UTR construct (Figure 4C, lane 2). The level of GFP expression was much higher in cells transfected with the mutated UTR construct, suggesting that the endogenous *miR-2* family miRNAs in S2 cells can repress expression of a reporter construct via the *reaper* 3′ UTR. The *grim* and *sickle* 3′ UTR sensor transgenes were expressed at detectable levels in transgenic flies and were both downregulated by expression of *miR-2b* in the wing disc (Figure 4D and 4E). The *miR-13* family is similar in sequence to the *miR-2* family. Experimental validation will be needed to determine whether *reaper*, *grim*, and *sickle* are also regulated by *miR-13*. Identification of *reaper*, *grim*, and *sickle* as targets suggests that *miR-2* family miRNAs may be involved in control of apoptosis.

Statistical Evaluation of Target Predictions

Although a number of the top-ranking sites identified in our screen have been experimentally validated, we wanted to assess the likelihood that sites with equivalent scores can be found by chance. To do so, we calculated *E*-values for the

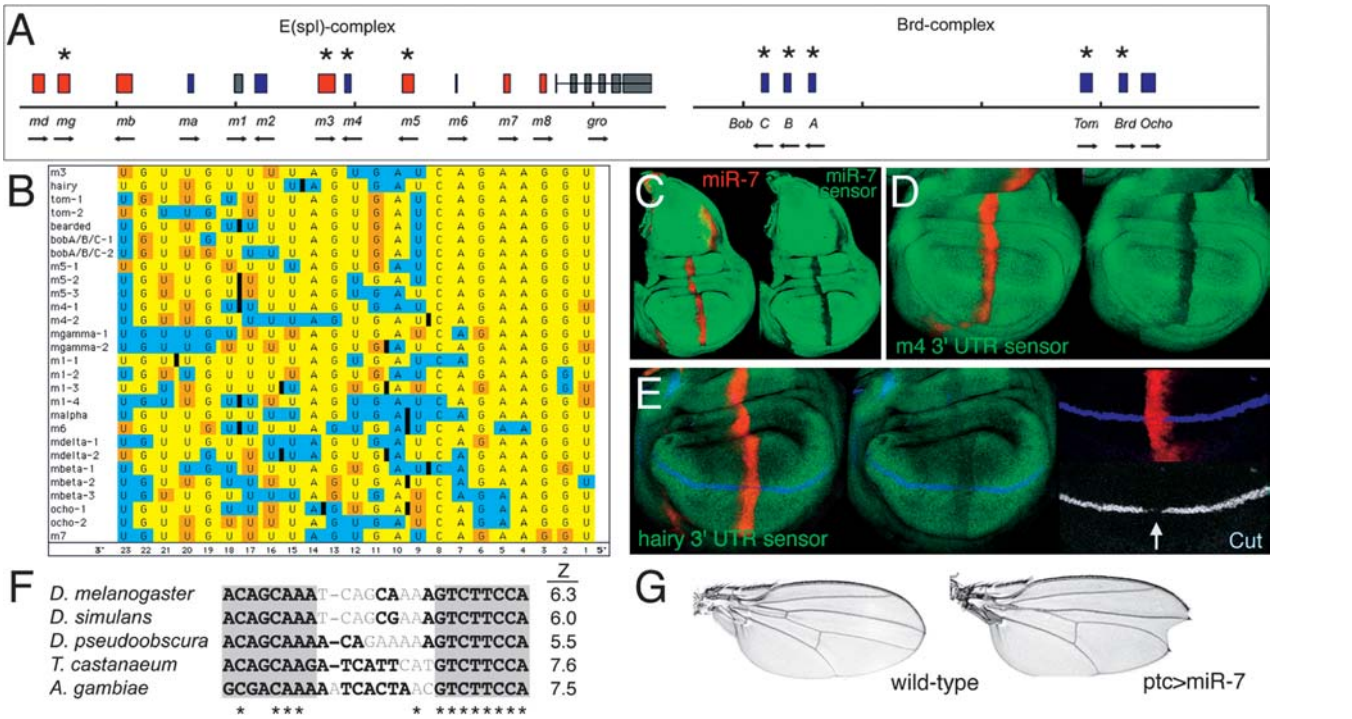

Figure 3. Experimental Validation of *miR-7* Targets

(A) Schematic representation of the *E(spl)* and *Brd* gene complexes, which contain multiple predicted *miR-7* target genes. bHLH-type transcriptional repressors are shown in red. Brd-type proteins are shown in blue. Other transcripts in the *E(spl)* cluster are in gray. Black asterisks indicate sites with no mismatch in the first eight residues (likely to be valid sites).

(B) *miR-7* miRNA sequence showing the pattern of basepairing with target sites in *E(spl)* and *Brd* complex genes sorted in order of predicted folding energy. Yellow indicates a conventional basepair. Orange indicates a G:U basepair. Blue indicates a mismatch. The black bars indicate the position of loops in the target sites.

(C) Expression of the *miR-7* sensor transgene is shown in green. Expression of the red fluorescent protein *miR-7* miRNA under ptc–Gal4 control is shown in red. The right panel shows the *miR-7* sensor alone.

(D and E) Expression of the *m4* 3′ UTR and *hairy* 3′ UTR sensor transgenes (green) were downregulated by *miR-7* (red). Expression of the *hairy* 3′ UTR sensor was much lower than the *m4* 3′ UTR sensor overall. Cut protein, shown in blue, was downregulated in *miR-7* expressing cells. The right panel shows a second example of Cut repression. The lower panel shows Cut channel alone.

