LETTERS

Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes

Oliver H. Tam[1]*, Alexei A. Aravin[1]*, Paula Stein[2], Angelique Girard[1], Elizabeth P. Murchison[1], Sihem Cheloufi[1], Emily Hodges[1], Martin Anger[2]†, Ravi Sachidanandam[1], Richard M. Schultz[2] & Gregory J. Hannon[1]

Pseudogenes populate the mammalian genome as remnants of artefactual incorporation of coding messenger RNAs into transposon pathways[1]. Here we show that a subset of pseudogenes generates endogenous small interfering RNAs (endo-siRNAs) in mouse oocytes. These endo-siRNAs are often processed from double-stranded RNAs formed by hybridization of spliced transcripts from protein-coding genes to antisense transcripts from homologous pseudogenes. An inverted repeat pseudogene can also generate abundant small RNAs directly. A second class of endo-siRNAs may enforce repression of mobile genetic elements, acting together with Piwi-interacting RNAs. Loss of Dicer, a protein integral to small RNA production, increases expression of endo-siRNA targets, demonstrating their regulatory activity. Our findings indicate a function for pseudogenes in regulating gene expression by means of the RNA interference pathway and may, in part, explain the evolutionary pressure to conserve argonaute-mediated catalysis in mammals.

Small-RNA-directed gene silencing pathways have been adapted to accept numerous inputs and to act on many types of downstream targets. In few places is this more apparent than in animal germ lines where two classes of small RNAs—microRNAs (miRNAs) and Piwi-interacting RNAs (piRNAs)—with distinct biogenesis mechanisms and biological functions have been reported. Although miRNAs, as a group, are ubiquitously expressed, piRNAs have thus far been found only in germ cells and in a few gonadal somatic cells types[2]. piRNAs repress the activity of mobile genetic elements, forming a small RNA-based, innate immune system with both genetically encoded and adaptive components[2–9].

In mice, a homozygous mutation in any single Piwi family member causes male sterility accompanied by gonadal hypotrophy[5,10,11]. In Mili and Miwi2 mutants, meiosis is not completed and germ cells are progressively lost[5]. This correlates with an activation of transposons, particularly the non-long terminal repeat (LTR) retrotransposon, L1 (refs 5, 12). DNA methylation of L1 elements is correspondingly lost. In contrast, females bearing homozygous mutations in individual Piwi genes are apparently normal and fertile[5,10,11]. Because female germ cells must also control transposons, we sought to characterize their small RNA profiles to determine whether a piRNA system, similar to that operating in spermatocytes, also exists in oocytes.

Approximately 6,000 fully grown oocytes, arrested in prophase of meiosis I, were collected. Small RNA fractions from 19–24 nucleotides (lower fraction) and 24–30 nucleotides (upper fraction) were gel purified and used to prepare small RNA libraries. These were deeply sequenced[3,4]. A total of 1,037,355 sequences was obtained that could be mapped to the mouse genome (753,981 from the lower fraction and 283,374 from the upper fraction; Supplementary Table 1). In the lower fraction, 126,515 non-redundant sequences were obtained, falling into 24,271 non-overlapping clusters. In the upper fraction, 97,807 non-redundant sequences fell into 15,032 non-overlapping clusters.

An examination of the small RNAs in the upper fraction of the oocyte library revealed a piRNA population that resembled those found in early-stage spermatocytes[3] (Fig. 1a, right). Roughly 62% of small RNAs correspond to annotated repeats (Supplementary Table 2), with 3% matching genic sequences and 3% matching unannotated, intergenic sites. The function of the latter species remains unknown. Roughly 30% of the library corresponded to presumed breakdown products of abundant, non-coding RNAs, such as ribosomal RNAs, transfer RNAs and small nucleolar RNAs.

As expected, oocyte piRNAs arise from discrete genomic loci in a strand-asymmetric fashion (Supplementary Table 3). A number of these loci share structural similarities to *Drosophila* piRNA loci, which act as master controllers of mobile elements[3]. One example (Fig. 1b) spans ~120 kb of chromosome 10 and contains an abundance of long interspersed elements (LINEs) and LTR elements. These have an orientation bias that results in the generation of predominantly antisense piRNAs (Fig. 1b, piRNA, weighted; see also Supplementary Fig. 1).

piRNAs have been proposed to act with transcripts from active transposons in a feed-forward amplification loop that confers signature features on a piRNA population that is mounting an ongoing transposon defence[2–4,6,7]. Primary piRNA-directed cleavage of transposon mRNAs creates the 5′ ends of secondary piRNAs[4,6]. This produces primary and secondary piRNA pairs that overlap by 10 nucleotides at their 5′ ends. The 5′ U bias of primary piRNAs thus leads to an enrichment of an A at position 10 of secondary piRNAs. These characteristics are prevalent in piRNA populations from mouse oocytes, particularly those that can be mapped to the L1 and intracisternal A particle (IAP) elements (Supplementary Fig. 2).

