SI Text

Short RNA analysis

Complete description of repeat-overlapping sequences. The genomic distribution of highly repetitive sequences correlated well with total repeat content along chromosomes, and specific high-density clusters of repeat overlapping reads were not observed (SI Fig. 10 *A* and *B*). Of note, the ratio of short RNA associated repeats to total chromosome repeat content in all 4 libraries is highest on chromosome X (SI Fig. 10 *B* and C). This increase is due to a large number of matches to LINE-associated short RNAs, and not an absolute increase in the number of distinct repetitive short RNA sequences on chromosome X (SI Fig. 10 *D* and E).

Because of their large number of corresponding genomic locations, the highly repetitive sequences were not searched comprehensively for prototype structures that could generate novel miRNAs (see *SI Methods*). Nevertheless, the total proportions of highly repetitive sequences compared to all novel reads were similar between the *Dicer*$^{+/+}$ and *Dicer*$^{-/-}$ libraries, indicating that as a class they exist independently of Dicer activity (SI Table 8). Further, these sequences share several descriptive characteristics with the set of repeat-overlapping novel sequences with less than 20 hits to the genome, described below, again indicating they exist as a class of sequences separate from miRNAs.

The repeat overlap in the set of novel sequences with less than 20 hits to the genome was analyzed (referred to as “nonrepetitive repeat-overlapping novel reads” below). Because these sequences had far fewer genomic hits than the highly repetitive sequences, a more in depth analysis of their relationship to surrounding genomic sequence was feasible. Foremost, these novel reads were comprehensively evaluated as potential miRNAs, and their surrounding genomic sequences are devoid of clear hairpin structures. Similar to the set of highly repetitive sequences, the proportions of SINE and Simple repeat overlap in this set are reduced in the *Dicer*$^{-/-}$ library as compared to the *Dicer*$^{+/+}$ library, further supporting the idea that a subset of repeat-associated short RNAs depend on Dicer for biogenesis (SI Fig. 11*A*). The nonrepetitive, repeat-overlapping novel sequences showed no clear strand bias with respect to overlapping repetitive elements, and had the same distribution of length and first nucleotides as the novel sequences that did not overlap repeats (SI Fig. 11 *A* and *B*, data not shown). As expected, the nonrepetitive, repeat-overlapping novel sequences were more frequently complementary to intergenic regions, and significantly less conserved when compared to the novel reads that did not overlap repeats (SI Fig. 11 *C* and *D*). Consistent with the broad chromosomal distribution of the set of highly repetitive novel sequences, there is no significant clustering of the repeat overlapping reads with less than 20 hits to the genome (data not shown).

No evidence for piRNAs in mouse ES cells. The recent description of piRNAs in the mouse, rat, and human testis prompted us to examine if ES cells expressed similar short RNAs (1). There was no evidence for a distinct class of 29-31 nt piRNA-like species in any of the four cDNA libraries. There were, however, 51 distinct sequences, represented by 112 reads, which uniquely overlapped 14 known piRNA clusters (2). 30 of these sequences, represented by 82 reads, fell into one piRNA cluster and were generated by a group of known and novel miRNAs that was identified on chromosome X (SI Table 7). Accordingly, the length distribution and first nucleotide bias of the reads falling into piRNA clusters were miRNA-like, with a major peak of reads surrounding 22 nt, and 60% of sequences beginning with a ‘U’ (SI Fig. 12 *A* and *B*).

No evidence for *C.elegans*-like siRNAs in mouse ES cells. The possibility that ES cells express endogenous siRNAs similar to those observed in *C.elegans* was examined (3-6). Such siRNAs are antisense to protein coding genes, and do not typically have the 5' mono-phosphates that were selected for in this study. Nevertheless, an analysis of 5' mono-phosphate-containing short RNAs from *C. elegans* did identify two distinct siRNA populations that had a strong ‘G’ first nucleotide bias and peaked at 22 and 26 nt, respectively (5). There was no evidence that these species exist in ES cells. In the four libraries analyzed here, there were 190 distinct sequences, represented by 261 reads, that were uniquely anti-sense to known protein coding exons and could be considered potential analogues of *C. elegans* siRNAs. However, unlike *C. elegans* siRNAs, these sequences had no distinct length or first nucleotide bias when compared to the set of 1319 distinct sequences (1,500 reads) that were uniquely sense to known protein coding genes (SI Fig. 12 *C* and *D*). This apparent absence of analogous siRNA species in ES cells is not entirely surprising considering that mammals do not have RdRPs homologous to those required for siRNA production in *C. elegans*. Anti-sense exon-overlapping short RNAs were present in all four libraries (SI Fig. 12*E*). The low abundance of these species in our libraries suggests that they have limited physiological roles; however, similar, more abundant molecules may be 5' end modified such that they were excluded in the libraries analyzed here.