(F) ClustalW alignment of *miR-7* target sites in the 3′ UTRs of *hairy* from several species. Asterisks indicate sequence identity. Black type indicates basepairs by Mfold (including G:U basepairs). Gray shading highlights the conserved miRNA–target binding region in all five species.

(G) Cuticle preparations of a wild-type adult wing and a wing expressing *miR-7* under ptc–Gal4 control in the region between veins 3 and 4. Note the notching of the wing and the reduction of the region between veins 3 and 4, leading to partial fusion proximally. The size of the posterior compartment was increased apparently to compensate for reduction of the vein 3–4 region.

DOI: 10.1371/journal.pbio.0000060.g003

bantam miRNA based on the tail of the cumulative distribution of ΔG values for 10,000 random matches. An *E*-value predicts the number of background matches with a similar or better score (*E*-values scale with database size and are applicable to any distribution profile). Values greater than 1 are not significant, while those close to 0 are very significant. The results are presented on a logarithmic scale for UTRs containing one, two, or four sites of a given ΔG value (Figure 5). The best single *bantam* site in the *hid* UTR had an *E*-value of 1.3. This means that background matches reach RNA–duplex energies similar to the best sites, even in the smaller conserved 3′ UTR database. Indeed, target sites predicted using shuffled *bantam* miRNA sequences give folding energy distributions very similar to the native sequence (data not shown). Although single sites are not statistically significant, the presence of multiple sites within a single UTR can greatly increase the significance of the prediction. Combining the three *bantam* sites ($Z > 3$) predicted in the *hid* 3′ UTR gives an *E*-value of 1.8×10^{-5}. Some single sites are sufficient to mediate regulation by a miRNA; however, we emphasize that the lack of statistical significance for even the best single site means that they require experimental validation.

Additional Validation by Cross-Genome Comparison

One means to improve the significance of the predictions would be to require conservation in a third genome. The two *Drosophila* species are separated by an estimated 30 million years. The mosquito *A. gambiae* is separated from *Drosophila* by 250 million years. Orthologous mosquito genes have been defined for approximately two-thirds of *Drosophila* genes; however, systematic comparison showed great differences in length between orthologous gene pairs (Zdobnov et al. 2002). Indeed, we were able to identify orthologous last exons with confidence for only half of these pairs, or one-third of *D. melanogaster* genes. We have therefore chosen to use conservation in *Anopheles* to provide more stringent evaluation of target site conservation, instead of requiring it generally. The presence of a conserved site with a high *Z* score across all three genomes increases the confidence that the site is functional. To illustrate the utility and limitations of this, we examined the top 100 predictions for *miR-7* and *miR-2*. The

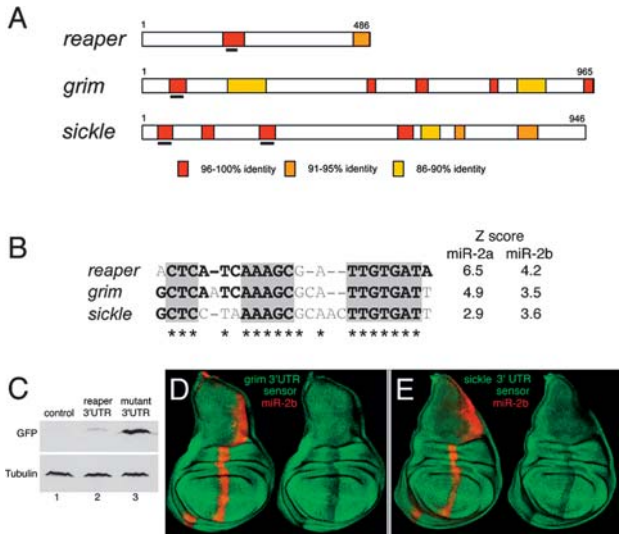

Figure 4. Experimental Validation of *miR-2* Targets

(A) Conservation of sequences in the 3′ UTRs of *reaper*, *grim*, and *sickle* genes of *D. melanogaster* and *D. pseudoobscura*. Blocks of high sequence similarity are color-coded. The location of predicted *miR-2/miR-13* family target sites are underlined in black.

(B) ClustalW alignment of predicted *miR-2/miR-13* family target sites in the *reaper*, *grim*, and *sickle* 3′ UTRs. *Z* scores for *miR-2a* and *miR-2b* are shown for each site. The first bases of the *grim* and *sickle* sites do not pair with the miRNAs. Because we use a hairpin-forming linker sequence, this causes a penalty in Mfold, which gives these sites lower *Z* scores than they should otherwise have.

(C) Immunoblot of S2 cells transfected to express a tubulin promoter–EGFP–reaper 3′ UTR construct (lane 2) or a comparable construct from which the *miR-2/miR-13*-binding site in the UTR was deleted (lane 3). Control cells were transfected with empty vector (lane 1). The blot was probed first with antibody to GFP and then reprobed with anti-tubulin as a loading control.

(D and E) Expression of the *grim* and *sickle* 3′ UTR sensor transgenes (green) was downregulated by *miR-2b* expressed under ptc–Gal4 control (red).

