Abundant and dynamically expressed miRNAs, piRNAs and other small RNAs in the vertebrate *Xenopus tropicalis*

Javier Armisen[1,2,3], Mike Gilchrist[1,3], Anna Wilczynska[2], Nancy Standart[2] and Eric A. Miska[1,2,4]

[1] Wellcome Trust Cancer Research UK Gurdon Institute, University of Cambridge, The Henry Wellcome Building of Cancer and Developmental Biology, Tennis Court Rd, Cambridge CB2 1QN, UK

[2] Department of Biochemistry, University of Cambridge, Tennis Court Rd, Cambridge CB2 1GA, UK

[3] These authors contributed equally to this work

[4] Corresponding author: Email e.miska@gurdon.cam.ac.uk; phone +44-1223-767225; fax +44-1223-767225

Running title: *Xenopus* small RNAs

Key words: miRNA, siRNA, endo-siRNA, piRNA, germline, *Xenopus*

Supplemental Methods

β-elimination

β-elimination was performed as described previously (Horwich et al. 2007) with the following modifications: 20*μ*g of total RNA from mixed stage oocytes or liver was incubated with 4*μ*l 5×borate buffer (148mM borax, 148mM boric acid, pH8.6) and 2.5*μ*l freshly dissolved 200mM $NaIO_4$ to a final volume of 20*μ*l, and incubating for 10min at room temperature. 2*μ*l glycerol was added to quench unreacted $NaIO_4$ and incubated for an additional 10min at room temperature. Samples were dried by centrifugation under vacuum for 30-60min at room temperature, dissolved in 50*μ*l of 1×borax buffer (30mM borax and 30mM boric acid, 50mM NaOH, pH9.5), and incubated for 90min at 45°C. To precipitate the RNA 20*μ*g of glycogen and 3 volumes ethanol were added and incubated at −80°C for 1h. Samples were collect by centrifugation (13,000 rpm, 4°C, 20min) and dissolved in denaturing gel loading buffer (95% formamide, 18mM EDTA and 0.025% each of SDS, xylene cyanol, and bromophenol blue) or H_2O.

Immunoprecipitation of piRNAs

piwil1-a/piwil1-b antiserum or pre-immune serum was incubated with pre-cleared *X. tropicalis* lysate overnight at 4 °C. 40 µl protein A-Sepharose beads (GE Healthcare) were washed in 1 ml blocking buffer (20 mM Tris-HCl, pH 7.5, 150 mM NaCl, 1.5 mM $MgCl_2$, 0.5% Triton-100, 0.5% NP-40, 1 mM DTT, 100 u/ml RNAse inhibitor, 1x Complete EDTA-free protease inhibitors (Roche Applied Science)). Beads were incubated with the lysate containing serum for 2 hours at

4 °C. Beads were washed 3 times in washing buffer (20 mM Tris-HCl, pH 7.5, 300 mM NaCl, 1.5 mM $MgCl_2$, 0.5% Triton-100, 0.5% NP-40, 1 mM DTT, 100 u/ml RNAse inhibitor). RNA was trizol extracted followed by ethanol precipitation.

5′ end-labeling

Following *β*-elimination RNA was dephosphorylated using antartic phosphatase (NEB, Ipswich, MA, USA) according to the manufacturer's protocol, followed by purification on an Illustra G-25 microspin column (GE Lifesciences, Uppsala, Sweden) to remove free phosphate groups. Samples were 5'-end labeling with *γ*-[^{32}P]-ATP using T4 Polynucleotide Kinase (NEB) and purified again on an Illustra G-25 microspin column to remove unincorporated nucleotides. Gels were analyzed using a PhosphorImager system.

Northern blotting

Small RNA northern blotting using DNA probes was performed as described previously (Miska et al. 2004; Pall and Hamilton 2008) using oocytes equivalents. 1-ethyl-3-[3-dimethylaminopropyl]carbodiimide hydrochloride (EDC, Perbio Science, Erembodegem, Belgium) cross-linking reactions were carried out for 2h at 60°C. Northern hybridizations were modified as follows: membranes were pre-hybridized at 40°C for 4h in hybridization buffer (0.36M Na_2HPO_4, 0.14M NaH_2PO_4, 7% SDS and 1mg of sheared, denatured, salmon sperm DNA) and hybridized at 40°C overnight using 20pmole of a *γ*-[^{32}P]-ATP radiolabeled probe. After hybridization, membranes were washed twice with 0.5×SSC, 0.1% SDS at 40°C for 10min and once with 0.1×SSC, 0.1% SDS at 40°C

for 5min. Probe sequences used were the reverse complement of the miRNA or piRNA target.

