Supplementary Data

Global analyses of endonucleolytic cleavage in mammals reveal expanded repertoires of cleavage-inducing small RNAs and their targets

Ashley A. Cass, Jae Hoon Bahn, Jae-Hyung Lee, Christopher Greer, Xianzhi Lin, Yong Kim, Yun-Hua Esther Hsiao, Xinshu Xiao

Supplementary Methods

Bioinformatic analysis of Ago2 dependence

Ago2 CLIP-Seq in wild type mESCs (3 samples) were used (accession GSE25310: SRR072951, SRR072952, SRR072953, SRR072954, SRR072955) (1). To evaluate the presence of sciRNAs in *Ago2* CLIP-Seq reads, the total number of CLIP reads containing each sciRNA was summed across the 3 samples. As a control, the presence of *Dicer*-independent, *Dgcr8*-independent small RNAs (2) in *Ago2* CLIP-Seq reads was calculated similarly.

To calculate the enrichment of Deg-Seq peak regions in *Ago2* CLIP-Seq data, the total number of CLIP reads at each nucleotide of the Deg-Seq peak (up to 4 nucleotides) was summed across each of the 3 wild type samples and normalized by the length of the Deg-Seq peak. For each peak with non-zero CLIP coverage, 100 random peaks with the same abundance were chosen by Fisher-Yates shuffle, and normalized CLIP coverage was calculated similarly as described above. Then, the average CLIP-Seq coverage of the 100 random peaks was calculated.

To calculate Deg-Seq abundance of sciRNA targets in wild type mESCs, the total number of Deg-Seq reads at the peak was normalized by the total number of reads in the gene. For each target gene, there is often no Deg-Seq peak present in *Ago2*-/- data, as expected. Thus, to calculate Deg-Seq abundance of targeted genes in *Ago2*-/- Deg-Seq (accession GSE21975 (3)), the maximum number of Deg-Seq reads at any position in the gene was normalized by the total Deg-Seq reads in the gene.

RNA-Seq analysis

The following public mESC RNA-Seq datasets were used: SRR921480, SRR921481, SRR921482, SRR921483 (GSE48252) (4). For both public and in-house RNA-Seq data, sequencing reads were aligned to mm10 using tophat2 (5) with parameters -a 9 -g 1 for all datasets. RPKM was calculated using in-house scripts for genes annotated in the databases described in Materials and Methods.

Pearson correlation analysis: The Pearson correlation was calculated using RNA-Seq and small RNA-Seq data of the following samples: mESCs, testis E14, testis E18, testis 3 weeks post-natal, testis 6 months post-natal. For mESCs, the average RPKM of the above four datasets from GSE48252 was used in the Pearson correlation analysis of target and sciRNA expression. If more than one sciRNA targeted a gene, the minimum Pearson correlation was kept.

sciRNA characterization and expression

sciRNAs were aligned to genome mm10 or hg19 using Bowtie 0.12.7 (6) with parameters -v 1 -a --best --strata. Because many sciRNAs may align to thousands of genomic locations, a stepwise prioritization process was used to annotate them (listed in descending priority): [1] the sciRNA was found to be *Dicer*-dependent and *Dgcr8*-independent (2), [2] the sciRNA overlaps a miRBase (7) miRNA or Rfam (8) pre-miRNA +/- 3nt, [3] the sciRNA overlaps a non-miRNA Rfam annotation +/- 3nt, [4] the sciRNA is a sub-sequence of a piRNABank (9) annotated piRNA, [5] the sciRNA overlaps a region from our custom merged annotation described earlier, [6] no annotation, [7] unmapped. If a category had 3 or fewer members, it was labeled as "other" in Figure 3a.

We next predicted whether a sciRNA was derived from a long hairpin RNA (hpRNA) structure. For each sciRNA, we applied RNAfold (10) to the region flanking its genomic alignment (+/-500nt). If the sciRNA was aligned to multiple genomic locations, the region (+/-500nt) with the highest read coverage was used. Then, RNAfold's dot bracket notation was used to examine whether the sciRNA aligned to the stem of a long secondary structure, i.e. "long hpRNA." Namely, two criteria were used: stem length ≥70, and ≥70% of the sciRNA nucleotides (length 19-24) were structured (brackets). Stem length was calculated as the distance to the next opposite-facing bracket (equal to zero if the sciRNA contained two opposite facing brackets). These thresholds were chosen by checking that previously identified endo-siRNAs (e.g. miR-1195 and miR-1965) in mESCs were included and manually checking the RNAfold structure prediction for some novel examples (e.g. *Ccdc30* shown in Figure 3b). In addition, several thresholds were tested, and although stem length ≥70 and ≥70% structured were the optimal thresholds, the results were largely robust to choice of stem length and percent nucleotides structured.

Small RNA expression was calculated using reads per million (RPM), where sequencing depth for each sample only included small RNAs that passed the preprocessing steps in the pipeline (i.e. masked small repeats and length [19,24]). sciRNAs identified in mESCs and/or testis 6M PN were clustered across all testis developmental stages, and similarly for mESCs and/or cerebellum 6M PN. sciRNAs with RPM ≥ 1 were grouped based on whether they targeted a transcript in mESCs, adult tissue, or both.

miRNAs with minimum read count 20 in at least one sample were clustered since this minimum read count was used in the preprocessing pipeline to identify sciRNAs. miRNAs with RPM ≥ 1 were labeled as adult tissue, mESCs, or both based on read count ≥ 20 in these samples.