DOI: 10.1371/journal.pbio.0000060.g004

Anopheles orthologs were identified for 52 of the top 100 predicted *miR-7* targets. Of these, 11 had conserved target sites ($Z \geq 3$), including four of the top ten predictions: *hairy*, *Tom*, *m4*, and *CG14989* (see Table 2). For *miR-2a*, forty of the top hundred predictions had a detectable ortholog in *Anopheles*. Of these sites, five were conserved in *Anopheles* ($Z \geq 3$), and none of these were among the top ten predictions. Conservation in *Anopheles* can be used to enrich for sites with a higher probability of being valid, but increases the risk of missing real targets. It is only useful in cases where the orthologous UTR region can be identified, which, for example, is not the case for the validated *miR-2a* targets *grim*, *reaper*, and *sickle*.

miR-277: A Metabolic Switch?

Table 3 shows predicted *miR-277* targets that are conserved ($Z \geq 3$) in *Anopheles*. Of the top 11, seven are enzymes involved in branched chain amino acid degradation (Figure 6). At more relaxed stringency ($Z \geq 2$), two additional enzymes were identified (Figure 6) along with a number of unrelated loci. This striking clustering of functionally related enzymes suggests that *miR-277* regulates the pathway for valine, leucine, and isoleucine degradation by downregulating many of its enzymes and thus acts as a metabolic switch. The degradation of these essential amino acids is presumably

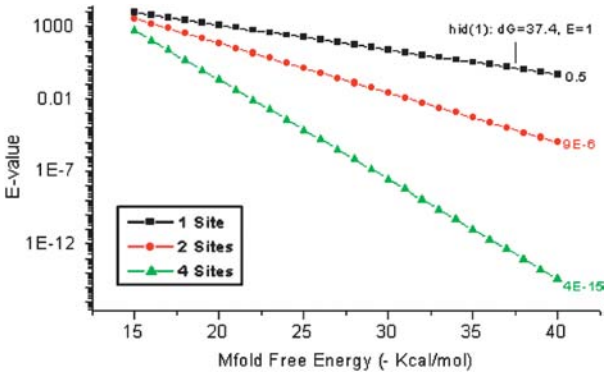

Figure 5. Statistical Evaluation of Predicted Targets

Plot of *E*-values as a function of free energy of folding. Y-axis: logarithmic scale of *E*-values. X-axis: free energy of folding calculated by Mfold. Calculations for one, two, and four sites are shown separately. The position of the best *bantam* site in *hid* is shown for reference.

DOI: 10.1371/journal.pbio.0000060.g005

regulated under conditions of starvation or excess dietary intake. *miR-277* expression has so far only been detected in adult flies (Aravin et al. 2003; Lai et al. 2003), suggesting a role in regulating metabolic responses to environmental conditions. Interestingly, the human homolog of *CG8199* is mutated in maple syrup urine disease. It remains to be determined whether these enzymes are regulated by miRNAs in vertebrates.

Features Shared by Validated Targets

Comparison of the five previously validated targets and the six new targets validated here revealed three features shared by all sites. First, cross-genome comparison showed perfect sequence identity in the target site residues that basepair with residues 2–8 of the miRNA (see Figure 1B). This was also true for the newly validated target sites (data not shown). Second, the pattern of basepairing between the miRNAs and their targets, shown in Figure 2A, suggested that a continuous helix of at least six of the first eight basepairs might be required (allowing G:U basepairs). This was also true for the newly validated target sites (see Figures 3B and 4B). Third, many transcripts in the *D. melanogaster* genome overlap other transcripts on the same strand or on the opposite strand of the DNA. There are many examples of alternate splicing that produce alternate 3′ UTRs so that one UTR variant may include coding exons from another variant. In such cases, the basis for the sequence conservation between genomes is unclear. None of the validated sites from *Drosophila* overlaps coding sequence (CDS) on either strand (this feature was not examined for the *C. elegans* sites).

Target sites that do not share these features are indicated in Table 2. These features can be used to increase the stringency of the screen, by discounting sites that differ from validated targets. For *miR-7* this would eliminate two of the top ten predictions so that the validated targets would constitute three of the remaining top eight predictions. For *miR-2a* this would eliminate four of the top ten predictions, so that the validated targets *reaper* and *grim* would rank in positions 2 and 6. We have chosen not to implement the flags as filters to exclude predictions because they are based on a relatively small set of 11 validated targets, although we note