Reverse transcription polymerase chain reaction (RT-PCR) and quantitative RT-PCR

Total RNA from oocyte equivalents was isolated and treated with RQ1 RNase-Free DNase (Promega, Madison, WI, USA) for 30min at 37°C. Samples were phenol/chloroform extracted and ethanol precipitated. Reverse transcription (RT) was performed according to the Superscript II protocol using an oligo-dT primer. The reaction was incubated at 42°C for 60min, 65°C for 10min and diluted with H_2O to 50μl final volume. 2μl were used for standard PCR (30 cycles of 95°C for 20s, 60°C for 30s and 72°C for 30s). qRT-PCR was performed using Quantitect SYBR green PCR mix (Qiagen, Hilden, Germany). qRT-PCR primer design was as described previously (Chen et al. 2005). For each reaction, 10μl of 2×Quantitect SYBR green PCR mix, 0.2μl of 100μM specific forward primer, 0.2μl of 100μM specific reverse primer, 8.27μl of RNase-free H_2O and 1.33μl of RT product was incubated at 95°C for 10min, followed by 40 cycles of PCR (95°C for 15s, 60°C for 1min) on a 7300 Real Time PCR System (Applied Biosystems, Foster City, CA, USA). All reactions were run in triplicate. PCR and qRT-PCR products were resolved on a 2% agarose gel. Sequences of oligonucleotides used for qRT-PCR are listed in Armisen_SupData4.xls.

Sequencing data analysis

Illumina reads were processed from the *fastq* format and loaded into a database

table. Raw reads were collapsed into a second database table, one lane at a time. Identical reads were only counted once, but the number of times they occur was recorded. Each unique read was named by the lexically lowest identifier of the group (this was done in the second collapse step also). Lane data was stored alongside the reads. The set of collapsed reads was exported to a *fasta* file, which was then BLASTed against databases consisting of either the 5' or 3' adaptor sequence (Altschul et al. 1997). The resultant blast data was used to identify the presence and end position of either (or both) adaptor sequence(s). Additionally, if the last 3 or 4 nt of the read were identical to the start of the 3' adaptor, then this was also assumed to be the adaptor. Reads were discarded at this stage if (a) the 5' adaptor was identified, (b) the 3' adaptor was not identified, (c) there were <18 bases before the 3' adaptor, and (d) the sequence up to the 3' adaptor were primarily poly(A). The sequence of the remaining reads from the first base to the last base before the 3' adaptor were termed the tag and extracted into another column in the same table. Tags were then collapsed into a third database table in groups of lanes; the three oocyte lanes in one group and the two somatic cell lanes in another. For each group tags were exported to a *fasta* file which was then BLASTed against the *Xenopus tropicalis* genome (JGI v4.1, http://www.jgi.doe.gov/) with appropriate E-value and word size to find all exact matches up to some initial limit (to prevent data overload). A second round of BLAST was performed with those tags (a much smaller number) that saturate at this limit, in such a way as to find all exact matches. Tags that failed to match the genome perfectly were discarded. The remaining, genome filtered tags were between 18 and 42 bases long and were extracted into another *fasta* file, and

matched using BLAST to known RNA families: miRNAs, ribosomal RNA, transfer RNAs, and other non-coding RNAs deposited in Rfam, using a cutoff of 80% sequence identity (Griffiths-Jones 2004; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008; Gardner et al. 2009).

RNA block analysis. Groups of neighboring tags were clustered together in blocks. A block was defined as a group of overlapping or neighboring tags with no more than 200 nucleotides (nt) between two tags in the block. Single tag locus blocks were defined as any block where the longest tag was within 10 nt of both ends. These were also referred to as isolated tags.