R function heatmap.2 was used for hierarchical clustering with Euclidean distance and complete linkage.

Target characterization

Cleavage sites (Deg-Seq peaks) were characterized by genic location and by type of transcript. If the Deg-Seq peak overlapped multiple transcripts, the genic regions were prioritized as follows:

coding exon (CDS exon) > 5'UTR exon > 3'UTR exon > exon in non-coding transcripts (NC exon) > intron in coding genes > 5'UTR intron > 3'UTR intron > intron in non-coding transcripts. Transcripts harboring the cleavage sites were also examined for their types: coding, pseudogene, lncRNA, or other non-coding RNAs. Finally, Repeatmasker was used to decide whether a cleavage site overlaps a repetitive sequence. To test if the observed gene region and Repeatmasker enrichment at cleavage sites significantly differed from the transcriptome overall, 100 random positions per Deg-Seq peak were chosen from any annotated transcript. If a random position overlapped multiple transcripts, the genic regions were prioritized using the above prioritization schemes (Figure 4b, 5b).

Functional analysis of target genes

Gene set enrichment. Ingenuity pathway analysis was conducted for sciRNA targets in the mouse and human data sets (IPA®, QIAGEN Redwood City, www.qiagen.com/ingenuity) using the default parameters. Gene Ontology (GO) analysis was carried out with our previous approach (11) for non-coding genes according to the functional annotations provided by NONCODEv4 (12).

SNP enrichment. The union of all sciRNA binding sites located within B1 regions (i.e. the Alu Repeatmasker family) in predicted target genes of mESC, testis, and cerebellum were used (n = 1,510). Each position in the targeted region and 100 nt flanking region were interrogated for SNPs (dbSNP 138 (13)). To calculate a SNP density per nucleotide, the sum of SNPs at each position was normalized by the total sequences interrogated at that position. These values were smoothed using a sliding window of size 10nt and step size 1nt (Figure 5b). The smoothed values were anchored on the rightmost nucleotide (e.g. the smoothed SNP density at position -100 is the average SNP density of the window -100 to -90). A chi-square p-value was calculated using a contingency table of total SNPs vs. the sum of length of B1 annotation (up to 100 nt) within vs. outside the target region.

Ago2 binding in canonical miRNA targets. Predicted canonical miRNA target sites were downloaded from microrna.org (14). Targets of expressed miRNAs in mESCs (miRNA read count ≥ 20) located within B1 (Alu family) regions were separated into two groups based on whether or not they overlapped a sciRNA target sequence. A total of 102 unique miRNA targets overlapped a sciRNA target, and 54,766 did not. Among the latter, 53,333 (97%) did not contain a Deg-Seq peak, as expected. We refer to these 53,333 targets as "putative canonical miRNA B1 targets" and the 102 unique miRNA targets overlapping a sciRNA target as "putative sciRNA targets". To compare *Ago2* CLIP-Seq overlap between the two groups, one "putative canonical miRNA B1 target" was randomly chosen for each "putative sciRNA target" (i.e. 102 vs. 102), and this process was repeated 1,000 times. *Ago2* CLIP-Seq overlap was calculated as the sum of reads in the target site, divided by the length of target sequence. For each of the 1,001 total target and control sets, targets with zero *Ago2* CLIP-Seq overlap were excluded and the average CLIP-Seq overlap of the remaining targets was calculated. An empirical p-value was calculated by counting the total sets of random "putative canonical miRNA B1 targets" with higher average *Ago2* CLIP-Seq density than the "putative sciRNA targets."

In Vitro Transcription

sciRNA targets were amplified using OneTaq DNA polymerase (NEB) from mouse genomic DNA followed by TOPO TA cloning (Life Technologies). The target-specific primers are listed in Supplementary Table S3. The TOPO cloning products were then transformed into DH5α competent cells and were later plated for overnight incubation at 37°C. PCR and Sanger sequencing were used to verify the constructs.

Target RNAs were *in vitro* transcribed using 1 μg of template DNA and the HiScribe™ T7 High Yield RNA Synthesis Kit (NEB). To remove template DNA, 20U RNase-free DNase I (Roche Diagnostics) was applied for 15 min at 37°C followed by phenol extraction. *In vitro* transcribed RNA was purified from 10% PAGE gel. RNA was dephosphorylated by 10U calf intestinal alkaline phosphatase (NEB) at 37°C for 60 min and then purified by phenol-chloroform extraction. Two μg dephosphorylated RNA was labeled with γ-^{32}P ATP 150Ci (MPbio) by T4 Polynucleotide Kinase (NEB) at 37°C for 60 min followed by 12% PAGE gel isolation.