SI Methods

Generation and characterization of Dicer$^{-/-}$ ES cells. All ES cells were cultured and transfected as described in ref. 7.

Dicer$^{+/+}$ ES cells were derived from mice homozygous for the floxed *Dicer* allele described in ref. 8, and floxed GFP described in ref. 9. To generate clonal *Dicer*$^{-/-}$ cell lines, Cre recombinase was transiently transfected into *Dicer*$^{+/+}$ ES cells. 24 h posttransfection cells were plated at clonal density onto feeder layers and individual GFP negative colonies were selected and cultured until growth recovered, then expanded, removed from feeder layers, and used for subsequent analysis. *Dicer* genotyping oligos, from 5' to 3' as illustrated in SI Fig. 4, are as follows: (1) 5'-CATGACTCTTCAACTCAAACT-3'; (2) 5'-CCTGACAGTGACGGTCCAAAG-3'; (3) 5'-AGCATGGGGGCACCCTGGTCCTGG-3'. Sex determination of ES cells was determined as described in ref. 10. Dicer antibody 1416 was from ref. 11. 5 μg of DNA was used for the Southern blot in SI Fig. 4. The minor satellite probe was described in ref. 12. The mitochondrial DNA probe and *DNMT1* null ES cells were gifts from R. Jaenisch. LINE L1 and SINE B1 probes were PCR amplified by using primers from ref. 12. The SINE B1 Northern blot was performed as in ref. 7.

J1aza treatment and analysis. J1 ES cells were treated with 30 µM 5-aza-2'-deoxycytidine, 5-aza-dC (Sigma) dissolved in DMSO, or DMSO only, for 24 h. DNA and RNA samples were collected for each sample approximately every two days for a total of 2 weeks. To determine the percentage cellular DNA methylation HPLC analysis was used as described in ref. 13. Samples were loaded onto a Vydac 218TP52 reverse phase C18 HPLC column and a 60 min isocratic run in 50 mM Ammonium phosphate dibasic buffer pH 4.1 was used for separation. As a control, dTMP, dAMP, dCMP, dGMP, and 5mCMP (Sigma and Reliable Biopharmaceutical Corporation) were mixed equally and eluted as above. MuERV-L primers are from ref. 14.

Calculation of mature miRNA processing variability. To calculate miRNA processing variability, the number of miRNA reads matching each annotated 5'/3' miRNA end was divided by the total number of reads overlapping the arm of the hairpin on which the mature miRNA was located. If previously unannotated, 5p and 3p miRNAs were assigned from miRNA* strands if the miRNA* species represented ≥20% of all reads originating from the hairpin. For miRNAs mapping to multiple hairpin precursors, only the sequences mapping to the hairpin with the most aligning reads were evaluated for miRNA processing variability.

Novel miRNA annotation. Only sequences with <20 matches to the genome were evaluated as potential miRNAs, with the exception of the 8 sequences with ≥20 hits to the genome that were sequenced ≥3x in the *Dicer*$^{+/+}$ library and absent in the *Dicer*$^{-/-}$ library. For each short RNA genomic location, two potential miRNA hairpins were defined: one encompassing 20 and 80 nt of sequence around the 5' and 3' ends of the short RNA, respectively, and the other encompassing 80 and 20 nt of sequence around the 5' and 3' ends of the short RNA, respectively. Potential miRNAs were evaluated for their ability to form hairpins that had secondary structures consistent with *Drosha* and *Dicer* processing (15-17). Those hairpins whose RNAfold output (18) exhibited base pairing over at least 70% of the length of the potential miRNA, base pairing over or adjacent to the processed ends of the putative premiRNA, symmetrical bulges, and double-strandedness existing approximately one helical turn past the *Drosha*-processed end of the miRNA–between 7 and 17 nt after the end–were annotated as novel miRNAs. Additionally, in order for a hairpin to be annotated as a novel miRNA, we required that there existed a sequence comprising the majority of all reads aligning to the novel hairpin (a dominant miRNA), and that this dominant miRNA was sequenced more than once and between 20 and 25 nt long. Requiring a dominant miRNA allowed for unambiguous assignment of the novel miRNA's seed.

miRNA-like hairpins with aligned reads that did not produce a dominant miRNA of the specified length, that had only one aligned read, or that had a predicted secondary structure that did not completely satisfy our requirements, were deemed miRNA candidates and not included in the final set of novel miRNAs (*SI Text*). This set of candidate miRNA hairpins had secondary structures and expression characteristics similar enough to known miRNAs that we believe short RNAs mapping to these loci were likely generated by the miRNA-processing pathway; however, the putative hairpins either lacked sufficient expression for confident annotation, or had characteristics that differed slightly from known miRNAs.