As expected, annotated miRNAs comprised the majority (69%) of 19–24-nucleotide RNAs (Fig. 1a, left; see also Supplementary Table 4). Among the highly abundant species are members of the let-7 family (let-7a/c/f), generally abundant miRNAs (miR-22, miR-16, miR-21, miR-26, miR-93 and miR-29a/b), and miRNAs abundant in ovary and placenta (miR-322, miR-503, miR-451). Finally, we detected miRNAs specifically expressed in male and female gonad (miR-103)[13].

A substantial fraction of 19–24-nucleotide RNAs matched annotated transposons (Supplementary Table 2). Many that mapped uniquely to the genome could be traced to oocyte piRNA loci (Fig. 1b, siRNA, weighted). These species might represent piRNA

[1]Cold Spring Harbor Laboratory, Watson School of Biological Sciences and Howard Hughes Medical Institute, 1 Bungtown Road, Cold Spring Harbor, New York 11724, USA. [2]Department of Biology, University of Pennsylvania, 433 South University Avenue, 205 Lynch Laboratories, Philadelphia, Pennsylvania 19104-6018, USA. †Present address: University of Oxford, Department of Biochemistry, South Parks Road, Oxford OX1 3QU, UK.
*These authors contributed equally to this work.

degradation products, or oocyte piRNA clusters might generate both siRNAs and piRNAs.

Therefore, we independently mapped piRNAs and candidate siRNAs to consensus L1 and IAP sequences (Supplementary Fig. 3). Each gave characteristic profiles. Moreover, piRNAs and candidate siRNAs show distinctly different nucleotide biases, with piRNAs displaying their characteristic enrichment for a 5′ uridine residue and an A at position 10 (Supplementary Fig. 2). Candidate siRNAs lack a 10A bias and show enrichment for both A and U residues at their 5′ ends (Supplementary Fig. 2). Finally, we gel purified 19–30-nucleotide RNAs from mouse oocytes as a single fraction and deeply sequenced this population. A length distribution of small RNAs that match the piRNA cluster shown in Fig. 1b yields two distinct peaks (Fig. 1c). siRNAs 21–22 nucleotides in length apparently predominate over the piRNA population, which averages ~27 nucleotides. We conclude that transposon-rich loci in oocytes give rise to both siRNAs and piRNAs. Although siRNAs are apparently more abundant, piRNA cloning frequencies could be reduced by the 2′-O-methyl modification that occurs on their 3′ termini[2]. Our results

raise the possibility that piRNA and siRNA systems may act redundantly to repress transposons in mouse oocytes, perhaps explaining the lack of substantial phenotypic consequences of individual Piwi mutations in females[5,10,11].

Although many transposons were targeted by both piRNAs and siRNAs, some relied more heavily on a particular pathway. For example, MTB and MTC, both LTR retrotransposons, matched almost exclusively to siRNAs. Moreover, the most prominent cluster that produces MTB/MTC small RNAs contains an inverted repeat with strong potential to produce a Dicer substrate. Notably, this transposon class showed increased expression in Dicer-null oocytes, consistent with its being regulated predominantly if not exclusively by the siRNA system[14].