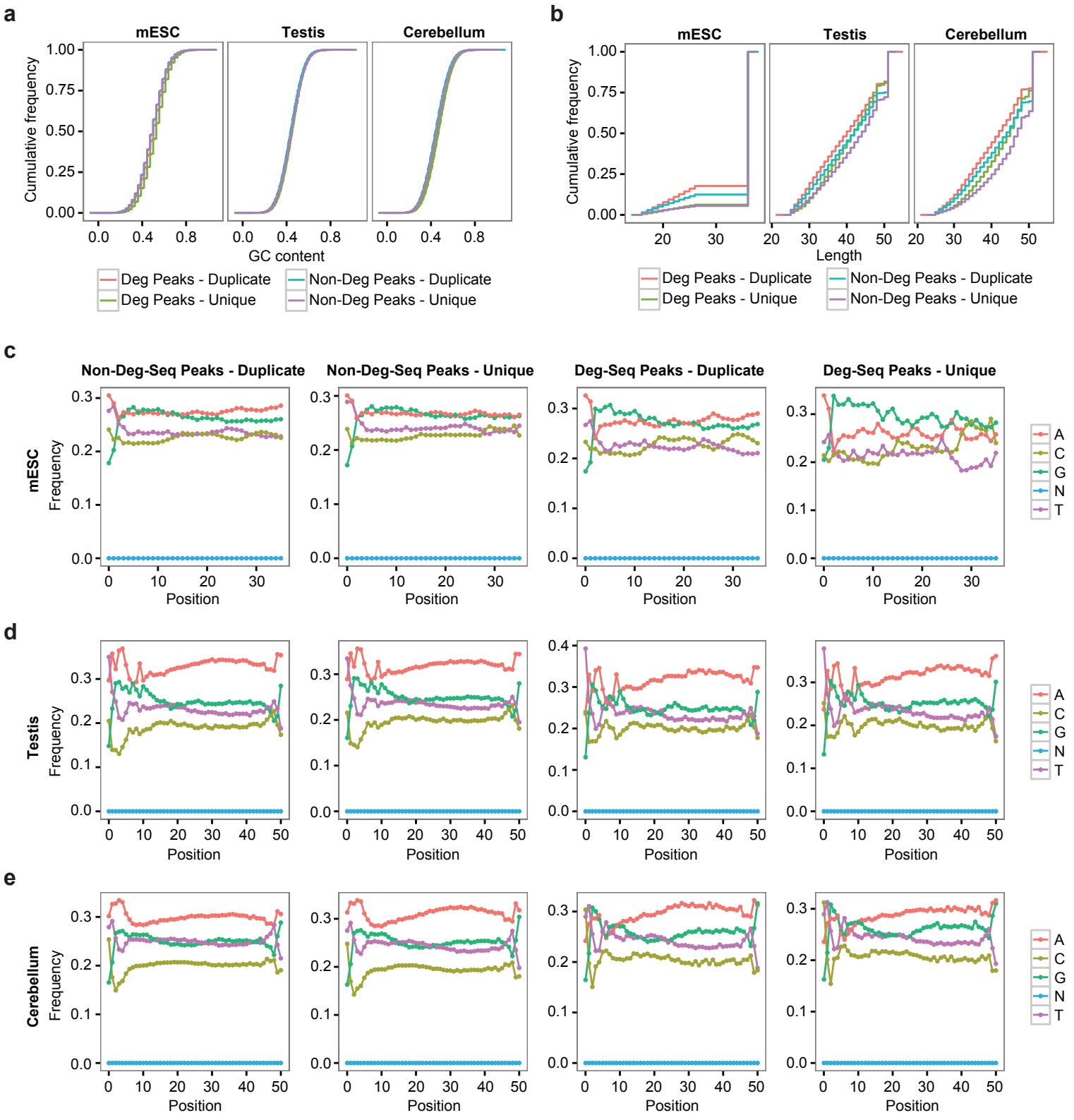
In Vitro Cleavage Assays

In vitro cleavage assays were performed as described previously (15). HeLa cytoplasmic S100 extract was obtained from Speed BioSystems. For endogenous sciRNA cleavage (miR-708-5p and miR-29c-3p), HeLa cytoplasmic S100 extract (0, 0.1, 0.5, and 2 μg respectively) was incubated with 200 ng of ^{32}P-labeled target RNA at 37°C for 30 min in the cleavage buffer (20mM HEPES KOH pH7.9, 100mM KCl, 1.5mM $MgCl_2$, 0.5mM DTT, 0.5mM PMSF, 1mM ATP, 0.2mM GTP). The cleavage reaction was terminated by adding 2X RNA gel loading buffer and incubated at 60°C for 5 min. Cleaved RNA was loaded onto 10-12% PAGE gel and exposed to X-ray film at –80°C.