Prediction of miRNA candidates. Tags to be tested for potential miRNA precursor folds were extracted from the genome sequence with either -12 and +48 or -48 and +12 bases either side of the tag locus. These candidate precursor sequences were then processed using RNAfold (Hofacker and Stadler 2006), and the output analyzed for hairpin folds with the tag aligned along one of the hairpin arms. Tags that showed acceptable folds (Ambros et al. 2003) were deemed candidate miRNAs unless already identified by similarity searches as known miRNAs deposited in the miRNA Registry (Griffiths-Jones 2004; Griffiths-Jones et al. 2006; Griffiths-Jones et al. 2008). All miRNA candidates that were represented in our libraries with abundant reads and for which we could also detect miR* reads were submitted to the miRNA registry.

Mapping of small RNAs to repeat elements. A *fasta* file of *Xenopus* repeats from Repbase (Jurka et al. 2005) was searched using BLASTn against the *Xenopus tropicalis* genome sequence (JGI v4.1, http://www.jgi.doe.gov/) using non-default parameters -e 1e-10 -b 20 -F mD. For each repeat sequence this yields all

the alignments on the 20 scaffolds with best single alignment. No attempt was made to check duplication of data within families of repeats. The genomic coordinates for all found alignments for each repeat in these scaffolds sequences were then compared with the set of exactly mapped positions for all tags on all scaffolds. This enabled us to identify and count all the tags in each mapped repeat sequence alignment. For each repeat sequence we report the number of scaffolds with at least one hit better than 1e-10 up to a maximum of 20, the number of separate alignments (mostly partial) for each repeat on these scaffolds, the total number of tag loci (tags may be counted more than once), the number of unique tags (tags only counted once), the number of reads yielding any of the tags mapped to the repeat, and the number of diluted reads (reads divided by total loci).

Access to sequencing data. All data were submitted to the GEO database and have the following the accession number: GSE14952 (sample series). Data are also available for download at:

http://informatics.gurdon.cam.ac.uk/online/solexa/X_trop_stage_1-2.fastq

http://informatics.gurdon.cam.ac.uk/online/solexa/X_trop_stage_3-4.fastq

http://informatics.gurdon.cam.ac.uk/online/solexa/X_trop_stage_5-6.fastq

http://informatics.gurdon.cam.ac.uk/online/solexa/X_trop_liver.fastq

http://informatics.gurdon.cam.ac.uk/online/solexa/X_trop_skin.fastq

http://informatics.gurdon.cam.ac.uk/online/solexa/genome-tags-somatic-JAG2.fasta

http://informatics.gurdon.cam.ac.uk/online/solexa/genome-tags-oocyte-JAG3.fasta

Supplemental References

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Ambros, V., Bartel, B., Bartel, D.P., Burge, C.B., Carrington, J.C., Chen, X., Dreyfuss, G., Eddy, S.R., Griffiths-Jones, S., Marshall, M. et al. 2003. A uniform system for microRNA annotation. *RNA* **9:** 277-279.

Chen, C., Ridzon, D.A., Broomer, A.J., Zhou, Z., Lee, D.H., Nguyen, J.T., Barbisin, M., Xu, N.L., Mahuvakar, V.R., Andersen, M.R. et al. 2005. Real-time quantification of microRNAs by stem-loop RT-PCR. *Nucleic Acids Res* **33:** e179.

Gardner, P.P., Daub, J., Tate, J.G., Nawrocki, E.P., Kolbe, D.L., Lindgreen, S., Wilkinson, A.C., Finn, R.D., Griffiths-Jones, S., Eddy, S.R. et al. 2009. Rfam: updates to the RNA families database. *Nucleic Acids Res* **37:** D136-140.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res* **32:** D109-111.

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A., and Enright, A.J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* **34:** D140-144.

Griffiths-Jones, S., Saini, H.K., van Dongen, S., and Enright, A.J. 2008. miRBase: tools for microRNA genomics. *Nucleic Acids Res* **36:** D154-158.

Hofacker, I.L. and Stadler, P.F. 2006. Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics* **22:** 1172-1176.

Horwich, M.D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., and Zamore, P.D. 2007. The Drosophila RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* **17:** 1265-1272.

Jurka, J., Kapitonov, V.V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* **110:** 462-467.

Miska, E.A., Alvarez-Saavedra, E., Townsend, M., Yoshii, A., Sestan, N., Rakic, P., Constantine-Paton, M., and Horvitz, H.R. 2004. Microarray analysis of microRNA expression in the developing mammalian brain. *Genome Biol* **5:** R68.