Small RNA libraries often contain genic sequences. In other tissues, these correspond exclusively to sense sequences that probably represent contaminating degradation products. However, in oocytes, numerous sense and antisense siRNAs corresponding to protein-coding genes could be identified (Supplementary Table 5). As mammals lack any identifiable RNA-dependent RNA
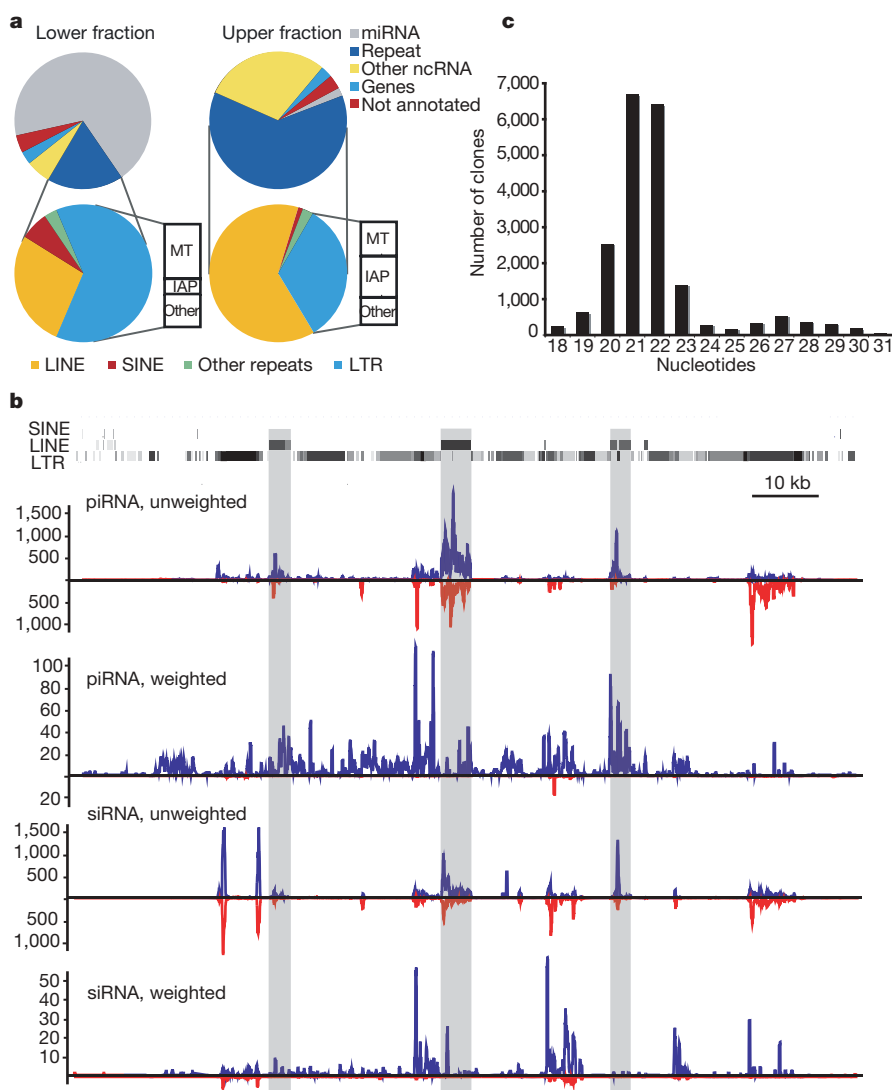

Figure 1 | Both piRNA and siRNA systems control transposons in mouse oocytes. a, Small RNA libraries from 19–24-nucleotide (lower fraction, left) and 24–30-nucleotide RNAs (upper fraction, right) were deeply sequenced. Reads were assigned an annotation as previously described[12]. The fraction of reads in each category is depicted. The repeat-annotated small RNAs were designated as LINE, SINE (short interspersed element), LTR and other, with the LTR category further divided between MT (mouse transcript), IAP and other. **b**, A representative piRNA locus on chromosome 10 is shown with the content of LINE, SINE and LTR fragments. Shading indicates the degree of match to the consensus element. Frequency plots for piRNAs and siRNAs from this locus are shown below. 'Unweighted' plots each match to the cluster; 'weighted' normalizes each match, dividing by its genomic frequency[4]. Blue and red lines indicate small RNAs mapping to the upper and lower genomic strand, respectively. **c**, RNAs from 19–30 nucleotides were deeply sequenced. Reads matching the piRNA cluster shown in **b** were used to construct a frequency plot by length.

polymerase, this raised the question of how antisense siRNAs might be generated.

On the basis of polymorphisms, uniquely mapping sense siRNAs could often be assigned to the functional protein-coding copy of a gene, whereas antisense siRNAs mapped to a homologous pseudogene (Fig. 2a and Supplementary Table 5). Thus, oocyte endo-siRNAs might be processed from double-stranded (ds)RNAs that form by hybridization of transcripts derived from two unlinked loci. A similar process in which transcripts from active transposons hybridize to antisense transposon fragments transcribed from piRNA clusters could explain the genesis of transposon siRNAs.

siRNAs from gene–pseudogene pairs arise exclusively from regions of complementarity between the partners. Because many sense-oriented siRNAs cross exon–exon junctions, we propose that mature, spliced mRNAs from genes interact with antisense pseudogene transcripts to form Dicer substrates (Supplementary Fig. 4). In one case (Fig. 2b), both sense and antisense siRNAs to the GTPase-activating protein for Ran (Ran-GAP) were produced from a pseudogene locus containing a ~300-base pair (bp) inverted repeat with an intervening ~800-base loop. siRNAs were derived only from the potentially double-stranded portion of this locus.