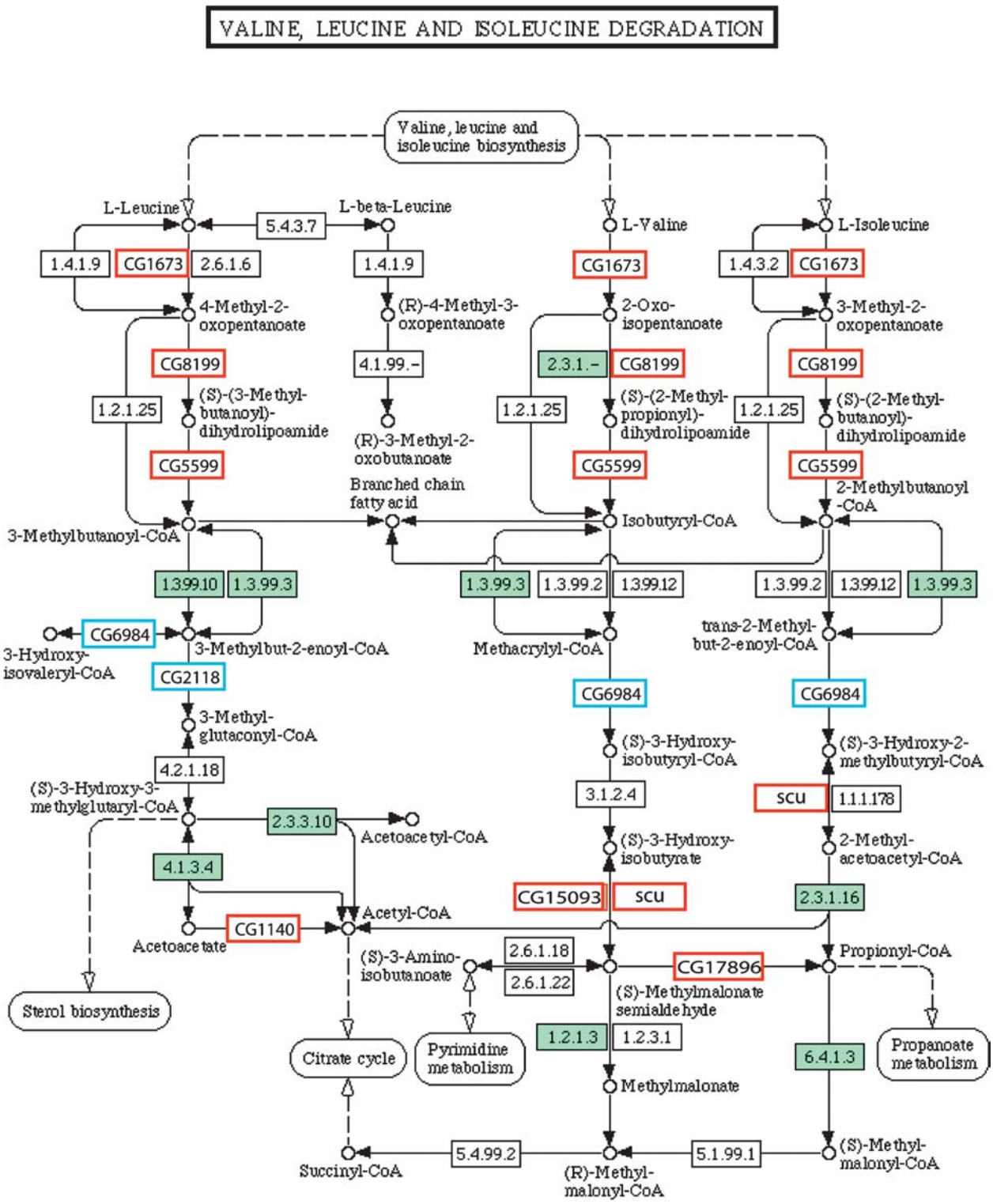

Figure 6. Valine, Leucine, and Isoleucine Catabolic Pathway

Enzymes identified as *miR-277* targets are boxed and identified by CG number. Red boxes required $Z > 3$ in *Anopheles*. Blue boxes required $Z > 2$ in *Anopheles*. In addition to the predicted targets, the other enzymes for which the gene has been identified in *Drosophila* are shaded in green. The metabolic pathway chart is from www.genome.ad.jp/kegg/pathway/map/map00280.html.

DOI: 10.1371/journal.pbio.0000060.g006

that all nine predicted *miR-277* targets would pass such a filter. When more targets are validated, we will learn whether these features have a general predictive value.

Discussion

One of the major limitations in studying animal miRNA function is the difficulty in identifying miRNA targets. Our screening strategy has proven to be useful for predicting new miRNA targets. Three new targets have been experimentally validated for *miR-7* and for *miR-2*, bringing the total number of validated targets of animal miRNAs to 11. In addition, we predict a number of miRNA–target pairs or target families that seem likely to be valid, but require experimental validation. Our study depended on the high-quality annotation of the *D. melanogaster* genome and the availability of the *D. pseudoobscura* genome sequence. Where possible, we have extended the analysis to include evaluation of predicted sites in the *A. gambiae* genome. More complete annotations of the fly and mosquito genomes, aided by cDNA sequencing projects, will increase the number of genes for which orthologous UTR sequences can be defined. This will permit more sensitive and more extensive cross-genome comparison. We also expect improvements to come from further knowledge of the structural requirements of miRNA-target pairing.

Evaluation of Target Predictions

In designing the screening strategy, we considered the balance between sensitivity and specificity. We chose a search strategy that was based on the known examples, but generalized to allow detection of similar targets. By doing so, we risk missing fewer valid targets at the expense of including more false positives, as indicated by our statistical analysis (see Figure 5). To help distinguish false positives from potentially valid targets, we identify features shared by valid targets and, where possible, test predictions for conservation in a third, more distantly related, genome. Both positive and negative results in tests of new predictions will provide a better understanding of how miRNAs bind their targets, perhaps highlighting positions that are particularly critical. This may permit us to achieve both high sensitivity and high specificity in target prediction.

Complete tables of target site predictions are available as Dataset S1 and at www.miRNA.embl.de. These tables report *Z* scores and sequences for the *D. melanogaster* and *D. pseudoobscura* target sites and, where possible, for the *Anopheles* target site. The tables contain flags to identify sites that share the features described above. We recommend using these flags to filter the predictions, but note that this may exclude valid targets. (Filtered tables containing the top 100 predictions are available as Dataset S2 and at www.miRNA.embl.de) We recommend making use of the *Anopheles* data to discount predictions where the othologous gene is identified and the site is absent or has a low *Z* score ($Z < 2$). We emphasize that the absence of an identified orthologous 3′ UTR in the mosquito should not be taken as evidence that a target prediction is not valid.