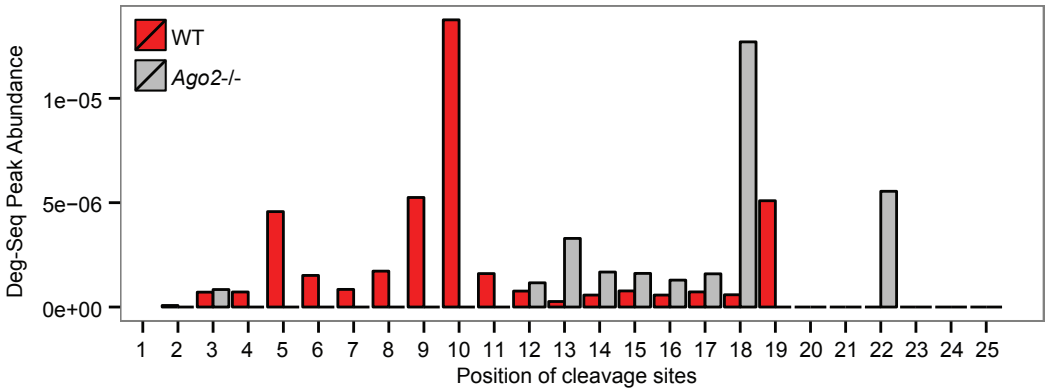
For cleavage of *Traf3ip2*-as and *Zfp389* target genes, endogenous sciRNA levels were relatively low. Thus, sciRNA +/- strands were annealed to form sciRNA duplexes as follows: [1] prepare the +/- strand sciRNAs (Supplementary Table S3) at a final concentration of 100 μM; [2] mix 2 μL of the two sciRNA strands with 5μL 10X Annealing Buffer (100 mM Tris-HCl, pH 7.5, 1 M NaCl, 10 mM EDTA); [3] add nuclease-free H_2O to reach a total volume of 50 μL; [4] heat at 94°C in water bath for 4 min, 70°C for 10 min, and then allow cooling to room temperature; [5] annealed sciRNA duplex was further purified by 10% PAGE gel and precipitated with 2.5 volumes of absolute ethanol. HeLa cytoplasmic S100 extract was then preincubated with 50 nM purified sciRNA duplex at 37°C for 30 min before adding the ^{32}P-labeled target RNA. The cleavage reaction was otherwise carried out in the same way as described above.

References

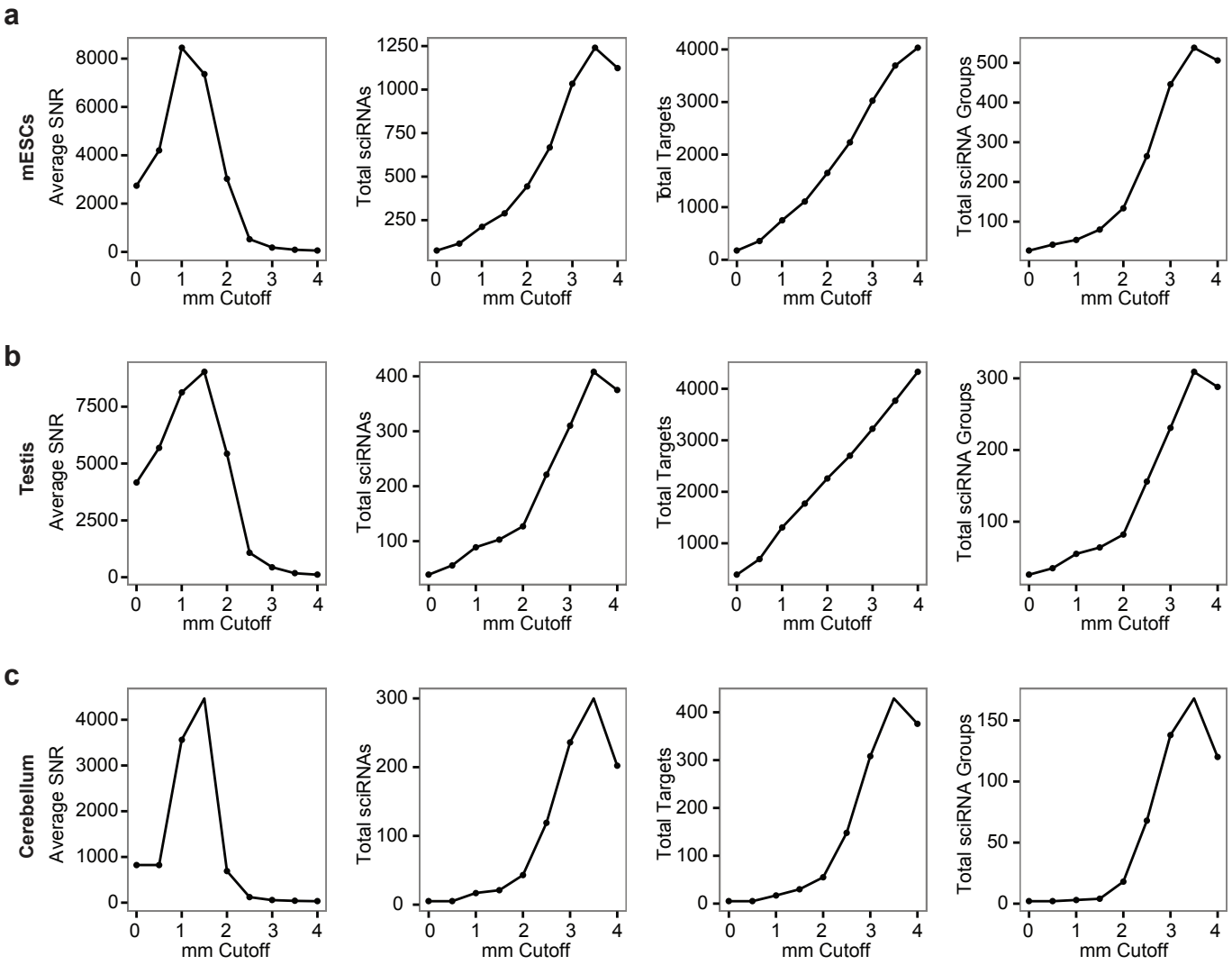
1. Leung,A.K.L., Young,A.G., Bhutkar,A., Zheng,G.X., Bosson,A.D., Nielsen,C.B. and Sharp,P. a (2011) Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nat. Struct. Mol. Biol.*, **18**, 237–44.
2. Babiarz,J.E., Ruby,J.G., Wang,Y., Bartel,D.P. and Blelloch,R. (2008) Mouse ES cells express endogenous shRNAs, siRNAs, and other Microprocessor-independent, Dicer-dependent small RNAs. *Genes Dev.*, **22**, 2773–85.
3. Karginov,F. V, Cheloufi,S., Chong,M.M.W., Stark,A., Smith,A.D. and Hannon,G.J. (2010) Diverse endonucleolytic cleavage sites in the mammalian transcriptome depend upon microRNAs, Drosha, and additional nucleases. *Mol. Cell*, **38**, 781–8.
4. Hou,P., Li,Y., Zhang,X., Liu,C., Guan,J., Li,H., Zhao,T., Ye,J., Yang,W., Liu,K., *et al.* (2013) Pluripotent Stem Cells Induced from Mouse Somatic Cells by Small-Molecule Compounds. *Sci. (New York, NY)*, **341**, 651–654.
5. Kim,D., Pertea,G., Trapnell,C., Pimentel,H., Kelley,R. and Salzberg,S.L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.*, **14**, R36.
6. Langmead,B., Trapnell,C., Pop,M. and Salzberg,S.L. (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.
7. Kozomara,A. and Griffiths-Jones,S. (2014) MiRBase: Annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.*, **42**, 68–73.
8. Burge,S.W., Daub,J., Eberhardt,R., Tate,J., Barquist,L., Nawrocki,E.P., Eddy,S.R., Gardner,P.P. and Bateman,A. (2013) Rfam 11.0: 10 years of RNA families. *Nucleic Acids Res.*, **41**, 1–7.
9. Sai lakshmi,S. and Agrawal,S. (2008) piRNABank: A web resource on classified and clustered Piwi-interacting RNAs. *Nucleic Acids Res.*, **36**, 173–177.
10. Hofacker,I.L., Fontana,W., Stadler,P.F., Bonhoeffer,L.S., Tacker,M. and Schuster,P. (1994) Fast folding and comparison of RNA secondary structures. *Monatshefte fur Chemie Chem. Mon.*, **125**, 167–188.
11. Lee,J.H., Gao,C., Peng,G., Greer,C., Ren,S., Wang,Y. and Xiao,X. (2011) Analysis of transcriptome complexity through RNA sequencing in normal and failing murine hearts. *Circ. Res.*, **109**, 1332–1341.
12. Xie,C., Yuan,J., Li,H., Li,M., Zhao,G., Bu,D., Zhu,W., Wu,W., Chen,R. and Zhao,Y. (2014) NONCODEv4: Exploring the world of long non-coding RNA genes. *Nucleic Acids Res.*, **42**, 1–6.
13. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
14. Betel,D., Wilson,M., Gabow,A., Marks,D.S. and Sander,C. (2008) The microRNA.org resource: targets and expression. *Nucleic Acids Res.*, **36**, D149–53.
15. Shin,C., Nam,J.-W., Farh,K.K.-H., Chiang,H.R., Shkumatava,A. and Bartel,D.P. (2010) Expanding the microRNA targeting code: functional sites with centered pairing. *Mol. Cell*, **38**, 789–802.