In some cases, Dicer processing of dsRNA substrates proceeds in an apparently processive fashion from a discrete initiation site, producing 'phased' small RNAs with an ~21-nucleotide periodicity[15]. Transposon-derived and genic siRNAs showed this property only very weakly (Supplementary Fig. 2). Notably, piRNAs also show a similar, very weak phasing signal, although with a period of ~27 nucleotides rather than ~21 nucleotides.

Pseudogenes have often diverged substantially from their functional ancestors. Thus, we wished to examine the possibility that pseudogene-derived antisense siRNAs could regulate corresponding protein-coding genes. We mapped antisense siRNAs to potentially relevant regulatory targets. Many small RNAs aligned to their targets either with no mismatches or with mismatches lying outside regions essential for slicer cleavage[16,17] (Fig. 2a, c). Thus, antisense, pseudo-gene-derived siRNAs might be capable of regulating homologous protein-coding genes through a conventional RNA-interference mechanism.

To test the regulatory potential of pseudogene-derived siRNAs, we assessed the effects of Dicer loss on their putative targets. We have previously shown that deletion of *Dicer* in growing oocytes causes the production of non-functional gametes with defects in spindle organization and chromosome segregation[14,18]. We compared the
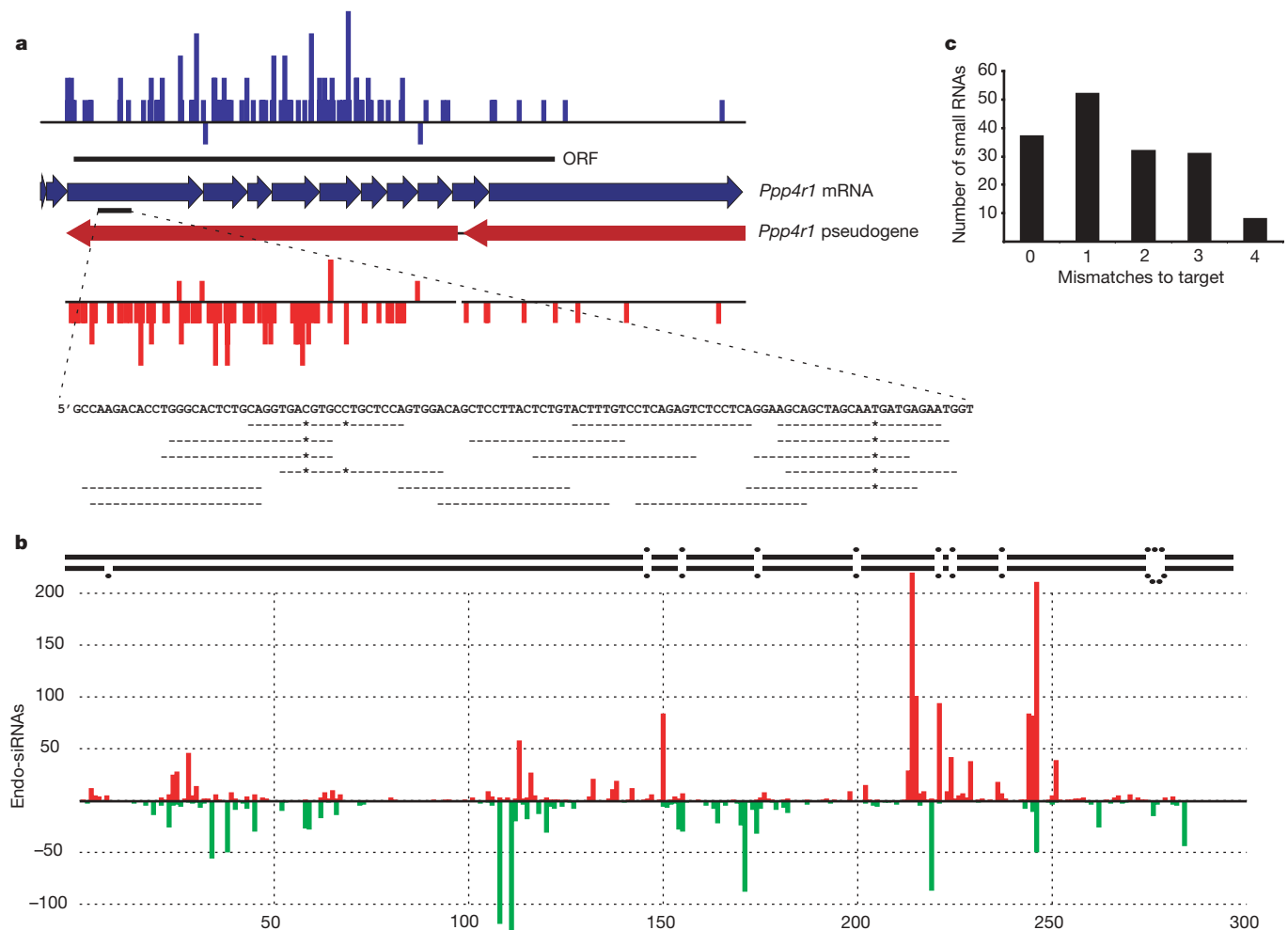