The presence of a conserved site in all three genomes increases the confidence that a predicted site is valid, as in the case of the *miR-7* sites in *hairy* and *Tom*. Also, *dpld*, the *Drosophila* homolog of the *let-7* target *lin-41*, ranks second among *Drosophila let-7* targets when conservation in *Anopheles* was required. A number of other target predictions that meet these requirements look quite promising. We have high confidence that the cluster of enzymes in the branched chain amino-acid degradation pathway will prove to be valid *miR-277* targets. Another promising candidate is the predicted *miR-9a* target *Lyra*. *Lyra* contains two predicted *miR-9a* sites. The best *Lyra* site ranks first among all predicted *miR-9a* targets that are conserved in *Anopheles*. Intriguingly, mutations affecting the *Lyra* 3′ UTR lead to a dominant phenotype and to increased Lyra protein levels, an observation that strongly suggests that *Lyra* is subject to translational regulation. *miR-9* is an excellent candidate to mediate this regulation. Many other miRNA–target pairs are identified with sites of a similar quality to those mentioned here (examples include four conserved sites for *miR-309* in *Ets65a* at $Z \geq 2$). As a cautionary note, we wish to emphasize that conservation of target sites in *Anopheles*, while compelling, should not be taken as sufficient evidence of function without experimental validation.

Although it is more difficult to distinguish functional sites from false positives in the cases where only two genomes are compared, we have made use of the clustering of related genes to identify real targets. *reaper*, *grim*, and *sickle* have been validated as *miR-2* targets. We note that the *Netrin* receptor *unc-5* and *Netrin-A* rank second and fourth among predicted *miR-288* targets. We also noted an abundance of transcription factors among the predicted targets of *miR-9*, *miR-279*, and *miR-286* for which orthologous UTRs were not identified in *Anopheles*. These predictions merit further study.

Single versus Multiple Sites

Our statistical analysis shows that the very best single predicted target sites are not statistically significant, even though we have used a reduced database consisting of conserved 3′ UTR sequences. This means that prediction of any single target site cannot be taken as evidence for regulation of a transcript by a miRNA without experimental validation. Sites that are not statistically significant alone can be significant when combined. For example, although none of the *bantam* sites are significant individually, their combined scores are highly significant and supported by experimental validation. 3′ UTRs with multiple predicted target sites are likely to be valid targets for regulation by the miRNA, particularly if their best single sites also rank high in the lists of predicted targets.

Despite the advantages conferred by multiple sites, single miRNA target sites can mediate regulation in vivo. The *C. elegans lin-4* miRNA appears to regulate its target *lin-28* through a single site (Moss et al. 1997). We have presented evidence that *miR-2* family miRNAs can regulate expression of transgenes containing the 3′ UTRs of *reaper* and *grim*, which have one predicted target site, as well as the *sickle* 3′ UTR, which has two predicted sites. Similarly, *miR-7* can regulate expression of transgenes containing the *HLHm3*, *m4*, and *hairy* 3′ UTRs, which have one predicted target site. Further work will be needed to gain insight into what makes some single sites functional and others not. One possibility is that a single site for one miRNA might function in conjunction with independent target sites for other miRNAs in the same UTR. Indeed, a survey of our lists of target predictions indicates that many 3′ UTRs are predicted to contain binding sites for more than one miRNA.

Materials and Methods

Conserved 3′ UTR database. *D. melanogaster* 3′ UTRs were obtained from the Berkeley Drosophila Genome Project (www.fruitfly.org/annot/release3.html) and those of >50 nt were selected. Duplicate UTRs from different splice variants of the same transcript were removed. For each of the resulting 10,196 nonredundant 3′ UTRs, we mapped the last 50 amino acids of the corresponding open reading frame to the *D. pseudoobscura* genome sequence with TBLASTN ($E \leq 10^{-5}$; hgsc.bcm.tmc.edu/projects/Drosophila). We selected UTR matches that included the last 10 residues and had a sequence identity ≥80% or $E \leq 10^{-10}$ and compared these UTRs to the 3,000 nt downstream of the putative *D. pseudoobscura* ortholog with BLASTN (word size = 7; $E \leq 10{,}000$, assuming a database the size of the whole *D. pseudoobscura* genome). Nonconserved nucleotides or those outside the matched regions were replaced by *N*s in the *D. melanogaster* 3′ UTR database to produce the conserved 3′ UTR database. The *D. pseudoobscura* genome has not been fully assembled. This means that some *D. pseudoobscura* genes are located close enough to the end of a contig that the UTR sequences may be missed. 386 *D. melanogaster* genes mapped to the *D. pseudoobscura* genome less than 1 kb from a contig end; 189 mapped less than 500 nt from a contig end. UTR conservation may be underestimated for these genes. For 3,564 genes, we did not detect a suitable ortholog using this protocol. Of these, 571 are known genes; the others are predicted genes about which little is known. For the 4,662 *D. melanogaster* genes lacking annotated UTRs, we assumed 3′ UTRs of 2 kb after of the stop codon and built a separate database of predicted UTRs. The search for *Anopheles* orthologs was done using TBLASTN for the last 50 amino acids of each *D. melanogaster* ORF. Owing to the more extensive sequence divergence, a lower cutoff threshold was allowed ($E < 0.05$) if the last exon of the predicted ORF mapped to the same location (±1 kb) in the annotated genome as the orthologous gene (Zdobnov et al. 2002). If not, the cutoff was $E \leq 10^{-5}$, as for *D. pseudoobscura*. The second, more stringent step of comparing the last 10 amino acids was omitted.