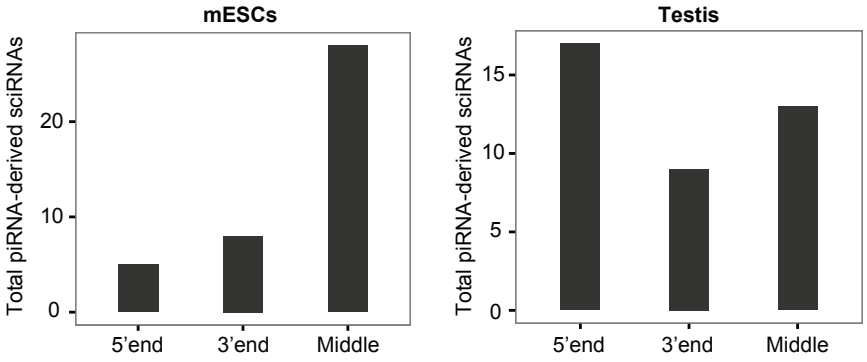
Supplementary Figure S1 | PCR amplification bias detection. All reads in each Deg-Seq library were separated into four groups: unique reads within Deg-Seq peaks, duplicate reads within Deg-Seq peaks, unique reads outside of Deg-Seq peaks, and duplicate reads outside of Deg-Seq peaks. Three criteria were used for comparison: (a) GC content, (b) read length after trimming adapters, and (c-e) nucleotide composition.