Figure 2 | Gene–pseudogene interactions produce endogenous siRNAs.
a, Endo-siRNAs unambiguously mapped to the functional *Ppp4r1* mRNA are plotted in blue, above the mRNA (individual exons indicated as thick arrows). The extent of the open reading frame (ORF) is indicated. Endo-siRNAs from the *Ppp4r1* pseudogene are shown in red below the pseudogene. Arrows indicate two segments of *Ppp4r1* homology. siRNAs plotted above each line are sense oriented, with respect to the functional mRNA, and those below are antisense. Shown below is an enlargement of one section of the mRNA with individual antisense siRNAs aligned. A dash indicates a match; an asterisk indicates a mismatch. The heights of the bars indicate the number of siRNAs starting at each position. **b**, Endogenous siRNAs homologous to Ran-GAP are shown below a schematic of the genomic inverted repeat structure from which they arise (the ~800-base loop is not depicted). Those shown above and below the x axis come from the upper or lower arm of the hairpin, respectively. No siRNAs were sequenced from the ~800-base loop that separates the two stem arms. **c**, siRNAs antisense to *Ppp4r1* counted and plotted by the number of mismatches to the functional mRNA.

expression of candidate endo-siRNA targets in wild-type and Dicer-null cells[14]. Many genes with abundant, pseudogene-derived siRNAs showed significant increases in expression following Dicer loss (Fig. 3a). We verified candidates derived from the array data by semi-quantitative polymerase chain reaction with reverse transcription (qRT–PCR, Fig. 3b).

Collectively, our data indicate that in mammalian oocytes, protein-coding mRNAs interact with pseudogene transcripts to form dsRNAs that are processed into endo-siRNAs. Examination of *Dicer* knockouts indicates a function for endo-siRNAs in gene regulation. At present, we cannot distinguish whether these small RNAs direct target cleavage or whether the act of siRNA production *per se*, which consumes the coding mRNA, is sufficient for repression. However, the specific case of HDAC1 may point to a RNA-induced silencing complex (RISC)-based mechanism. Few uniquely mapping

siRNAs are generated from the *Hdac1* gene itself, suggesting that it is not used prominently as a Dicer substrate. Instead, most uniquely mapping sense and antisense siRNAs can be assigned to a series of *Hdac1* pseudogenes. On the basis of its increased expression in Dicer-null oocytes, we propose that pseudogene-derived, antisense siRNAs direct RISC to cleave *Hdac1* mRNAs.

The catalytic potential of at least one argonaute protein has been conserved through mammalian evolution from platypus to humans[19,20]. However, mammalian miRNAs, with one known exception, act through translational mechanisms without the need for mRNA cleavage[21]. The discovery of endogenous siRNAs in mammalian oocytes not only expands the realm of mammalian small RNA classes but also provides one possible explanation for the evolutionary pressure to conserve argonaute enzymatic activity.

Pseudogenes have long been considered to be non-functional artefacts of transposition pathways acting on protein-coding mRNAs. In a few cases, regulatory roles have been posited for pseudogenes, largely through antisense mechanisms[22–24]. Our findings, and those of the accompanying paper[25], provide a role for a subset of mammalian pseudogenes in the production of functional siRNAs. The production of dsRNAs by interaction between sense and antisense transcripts from distinct loci has not been observed in other tissues and may require the unique environment of oocytes, which substantially lack a protein kinase R response (a dsRNA-induced general translational repression pathway) and are geared for mRNA stabilization and storage[26–28]. The fact that many targets of this pathway are related to microtubule dynamics (including microtubule-based processes, $P = 0$; kinesin complex, $P = 0$; motor activity, $P < 1 \times 10^{-254}$; spindle, $P < 8 \times 10^{-239}$; and microtubule-associated complex, $P < 3 \times 10^{-60}$; Supplementary Fig. 5) suggests that the regulatory circuits that we describe may have important biological roles, as the consequences of *Dicer* loss in growing oocytes is disruption of proper spindle formation and defects in chromosome segregation[14,18].