Pall, G.S. and Hamilton, A.J. 2008. Improved northern blot method for enhanced detection of small RNA. *Nat Protoc* **3:** 1077-1084.

List of Supplemental Data Files

Armisen_SupData1.xls	known miRNAs (MS Excel)
Armisen_SupData2.xls	repeat-associated RNAs in oocytes (MS Excel)
Armisen_SupData3.xls	repeat-associated RNAs in the soma (MS Excel)
Armisen_SupData4.xls	RT-PCR primer sequences (MS Excel)

Supplemental Tables

Supplemental Table 1. Small RNA library sequencing summary

Name	Source	# primary reads	# mapped reads	# unique tags
X_trop_stage_1-2	*X. tropicalis* oocytes stage I, II	6,691,824	3,641,616	549,672
X_trop_stage_3-4	*X. tropicalis* oocytes stage III, IV	7,214,118	2,851,821	566,626
X_trop_stage_5-6	*X. tropicalis* oocytes stage V, VI	7,397,640	1,703,200	385,001
X_trop_liver	*X. tropicalis* adult liver	6,573,612	758,445	87,720
X_trop_skin	*X. tropicalis* adult skin	6,009,764	1,670,064	87,960

Supplemental Table 2. Most frequently sequenced known miRNAs from somatic libraries

Library	**miRNA**	**Most common tag**	**# reads**
liver	miR-122	TGGAGTGTGACAATGGTGTTTG	135435
	miR-146b	TGAGAACTGAATTCCATGGACT	3805
	miR-101	TACAGTACTGTGATAACTGAAG	3573
	let-7f	TGAGGTAGTAGATTGTATAGTT	2706
	let-7e	TGAGGTAGTAGGTTGTTTAGTT	1968
	miR-143	TGAGATGAAGCACTGTAGCTC	1920
	miR-451	AAACCGTTACCATTACTGAGTTT	1703
	miR-30e	CTTTCAGTCGGATGTTTACAGC	1539
	let-7a	TGAGGTAGTAGGTTGTATAGTT	759
	miR-26a	TTCAAGTAATCCAGGATAGGCT	695
skin	miR-451	AAACCGTTACCATTACTGAGTTT	113648
	miR-10b	TACCCTGTAGAACCGAATTTGT	36498
	miR-146b	TGAGAACTGAATTCCATGGACT	29433
	miR-204	TTCCCTTTGTCATCCTATGCCT	14040
	miR-27b	TTCACAGTGGCTAAGTTCTGC	11653
	let-7e	TGAGGTAGTAGGTTGTTTAGTT	8488
	miR-143	TGAGATGAAGCACTGTAGCTC	5556
	miR-1	TGGAATGTAAAGAAGTATGTAT	4324
	let-7a	TGAGGTAGTAGGTTGTATAGTT	3724
	miR-214	TACAGCAGGCACAGACAGGCAGT	3525

Supplemental Table 3. Novel miRNAs identified in this study

	xtr-miR-2184
mature	AACAGUAAGAGAUUAUGUGCUG
precursor	CUUGCGUCUCGGaacaguaagagauuaugugcugUGUUAUCAGGCAGCCGGCACAUGGCCUUUUACUGCUCAGAGAGGCAGG
fold	(((((.((((.((.((((((((((..((((((((((.(((.......)))))))))))))..))))))))).)).)))).))))))
reads of miR	543
reads of miR*	58
	xtr-miR-2188
mature	AAGGUCCAGCCUCAUAUGUCCU
precursor	GGGCGUGUGGGAaagguccagccucauauguccuGUGAUCCUGAGGGGGAGAUAUGUGGUCAGACCUGUCCCACAGGCCGUG
fold	.(((.(((((((.((((((..(((.((((((.((((............)))).)))))).)))..))))))).)))))))...)
reads of miR	1414
reads of miR*	3

RNA secondary structure prediction was done using RNAfold (Hofacker and Stadler 2006). mature, refers to the most abundant read.

Supplemental Table 4. Strand bias of piRNA clusters

bias	clusters
100	2279
99	12
95	16
94	17
93	24
92	23
91	12
90	20
89	14
88	17
87	11
86	22
85	20
84	18
83	16
81	25
76	12
63	13
50	15

Bias is measured as 100*sum(strand)/tags in each cluster. Bias values have been rounded down to the nearest integer, so all 100% values are really 100%. Only clusters with ≥10 tags and > 50 nt length were considered. The analysis (both clusters and bias) is based on unique mapping tags with length 25 - 30.