Sequence search. HMMer (Eddy 1996) profiles were constructed for each of two alignments per miRNA containing copies of the reverse complement of the first (5′) 8 nt of the miRNA. The first alignment contained five copies of the exact complement; the second had an additional five copies with C replaced by T and A replaced by G to allow for G:U mismatches. We searched the conserved 3′ UTR database with both profiles and lenient domain bit score threshold (domT ≥ 3) and combined the results. Sequence matches were extended to miRNA length plus 5 nt, the hairpin loop and miRNA sequence were added, and the sequence was evaluated using Mfold. For *Anopheles*, predicted UTRs were searched for the presence of residues 2–7 of the predicted target site. The target sequences were extended and evaluated using Mfold. Only the best site in the *Anopheles* UTR was reported.

Statistics. For each miRNA, we calculated the mean and SD of a background distribution, i.e., the Mfold free energy ΔG of 10,000 randomly selected sequences from the conserved UTR database with lengths of miRNA plus 5 nt. For each prediction, we calculated the *Z* score as the number of SDs above the mean. To compute *E*-values, we fit an exponential function to the cumulative background distributions extrapolated it to give a value for any observed energy and database size. *E*-values are not restricted to normal distributions and scale with database size, so different searches can be compared.

Constructs. The *hairy*, *HLHm3*, *reaper*, *grim*, and *sickle* 3′ UTRs were amplified by polymerase chain reaction from genomic DNA and cloned into tubulin–EGFP as described (Brennecke et al. 2003). *m4* lacks an annotation for its 3′ UTR. A predicted UTR consisting of 900 nt was used. The *miR-7* hairpin and the *miR-2b* hairpin from the *spitz* intron were cloned downstream of DSred2 (Clontech, Palo Alto, California, United States) in pUAST.

Antibodies. Rabbit anti-GFP (TP401) was purchased from Torrey Pines Biolabs (Houston, Texas, United States). Mouse anti-tubulin (T-9026) was purchased from Sigma (St. Louis, Missouri, United States). Mouse anti-Cut is described in Blochlinger et al. (1993).

Supporting Information

Dataset S1. Complete Lists
These contain all predicted targets for each miRNA. There is one file per miRNA. No filtering has been done. From left, the columns are as follows:
gene: FlyBase identifier.
name: if there is one, in addition to the identifier.
GO term: gene ontology information about the gene product.
Z (me): score for the *D. melanogaster* site.
dG(me): folding energy for the *D. melanogaster* site.
Alignment (me): of the miRNA to the target site; asterisk, conventional basepair; plus sign, G:U basepair; minus sign, mismatch.
start and *stop*: position of the site in the 3′ UTR.
mfold (me): the target site, linker, and miRNA sequence as submitted to Mfold.
#sites: total number of sites for the miRNA found in that UTR.
Z(UTR): sum of *Z* for all sites $Z \geq 3$ in that UTR.
Z(max): the best *Z* score for the UTR.
#Z>3: total number of sites of $Z \geq 3$ found in that UTR.
UTR: whether the site is an experimentally defined or predicted UTR.
5′ conservation, *5′ helix*, *CDS+*, *CDS−*, and *mispairing* flags are described in text and in Table 2.
Z(ps): *Z* score for the corresponding site from *D. pseudoobscura*. It is not always possible to unambiguously identify the corresponding site from *D. pseudoobscura*, even though the sequence is in the database (because the genome is not assembled and some regions appear two or more times). We chose to omit ambiguous cases. This problem will disappear when the *D. pseudoobscura* genome sequence is assembled so that unambiguous gene assignment is possible.
dG(ps): folding energy for the site in the *D. pseudoobscura* UTR. The folding energy can be used to determine a *Z* score on the same scale as the site from *D. melanogaster*. This allows a direct comparison of site quality.
Alignment (ps): as above, for the *D. pseudoobscura* site.
Mfold (ps): as above, for the *D. pseudoobscura* site.
Ano-tblastn: TBLASTN score for the *Anopheles* ortholog (explained in the text).
Z(ano): as above, for *Anopheles*.
dG(ano): as above, for *Anopheles*.
Aln(ano): as above, for *Anopheles*.
Mfold(ano): as above, for *Anopheles*.
DOI: 10.1371/journal.pbio.0000060.sd001

Dataset S2. Top 100 List
This dataset presents our recommended filtering of the lists. Only sites in experimentally validated UTRs were included. The 5′ conservation, 5′ helix, CDS+, and CDS− flags were used to discard sites that we consider less likely to be valid. This allows use of a lower *Z* score cutoff ($Z \geq 2$). Sites of $Z < 2$ were removed and Z_{UTR} includes all sites $Z \geq 2$. Columns are as above, except the flags have been removed.
DOI: 10.1371/journal.pbio.0000060.sd002

Accession Numbers
The accession numbers for the miRNAs discussed in this paper are *bantam* (AJ550546, Rfam MI0000387), *let-7* (NR__000938), *lin-4* (NR__000799), *miR-2a* (RF00047, AJ421757), *miR-4* (AJ421762), *miR-7* (AJ421767), *miR-9* (AJ421769), *miR-11* (AJ421771), *miR-13a* (AJ421773), *miR-14* (AJ42177), and *miR-277* (Rfam MI0000360).