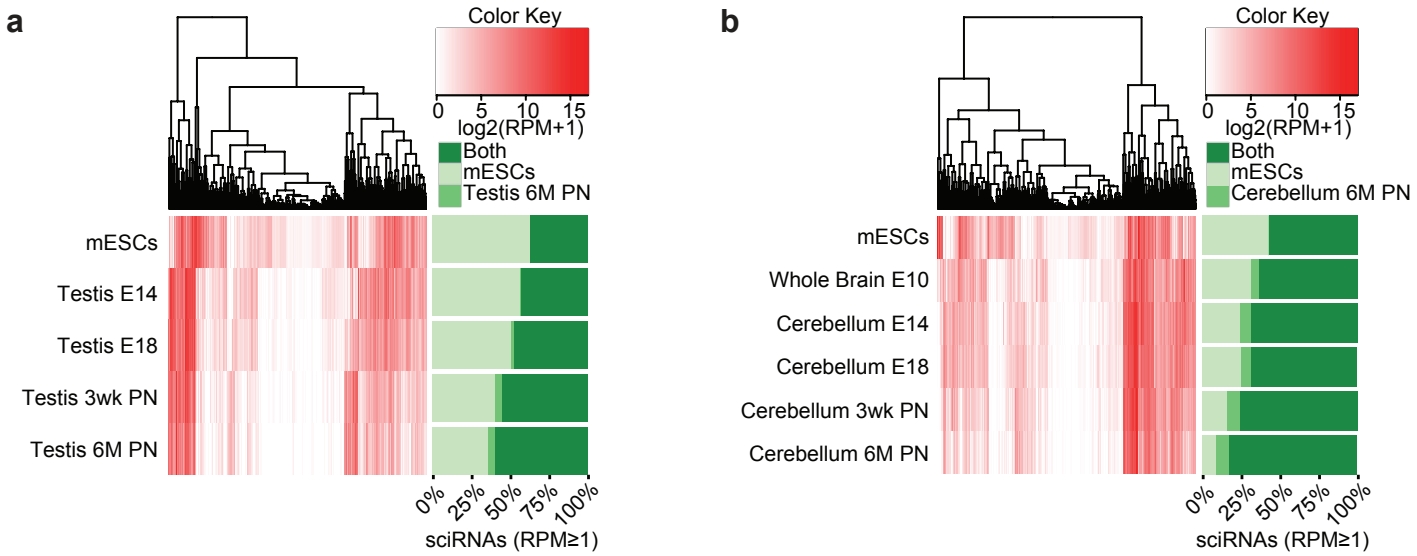
Supplementary Figure S2 | Alignment of predicted endo-siRNAs to Degradome-Seq supports existence of small RNA-guided cleavage. Deg-Seq peak abundance (total reads in the Deg-Seq peak / total mapped reads) per nucleotide along predicted predicted endo-siRNAs (Babiarz et al, 2008) from 5' to 3' (left to right on the x-axis) in wild type (WT, red) and *Ago2* knockout mESCs (*Ago2*-/-, grey). Results are shown for Deg-Seq peaks that aligned to endo-siRNAs with up to 1 mismatch. The enrichment of reads at nt 9-11 in WT is eliminated in *Ago2*-/-. Moreover, the Deg-Seq abundance is within the level of background noise for most positions in *Ago2*-/- with the exception of nt 18 and 22. These could be due to unknown artifacts or mechanisms.



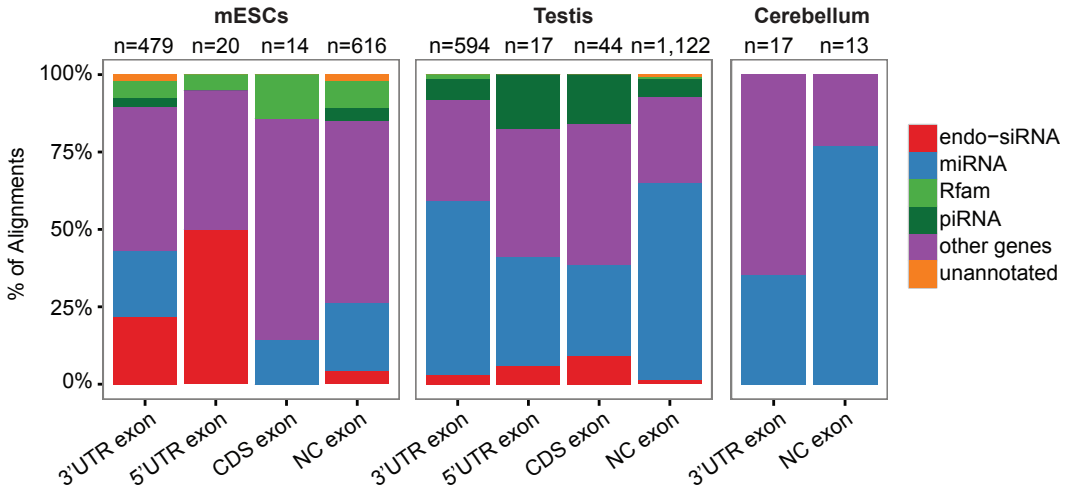
Supplementary Figure S3 | sciRNA-target prediction at varying mismatch cutoffs. The average SNR (Methods), total sciRNAs, total targets, and total sciRNA groups for each mismatch (mm) cutoff varying from 0 to 4 in 0.5 intervals in (a) mESCs, (b) testis, and (c) cerebellum.