Supplemental Table 5. Repetitive elements matching small RNAs in *X. tropicalis* oocytes libraries

repeat	type	hits < 1e-20	distinct tags	tag loci	reads	unique	unique bias
Polinton-2	DTN	217	8839	21281	506879	3450	-2194
Polinton-1	DTN	206	8641	29131	70324	2971	-2505
Harbinger-2	DTA	524	5296	51587	68542	338	-138
hAT-9	DTA	503	4162	81922	63950	152	146
hAT-10	DTA	1653	5833	148479	51713	749	-625
L1-55	LINE	101	5453	19776	42133	1775	-1057
piggyBac-1	DTA	620	4624	138735	36720	110	-108
ERV1-3-LTR	LTR	36	570	6752	20584	13	-13
piggyBac-1N1	DTA	996	2418	178823	15073	41	-39
Harbinger-5	DTA	292	1222	2187	12003	127	-67
Harbinger-4	DTA	83	1468	3126	9864	740	-372
Harbinger-1	DTA	845	3197	20081	9313	234	-48
Tc1-8Xt	DTA	107	1259	5800	9124	340	-222
piggyBac-2	DTA	486	2531	45858	8667	238	-220
ERV1-2-LTR	LTR	29	463	2103	8139	84	-72

Repeat, name of repeat from GIRI data file; hits < 1e-20, number of matches in up to 20 scaffolds; distinct tags, number of distinct tags found in repeat; tag loci, number of tag loci found in repeat (all types); reads, total number of reads for tags that match within the given repeat; unique, number of tags which map only once found in repeat; unique bias, sum (unique tags x strand relative to repeat). DTN, DNA transposon non-autonomous. DTA, DNA transposon autonomous. LINE, Long interspersed elements. LTR, retrotransposon.

Supplemental Table 6. Repetitive elements matching small RNAs in *X. tropicalis* somatic libraries

repeat	type	hits < 1e-20	distinct tags	tag loci	reads	unique	unique bias
Polinton-2	DTN	217	217	527	11242	74	-52
Harbinger-2N1	DTA	816	756	21937	4581	5	-5
Harbinger-N5	DTA	491	372	20874	3209	0	0
Harbinger-N1	DTA	431	350	22263	3039	11	-11
Harbinger-2	DTA	524	344	3501	2190	20	-14
Harbinger-2N2	DTA	1055	267	19245	2092	5	1
hAT-10	DTA	1653	412	16876	1512	15	-11
Tc1_XL	DTA	495	123	6735	1080	6	0
hAT-N2	DTA	976	175	5452	1054	0	0
Harbinger-N6	DTA	423	197	9658	1010	0	0
Harbinger-N3	DTA	1088	344	4788	991	6	-2
hAT-N2A	DTA	552	72	1449	801	0	0
Harbinger-N9	DTA	770	224	6485	717	14	0
piggyBac-N1	DTA	1418	147	4542	640	8	4
hAT-9	DTA	503	178	3770	620	1	1

Repeat, name of repeat from GIRI data file; hits < 1e-20, number of matches in up to 20 scaffolds; distinct tags, number of distinct tags found in repeat; tag loci, number of tag loci found in repeat (all types); reads, total number of reads for tags that match within the given repeat; unique, number of tags which map only once found in repeat; unique bias, sum (unique tags x strand relative to repeat). DTN, DNA transposon non-autonomous. DTA, DNA transposon autonomous. LINE, Long interspersed elements. LTR, LTR retrotransposon.

Supplemental Figure Legends

Supplemental Figure 1. (*A*) Small RNA expression throughout oogenesis in *Xenopus tropicalis* and *Xenopus laevis* was analyzed. For each experiment total RNA from 150 oocytes was extracted, size-selected using the miRVana kit and loaded on a 15% denaturing gel. RNA was stained using SYBR green. Bands likely representing miRNAs, endo-siRNAs and piRNAs are indicated with an asterix.