METHODS SUMMARY

Mouse oocytes were collected from primed mice and used to prepare small RNA fractions. These were cloned and deeply sequenced as previously described[2,4]. Bioinformatic analysis of the sequences was performed as described in the Methods. For semi-quantitative RT–PCR, RNA was extracted from fully grown oocytes from *Dicer^{flox/flox}* and *Dicer^{flox/flox} Zp3-cre* mice. Quantitative PCR was performed using TaqMan probes.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.
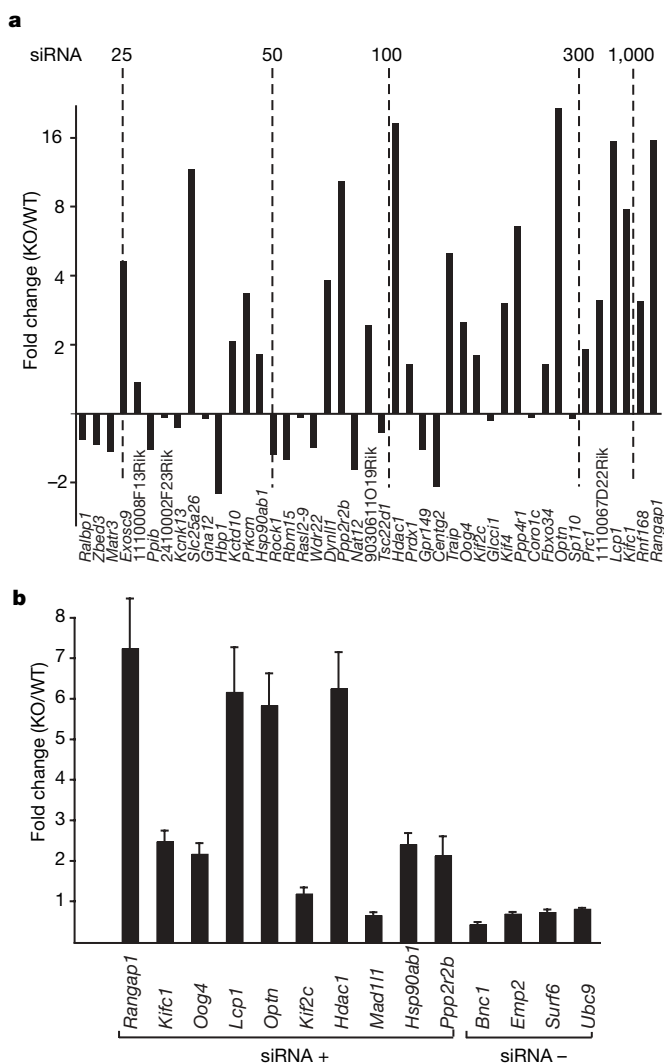
Figure 3 | Endo-siRNAs have a role in gene regulation. a, We had previously compared expression levels in wild-type and Dicer-null oocytes by microarray[14]. Genes with a large number of siRNAs were examined. For those with significant changes in expression ($P < 0.1$), the fold change in Dicer-null (KO) versus wild type (WT) was plotted. The graph was arranged according to the number of antisense siRNAs per gene in our data set, increasing from left to right (with benchmarks shown). The identity of the gene represented by each bar is given below. **b,** Fourteen genes (with and without siRNAs in our data set, indicated by siRNA+ and siRNA−, respectively), comprising a set in which some showed significant changes and some did not, were tested for changes in mRNA levels in Dicer-null versus wild-type oocytes by quantitative PCR. Gene names are indicated below each bar; error bars represent the mean ± s.e.m. (three replicates).

1. D'Errico, I., Gadaleta, G. & Saccone, C. Pseudogenes in metazoa: origin and features. *Brief. Funct. Genomics Proteomics* **3,** 157–167 (2004).
2. Aravin, A. A., Hannon, G. J. & Brennecke, J. The Piwi-piRNA pathway provides an adaptive defense in the transposon arms race. *Science* **318,** 761–764 (2007).
3. Aravin, A. A., Sachidanandam, R., Girard, A., Fejes-Toth, K. & Hannon, G. J. Developmentally regulated piRNA clusters implicate MILI in transposon control. *Science* **316,** 744–747 (2007).
4. Brennecke, J. *et al.* Discrete small RNA-generating loci as master regulators of transposon activity in *Drosophila. Cell* **128,** 1089–1103 (2007).
5. Carmell, M. A. *et al.* MIWI2 is essential for spermatogenesis and repression of transposons in the mouse male germline. *Dev. Cell* **12,** 503–514 (2007).
6. Gunawardane, L. S. *et al.* A slicer-mediated mechanism for repeat-associated siRNA 5′ end formation in *Drosophila. Science* **315,** 1587–1590 (2007).
7. Houwing, S. *et al.* A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in zebrafish. *Cell* **129,** 69–82 (2007).
8. Saito, K. *et al.* Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev.* **20,** 2214–2222 (2006).
9. Vagin, V. V. *et al.* A distinct small RNA pathway silences selfish genetic elements in the germline. *Science* **313,** 320–324 (2006).
10. Kuramochi-Miyagawa, S. *et al.* Mili, a mammalian member of piwi family gene, is essential for spermatogenesis. *Development* **131,** 839–849 (2004).
11. Deng, W. & Lin, H. miwi, a murine homolog of piwi, encodes a cytoplasmic protein essential for spermatogenesis. *Dev. Cell* **2,** 819–830 (2002).
12. Aravin, A. *et al.* A novel class of small RNAs bind to MILI protein in mouse testes. *Nature* **442,** 203–207 (2006).