From the set of eight sequences with ≥20 hits to the genome that were sequenced ≥3x in the *Dicer*$^{+/+}$ library and absent in the *Dicer*$^{-/-}$ library, three novel miRNA hairpins were identified. The four sequences aligning to these three hairpins had 120, 306, 379, and 1416 hits to the genome, respectively.

Repeat analysis. Repeat overlap was determined by using the Repeatmasker track of the UCSC table browser (19). For sequences with ≥20 hits to the genome, repeat-identity was assigned to the repeat and class that overlapped most frequently with each short RNA sequence. For those sequences with ≥500 hits to the genome, 250 hits were randomly selected 5 times, and specific repeat and class was assigned if the majority of the 250 hits overlapped the same repeat and class all 5 times. Otherwise the specific repeat and class were designated "can't distinguish". For sequences with <20 hits to the genome, the specific repeat and class were annotated for each genomic hit and normalized to the number of genomic hits for each sequence.

Conservation and motif analysis. Four-way mammalian alignments (mm7, hg17, canFam2 and rn3) were extracted from the UCSC genome browser's 17-way mammalian alignments for a given region of length L. The conservation score for a region was determined by $(\sum_{i=1:L,} F_i)/L$, where F_i is the number of bases in the other species (hg17, canFam2 or rn3) that is identical to that of mm7 at position i, divided by the number of aligned species (maximum of 3) at position i. Information content, I_M of motif M of length N was computed as $\sum_{i=1:N, \forall j} f_{i,j} \log_2(f_{i,j}/g_j)$, where $f_{i,j}$ was the frequency of nucleotide j at position i, and g_j was the background frequency of nucleotide j. Nucleotide background frequencies (g_A, g_T, g_G, g_C) were defined as (0.3, 0.3, 0.2, 0.2). For small sets of sequences, the algorithm MEME was utilized to identify motifs with a minimum width of 6, and an e-value cutoff of 0.001 (20).

1. O'Donnell KA, Boeke JD (2007) *Cell* 129:37-44.

2. Lau NC, Seto AG, Kim J, Kuramochi-Miyagawa S, Nakano T, Bartel DP, Kingston RE (2006) *Science* 313:363-367.

3. Ambros V, Lee RC, Lavanway A, Williams PT, Jewell D (2003) *Curr Biol* 13:807-818.

4. Pak J, Fire A (2007) *Science* 315:241-244.

5. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP (2006) *Cell* 127:1193-1207.

6. Sijen T, Steiner FA, Thijssen KL, Plasterk RH (2007) *Science* 315:244-247.

7. Calabrese JM, Sharp PA (2006) *Rna* 12:2092-2102.

8. Harfe BD, McManus MT, Mansfield JH, Hornstein E, Tabin CJ (2005) *Proc Natl Acad Sci USA* 102:10898-903.

9. Ventura A, Meissner A, Dillon CP, McManus M, Sharp PA, Van Parijs L, Jaenisch R, Jacks T (2004) *Proc Natl Acad Sci USA* 101:10380-5.

10. Conner DA (2000) *Curr Protocol Mol Biol* 23.4.1.

11. Kanellopoulou C, Muljo SA, Kung AL, Ganesan S, Drapkin R, Jenuwein T, Livingston DM, Rajewsky K (2005) *Genes Dev* 19:489-501.

12. Martens JH, O'Sullivan RJ, Braunschweig U, Opravil S, Radolf M, Steinlein P, Jenuwein T (2005) *EMBO J* 24:800-812.

13. Ramsahoye BH (2002) *Methods* 27:156-161.

14. Peaston AE, Evsikov AV, Graber JH, de Vries WN, Holbrook AE, Solter D, Knowles BB (2004) *Dev Cell* 7:597-606.

15. Ambros V, Bartel B, Bartel DP, Burge CB, Carrington JC, Chen X, Dreyfuss G, Eddy SR, Griffiths-Jones S, Marshall M, *et. al.* (2003) *RNA* 9:277-279.

16. Han J, Lee Y, Yeom KH, Nam JW, Heo I, Rhee JK, Sohn SY, Cho Y, Zhang BT, Kim VN (2006) *Cell* 125:887-901.

17. Zeng Y, Yi R, Cullen BR (2005) *EMBO J* 24:138-148.

18. Hofacker IL, Fontana W, Stadler PF, Bonhoeffer LS, Tacker M, Schuster P (1994) *Monatsh Chem* 125:167-188.

19. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, Kent WJ (2004) *Nucleic Acids Res* 32:D493-6.

20. Bailey TL a. E., C (1994) *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, August*, 28-36.

21. Murchison EP, Partridge JF, Tam OH, Cheloufi S, Hannon GJ (2005) *Proc Natl Acad Sci USA* 102:12135-40.