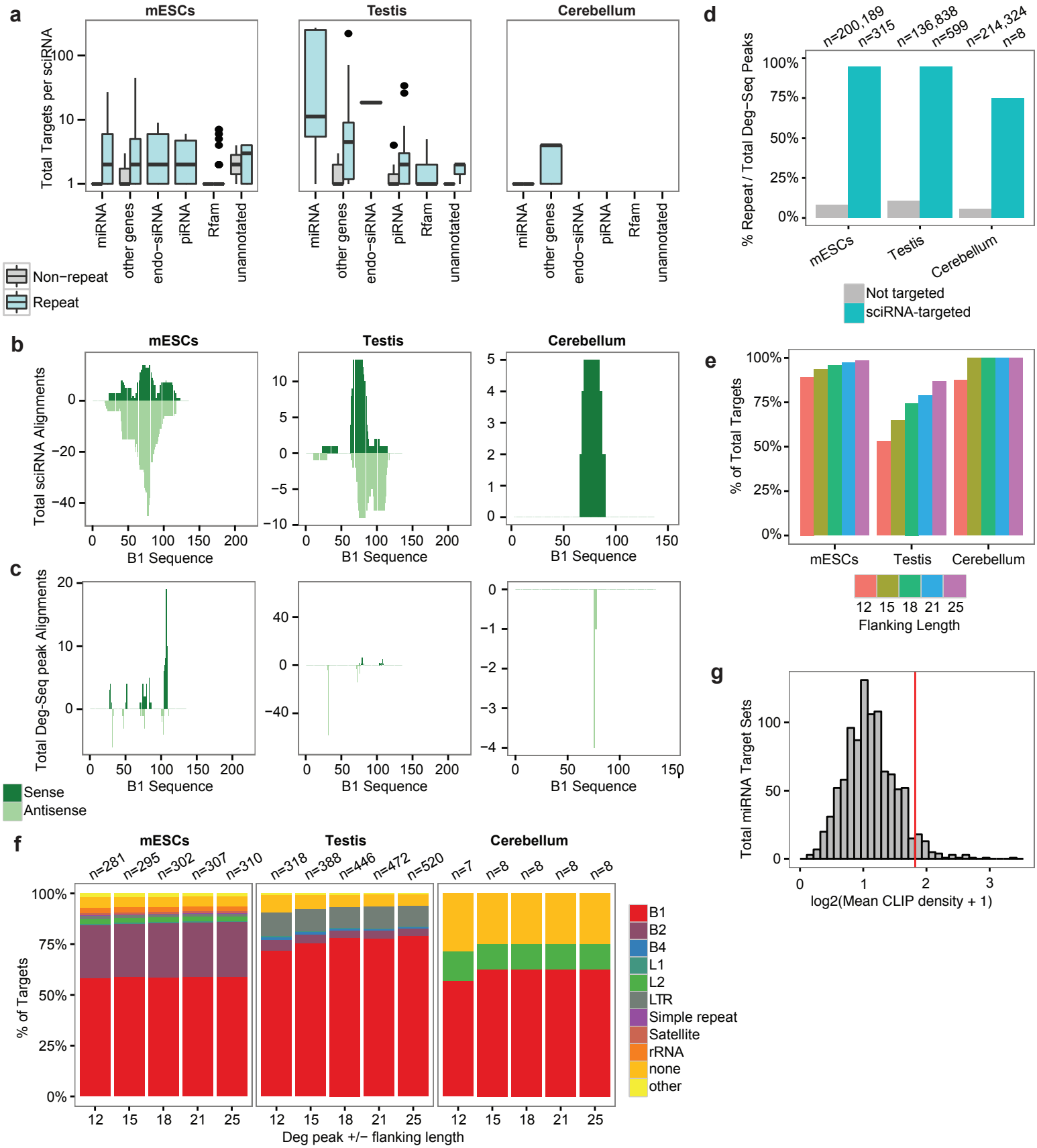
Supplementary Figure S4 | Characterization of piRNA-derived sciRNAs. piRNA-derived sciRNAs grouped by the sciRNA's relative location in the piRNA in mESCs and testis. 5' end: sciRNA starts at the first nt of the piRNA; 3' end: sciRNA ends at the last nt of the piRNA; middle: sciRNA starts and ends at internal nt of the piRNA.