13. Landgraf, P. *et al.* A mammalian microRNA expression atlas based on small RNA library sequencing. *Cell* **129**, 1401–1414 (2007).
14. Murchison, E. P. *et al.* Critical roles for Dicer in the female germline. *Genes Dev.* **21**, 682–693 (2007).
15. Allen, E., Xie, Z., Gustafson, A. M. & Carrington, J. C. microRNA-directed phasing during trans-acting siRNA biogenesis in plants. *Cell* **121**, 207–221 (2005).
16. Martinez, J. & Tuschl, T. RISC is a 5′ phosphomonoester-producing RNA endonuclease. *Genes Dev.* **18**, 975–980 (2004).
17. Schwarz, D. S., Tomari, Y. & Zamore, P. D. The RNA-induced silencing complex is a Mg^{2+}-dependent endonuclease. *Curr. Biol.* **14**, 787–791 (2004).
18. Tang, F. *et al.* Maternal microRNAs are essential for mouse zygotic development. *Genes Dev.* **21**, 644–648 (2007).
19. Joshua-Tor, L. The Argonautes. *Cold Spring Harb. Symp. Quant. Biol.* **71**, 67–72 (2006).
20. Murchison, E. P. *et al.* Conservation of small RNA pathways in platypus. *Genome Res.* (in the press).
21. Bartel, D. P. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* **116**, 281–297 (2004).
22. Korneev, S. A., Park, J. H. & O'Shea, M. Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J. Neurosci.* **19**, 7711–7720 (1999).
23. Weil, D., Power, M. A., Webb, G. C. & Li, C. L. Antisense transcription of a murine FGFR-3 psuedogene during fetal developement. *Gene* **187**, 115–122 (1997).
24. Zhou, B. S., Beidler, D. R. & Cheng, Y. C. Identification of antisense RNA transcripts from a human DNA topoisomerase I pseudogene. *Cancer Res.* **52**, 4280–4285 (1992).
25. Watanabe, T. *et al.* Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* doi:10.1038/nature06908 (this issue).
26. Stein, P., Zeng, F., Pan, H. & Schultz, R. M. Absence of non-specific effects of RNA interference triggered by long double-stranded RNA in mouse oocytes. *Dev. Biol.* **286**, 464–471 (2005).
27. Bettegowda, A. & Smith, G. W. Mechanisms of maternal mRNA regulation: implications for mammalian early embryonic development. *Front. Biosci.* **12**, 3713–3726 (2007).
28. Stitzel, M. L. & Seydoux, G. Regulation of the oocyte-to-zygote transition. *Science* **316**, 407–408 (2007).

METHODS

Mouse strains. Either CF-1 or CD-1 wild-type mice of 4–6 weeks were purchased from Harlan or Charles River Laboratories, respectively, and used to obtain oocytes for small RNA isolation. The *Dicer*$^{flox/flox}$ and *Dicer*$^{flox/flox}$ *Zp3-cre* mice, as previously reported[14], were used to obtain Dicer-deficient oocytes.

Generation of oocyte small RNA libraries. Wild-type mice were primed with 5 IU (international unit) PMSG 48 h before being killed, and fully grown germinal vesicle oocytes were collected as previously described[29]. Total RNA was extracted using Trizol (Invitrogen) according the manufacturer's protocol, and small RNA cloning was performed as described[4].