The accession numbers for the target genes discussed in this paper are *CG1140* (NM__167928), *CG1673* (NM__132656), *CG5599* (NM__132772), *CG8199* (NM__141648), *CG15093* (NM__166306), *CG17896* (NM__130489), *dpld* (NM__080033), *Drice* (NM__079827), *grim* (NM__079413), *hairy* (NM__079253), *hairy* (*D. simulans*) (AY055843), *hairy* (*T. castaneum*) (AJ457831), *hid* (NM__079412), *HLHm3* (NM__079785), *lin-14* (NM__077516), *lin-28* (NM__059880), *lin-41* (NM__060087), *lin-57* (NM__076575), *Lyra* (NM__080079), *m4* (NM__079786), *reaper* (NM__079414), *scu* (NM__078672), *sickle* (AF460844), and *Tom* (NM__079349).

Acknowledgments

We thank Ann-Mari Voie for preparing transgenic strains.

Conflicts of Interest. The authors have declared that no conflicts of interest exist.

Author Contributions. AS, JB, RBR, and SMC conceived and designed the experiments. AS and JB performed the experiments. AS, JB, RBR, and SMC analyzed the data. AS, JB, RBR, and SMC wrote the paper. ∎

References

Abrahante JE, Daul AL, Li M, Volk ML, Tennessen JM, et al. (2003) The *Caenorhabditis elegans hunchback*-like gene *lin-57/hbl-1* controls developmental time and is regulated by microRNAs. Dev Cell 4: 625–637.

Ambros V (2001) MicroRNAs: Tiny regulators with great potential. Cell 107: 823–826.

Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D (2003) MicroRNAs and other tiny endogenous RNAs in *C. elegans*. Curr Biol 13: 807–818.

Aravin AA, Lagos-Quintana M, Yalcin A, Zavolan M, Marks D, et al. (2003) The small RNA profile during *Drosophila melanogaster* development. Dev Cell 5: 337–350.

Baonza A, Garcia-Bellido A (1999) *Notch* signaling directly controls cell proliferation in the *Drosophila* wing disc. Proc Natl Acad Sci U S A 97: 2609–2614.

Blochlinger K, Jan LY, Jan YN (1993) Post-embryonic patterns of expression of *cut*, a locus regulating sensory organ identity in *Drosophila*. Development 117: 441–450.

Brennecke J, Hipfner DR, Stark A, Russell RB, Cohen SM (2003) *bantam* encodes a developmentally regulated microRNA that controls cell proliferation and regulates the pro-apoptotic gene *hid* in *Drosophila*. Cell 113: 25–36.

de Celis JF, Garcia Bellido A (1994) Roles of the *Notch* gene in *Drosophila* wing morphogenesis. Mech Dev 46: 109–122.

de Celis JF, de Celis J, Ligoxygakis P, Preiss A, Delidakis C, et al. (1996) Functional relationships between *Notch*, *Su(H)* and the bHLH genes of the *E(spl)* complex: The *E(spl)* genes mediate only a subset of *Notch* activities during imaginal development. Development 122: 2719–2728.

Diaz-Benjumea FJ, Cohen SM (1995) Serrate signals through Notch to establish a Wingless-dependent organizer at the dorsal/ventral compartment boundary of the *Drosophila* wing. Development 121: 4215–4225.

Doench JG, Petersen CP, Sharp PA (2003) siRNAs can function as miRNAs. Genes Dev 17: 438–442.

Eddy SR (1996) Hidden Markov models. Curr Opin Struct Biol 6: 361–365.

Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, et al. (2003) Computational and experimental identification of *C. elegans* microRNAs. Mol Cell 11: 1253–1263.

Ha I, Wightman B, Ruvkun G (1996) A bulged *lin-4/lin-14* RNA duplex is sufficient for *Caenorhabditis elegans lin-14* temporal gradient formation. Genes Dev 10: 3041–3050.

Hutvagner G, Zamore PD (2002) A microRNA in a multiple-turnover RNAi enzyme complex. Science 297: 2056–2060.

Kasschau KD, Xie Z, Allen E, Llave C, Chapman EJ, et al. (2003) P1/HC-Pro, a viral suppressor of RNA silencing, interferes with *Arabidopsis* development and miRNA function. Dev Cell 4: 205–217.

Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T (2001) Identification of novel genes coding for small expressed RNAs. Science 294: 853–858.

Lagos-Quintana M, Rauhut R, Yalcin A, Meyer J, Lendeckel W, et al. (2002) Identification of tissue-specific microRNAs from mouse. Curr Biol 12: 735–739.

Lai EC (2002) MicroRNAs are complementary to 3′ UTR sequence motifs that mediate negative posttranscriptional regulation. Nat Genet 30: 363–364.

Lai EC, Posakony JW (1997) The *Bearded* box, a novel 3′ UTR sequence motif, mediates negative posttranscriptional regulation of *Bearded* and *Enhancer of split* complex gene expression. Development 124: 4847–4856.

Lai EC, Posakony JW (1998) Regulation of *Drosophila* neurogenesis by RNA:RNA duplexes? Cell 93: 1103–1104.

Lai EC, Burks C, Posakony JW (1998) The K box, a conserved 3′ UTR sequence motif, negatively regulates accumulation of enhancer of split complex transcripts. Development 125: 4077–4088.

Lai EC, Tomancak P, Williams RW, Rubin GM (2003) Computational identification of *Drosophila* microRNA genes. Genome Biol 4: R42.

Lau NC, Lim LP, Weinstein EG, Bartel DP (2001) An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. Science 294: 858–862.

Lee RC, Ambros V (2001) An extensive class of small RNAs in *Caenorhabditis elegans*. Science 294: 862–864.