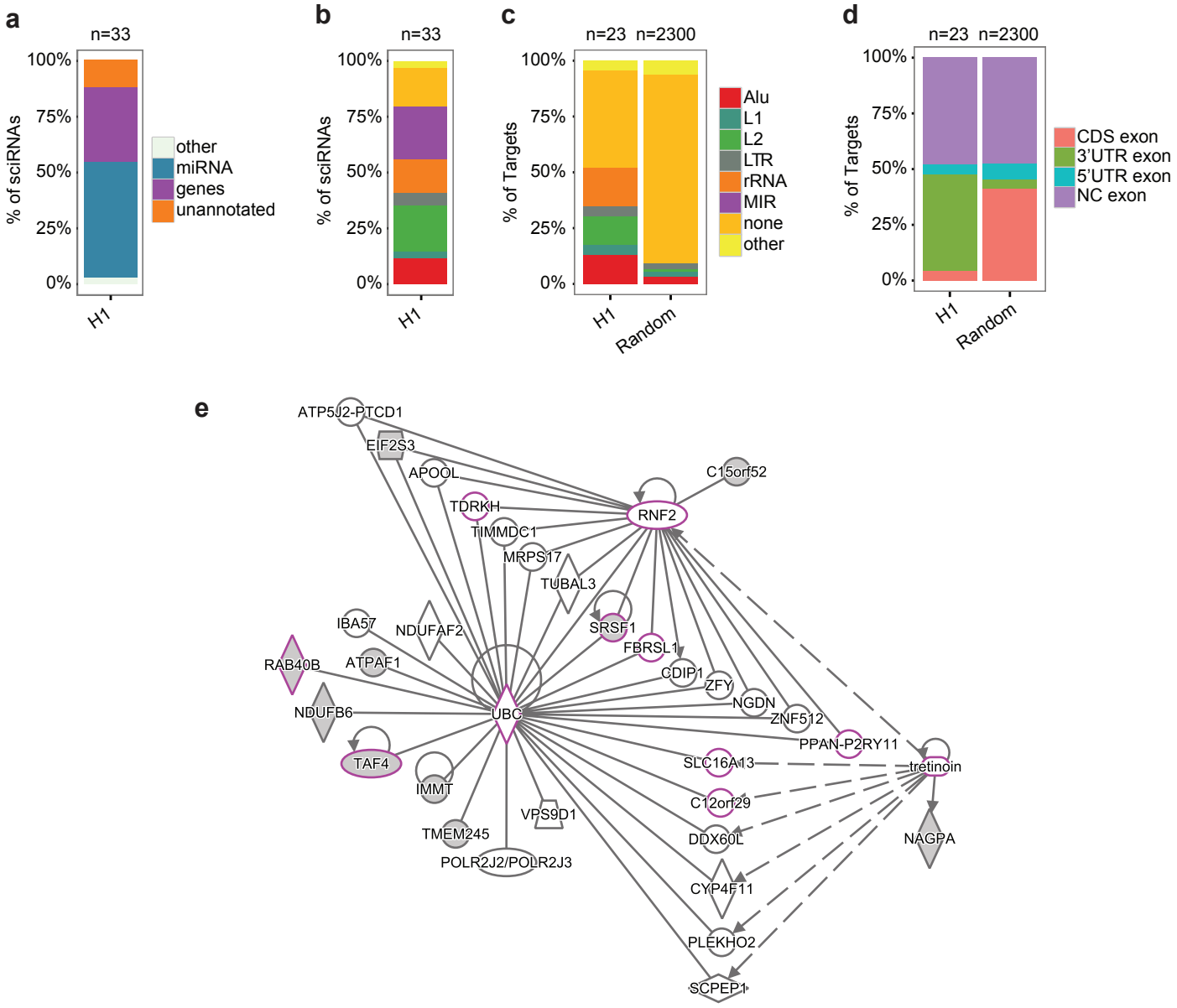
Supplementary Figure S5 | miRNA expression in testis and cerebellum development. Similar to Fig. 3c and 3d, but instead clustering all non-sciRNA miRNAs and labeling them based on having read count ≥ 20 in adult tissue (Testis 6M PN or Cerebellum 6M PN), mESCs, or both.



Supplementary Figure S6 | Characterization of sciRNA targets. The type of sciRNAs targeting each type of gene region is shown for each sciRNA-target alignment. CDS: coding sequence; NC exon: exon in non-coding transcript. The total number of sciRNA-target alignments is shown above each bar.



Supplementary Figure S7 | Repetitive nature of sciRNAs and targets. (a) Total number of targets per sciRNA separated by the subtype and repetitive vs. non-repetitive nature (Repeatmasker) of sciRNAs. Alignment of sciRNA (b) or Deg-Seq peak +/-5nt (c) to the B1 consensus sequence using BLASTN. Dark (light) green and positive (negative) values indicate sense (antisense) alignment. (d) Percentage of Deg-Seq peaks overlapping repetitive regions (any type in Repeatmasker) among all Deg-Seq peaks. Two groups of Deg-Seq peaks are shown: those targeted by sciRNAs (sciRNA-targeted) or otherwise (Not targeted). (e) Percent of total targeted cleavage sites with unique flanking sequence. Results shown for varying +/-12 to 25 nt. (f) Distribution of Repeatmasker annotations for target cleavage sites with unique flanking sequence for varying flanking sequence length (x-axis). For targets with the same flanking sequence, the following prioritization was used: B1 > B2 > B4 > others. (g) Histogram of 1,001 sets of 102 microrna.org miRNA targets based on their average *Ago2* CLIP-Seq density (empirical P = 0.049). 1,000 controls in grey; sciRNA target-overlapping miRNA targets in red.



Cancer, Cardiovascular Disease, Cell Cycle

Supplementary Figure S8 | Small RNA guided endonucleolytic cleavage in H1 ESCs. (a) sciRNA annotation. One unmapped sciRNA was excluded. (b) Repeatmasker family annotation of sciRNAs (c) Repeatmasker family annotation of target cleavage sites. Random: random positions from any transcript were chosen as a control cleavage site (see Methods). (d) Distribution of target cleavage sites (Deg-Seq peaks) in different regions of the transcriptome. CDS: coding sequence; NC exon: exon in non-coding transcript. Random: similarly defined as in (c). (e) One network was identified by Ingenuity Pathway Analysis of sciRNA targets. Grey-shaded nodes: sciRNA targets; magenta-outlined nodes: sciRNA targets associated with the top three diseases/functions shown below the network.