Quantitative real-time PCR. Total RNA was extracted from fully grown oocytes from *Dicer*$^{flox/flox}$ and *Dicer*$^{flox/flox}$ *Zp3-cre* mice using the Absolutely RNA Microprep Kit (Stratagene). cDNA was prepared by reverse transcription of total RNA with Superscript II and random hexamer primers. One oocyte equivalent of the resulting cDNA was amplified using TaqMan probes and the ABI Prism Sequence Detection System 7000 (Applied Biosystems). Three replicates of 45 oocytes each were used for RNA isolation and two replicates were run for each real-time PCR reaction; a minus template served as control. Quantification was normalized to the endogenous upstream binding factor (*Ubf*) within the log-linear phase of the amplification curve obtained for each probe/primer using the comparative C_T method (ABI PRISM 7700 Sequence Detection System, User Bulletin 2, Applied Biosystems, 1997). The TaqMan gene expression assays used were: Mm00441071_m1 (*Rangap1*), Mm00835842_g1 (*Kifc1*), Mm00620601_m1 (*Oog4*), Mm00786153_s1 (*Lcp1*), Mm00728630_s1 (*Kif2c*), Mm02391771_g1 (*Hdac1*), Mm00487521_m1 (*Mad1l1*), Mm00725286_m1 (*Optn*), Mm00833431_g1 (*Hsp90ab1*), Mm00511698_m1 (*Ppp2r2b*), Mm00801709_m1 (*Emp2*), Mm00486494_m1 (*Surf6*), Mm00456972_m1 (*Ubf*). For *Bcn1* and *Ubc9*, custom TaqMan Gene Expression Assays were used that had the following primers and probes: *Bcn1* forward primer 5′-ACTGGACGCTTCAGGATTACATC-3′, *Bcn1* reverse primer 5′-GTCATGATGCTCCAGTGATCCA-3′, *Bcn1* probe 5′FAM-TTCCCAGAGGCATCCTG-3′; *Ubc9* forward primer 5′-CAGGTGAGAGCC-AAGGACAAA-3′, *Ubc9* reverse primer 5′-GGCCCACTGTACAGCTAACA-3′, *Ubc9* probe 5′FAM-CTGGCCTGCATTGATC-3′.

Bioinformatic analysis. Small RNAs were sequenced using the Illumina 1G platform. Sequencing of the upper and lower fraction libraries produced 2,785,080 reads, of which 1,037,355 (37%) could be mapped to the mouse genome (release mm9, July 2007) with no mismatches. The small RNAs are matched to a suffix array generated from the mouse genome, keeping track of exact matches to genome. Repeat masking was not used, but small RNA sequences with more that ten identical nucleotides in a row were removed from consideration. Annotation categories were assigned based on the annotation of corresponding genomic sequences extracted from the UCSC genome browser. The genome was annotated with mRNAs, non-coding RNAs and repeats. The annotations at the mapping positions (up to five mappings per small RNA) were used, along with a majority rule, to assign an annotation to each small RNA. The sequences were also re-analysed to allow 1–2 mismatches to the genome. Although the number of mapped sequences increased (from 37% to 54%), the genomic origin of repeat-associated small RNAs became ambiguous (Supplementary Fig. 1), and therefore non-informative.

To extract small RNA clusters (both piRNA and siRNA), the genome was scanned to look for regions that had more than ten uniquely mapping small RNAs, and the boundaries were defined as the location of the first/last small RNA in the cluster. To identify sequences that match the consensuses for transposable elements, the small RNAs were aligned to consensus sequences from release 11.08 of Repbase (http://www.girinst.org). The following consensuses were used: L1_MM for LINE L1 and IAPLTR1a_I_MM for the IAP retrotransposon. Matches to consensus sequences with up to three mismatches were recovered and included in the analysis. Nucleotide biases were calculated for small RNAs matching L1 and IAP consensuses as described[4]. To identify gene–pseudogene pairs, the genomic sequences of the siRNA clusters were extracted from the UCSC genome browser, and re-matched to the genome using BLAT (http://www.genome.ucsc.edu). Genomic regions with greater than 95% identity were identified, and small RNAs (both sense and antisense) mapped to these locations were extracted.

Gene ontogeny analysis of endo-siRNA targets was carried out as previously described[30] using GOBAR, which uses a hypergeometric statistic to identify nodes that are significantly enriched. A bootstrapping technique, involving repeated sampling from the reference set, is used to assign significance values to the results.

29. Schultz, R. M., Montgomery, R. R. & Belanoff, J. R. Regulation of mouse oocyte meiotic maturation: implication of a decrease in oocyte cAMP and protein dephosphorylation in commitment to resume meiosis. *Dev. Biol.* **97**, 264–273 (1983).

30. Lee, J. S., Katari, G. & Sachidanandam, R. GObar: a gene ontology based analysis and visualization tool for gene sets. *BMC Bioinformatics* **6**, 189 (2005).