Lee RC, Feinbaum RL, Ambros V (1993) The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. Cell 75: 843–854.

Ligoxygakis P, Bray SJ, Apidianakis Y, Delidakis C (1999) Ectopic expression of individual *E(spl)* genes has differential effects on different cell fate decisions and underscores the biphasic requirement for notch activity in wing margin establishment in *Drosophila*. Development 126: 2205–2214.

Lim LP, Glasner ME, Yekta S, Burge CB, Bartel DP (2003a) Vertebrate microRNA genes. Science 299: 1540.

Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, et al. (2003b) The microRNAs of *Caenorhabditis elegans*. Genes Dev 17: 991–1008.

Lin SY, Johnson SM, Abraham M, Vella MC, Pasquinelli A, et al. (2003) The *C. elegans hunchback* homolog, *hbl-1*, controls temporal patterning and is a probable microRNA target. Dev Cell 4: 639–650.

Llave C, Xie Z, Kasschau KD, Carrington JC (2002) Cleavage of *scarecrow*-like mRNA targets directed by a class of *Arabidopsis* miRNA. Science 297: 2053–2056.

Martinez J, Patkaniowska A, Urlaub H, Luhrmann R, Tuschl T (2002) Single-stranded antisense siRNAs guide target RNA cleavage in RNAi. Cell 110: 563–574.

Mathews DH, Sabina J, Zuker M, Turner DH (1999) Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. J Mol Biol 288: 911–940.

Micchelli CA, Rulifson EJ, Blair SS (1997) The function and regulation of *cut* expression on the wing margin of *Drosophila*: Notch, Wingless and a dominant negative role for Delta and Serrate. Development 124: 1485–1495.

Moss EG, Tang L (2003) Conservation of the heterochronic regulator Lin-28, its developmental expression and microRNA complementary sites. Dev Biol 258: 432–442.

Moss EG, Lee RC, Ambros V (1997) The cold shock domain protein LIN-28 controls developmental timing in *C. elegans* and is regulated by the *lin-4* RNA. Cell 88: 637–646.

Mourelatos Z, Dostie J, Paushkin S, Sharma A, Charroux B, et al. (2002) miRNPs: A novel class of ribonucleoproteins containing numerous microRNAs. Genes Dev 16: 720–728.

Olsen PH, Ambros V (1999) The *lin-4* regulatory RNA controls developmental timing in *Caenorhabditis elegans* by blocking LIN-14 protein synthesis after the initiation of translation. Dev Biol 216: 671–680.

Palatnik JF, Allen E, Wu X, Schommer C, Schwab R, et al. (2003) Control of leaf morphogenesis by microRNAs. Nature 425: 257–263.

Pasquinelli AE, Reinhart BJ, Slack F, Martindale MQ, Kuroda MI, et al. (2000) Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. Nature 408: 86–89.

Reinhart BJ, Slack FJ, Basson M, Pasquinelli AE, Bettinger JC, et al. (2000) The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. Nature 403: 901–906.

Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP (2002) MicroRNAs in plants. Genes Dev 16: 1616–1626.

Rhoades MW, Reinhart BJ, Lim LP, Burge CB, Bartel B, et al. (2002) Prediction of plant microRNA targets. Cell 110: 513–520.

Rulifson EJ, Blair SS (1995) Notch regulates *wingless* expression and is not required for reception of the paracrine wingless signal during wing margin neurogenesis in *Drosophila*. Development 121: 2813–2824.

Seggerson K, Tang L, Moss EG (2002) Two genetic circuits repress the *Caenorhabditis elegans* heterochronic gene *lin-28* after translation initiation. Dev Biol 243: 215–225.

Slack FJ, Basson M, Liu Z, Ambros V, Horvitz HR, et al. (2000) The *lin-41* RBCC gene acts in the *C. elegans* heterochronic pathway between the *let-7* regulatory RNA and the LIN-29 transcription factor. Mol Cell 5: 659–669.

Tang G, Reinhart BJ, Bartel DP, Zamore PD (2003) A biochemical framework for RNA silencing in plants. Genes Dev 17: 49–63.

Wightman B, Ha I, Ruvkun G (1993) Posttranscriptional regulation of the heterochronic gene *lin-14* by *lin-4* mediates temporal pattern formation in *C. elegans*. Cell 75: 855–862.

Xie Z, Kasschau KD, Carrington JC (2003) Negative feedback regulation of Dicer-Like1 in *Arabidopsis* by microRNA-guided mRNA degradation. Curr Biol 13: 784–789.

Xu P, Vernooy SY, Guo M, Hay BA (2003) The *Drosophila* microRNA *miR-14* suppresses cell death and is required for normal fat metabolism. Curr Biol 13: 790–795.

Zdobnov EM, von Mering C, Letunic I, Torrents D, Suyama M, et al. (2002) Comparative genome and proteome analysis of *Anopheles gambiae* and *Drosophila melanogaster*. Science 298: 149–159.

Zeng Y, Wagner EJ, Cullen BR (2002) Both natural and designed micro RNAs can inhibit the expression of cognate mRNAs when expressed in human cells. Mol Cell 9: 1327–1333.

Zuker M, Mathews DH, Turner DH (1999) Algorithms and thermodynamics for RNA secondary structure prediction. In: Barciszewski J, Clark BFC, editors. A practical guide in RNA biochemistry and biotechnology. Dordrecht, The Netherlands: Kluwer Academic Publishers. pp. 11–43.