Supplemental Figure 2. piwil1-a/piwil1-b protein is expressed in the germline. Oocyte samples were prepared as described previously (Wilczynska et al. RNA 2009). *Xenopus tropicalis* tissue samples were lysed in RIPA buffer. For western blotting, 2 oocyte equivalents and similar protein amounts from tissue samples (assessed by Coomassie stain) were resolved in a 10% SDS-PAGE and probed with anti-piwil1-a/piwil1-b and anti-actin as a loading control. No piwil1-a/piwil1-b was detected in any of the adult tissues analyzed.

Supplemental Figure 3. (*A-D*) Length distributions of short RNAs. Data shown were grouped into germline (*A*,*B*) and somatic (*C*,*D*) libraries and tags (*A*,*C*) or reads (*B*,*D*), respectively. Blocks were defined by groups of tags of a given type with no gaps between neighbors greater than a fixed value (200 bases). Tag types used were single locus tags, tags with 10+ loci, and all tags, defining high copy number, low copy number and mixed blocks, respectively. Tags not in blocks were termed isolated, tags were grouped into low copy number or high copy number tags.

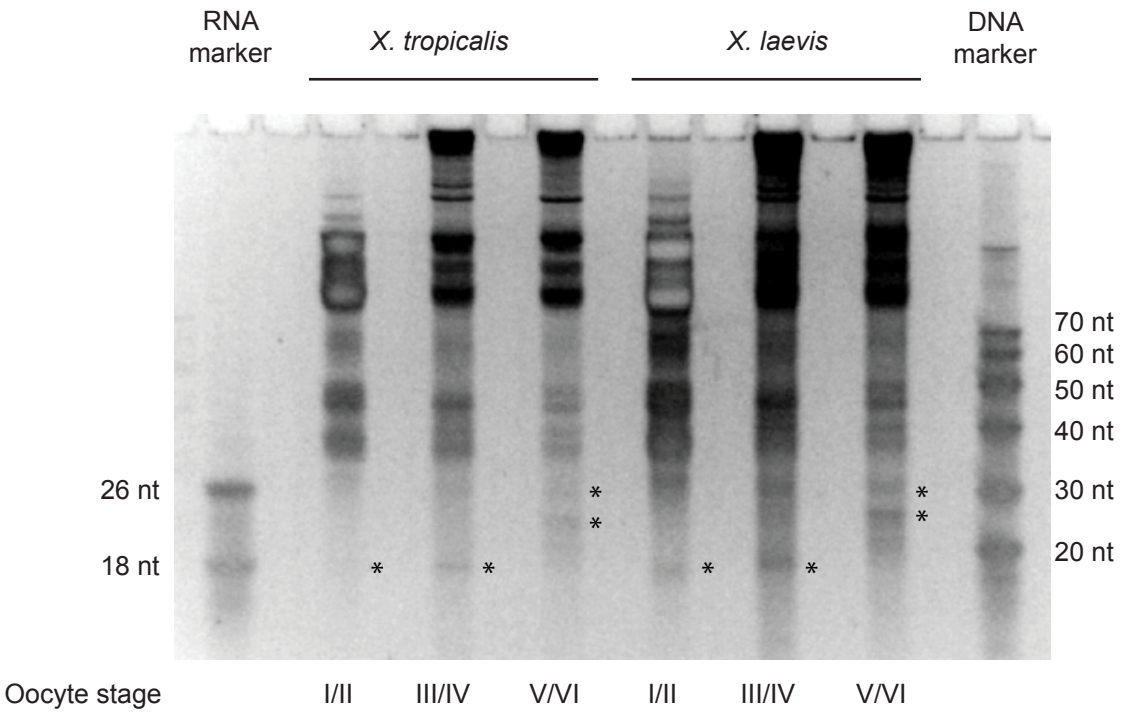
Supplemental Figure 4. Expression of miR-148a was assessed in oocytes, eggs and follicular cells separated from ovarian lobes. The northern blot was also hybridized with a probe for 5S rRNA to control for loading.

Supplemental Figure 5. (A-C) Expression of the three previously identified miRNAs (miR-101, miR-202-5p, miR-148a) in *Xenopus laevis* oocytes. 150 oocyte stages V and VI were pooled and used for northern blot analysis. EDC was used to crosslink small RNAs to the membrane prior to hybridization. Membrane was hybridized with 5S rRNA to control for loading.

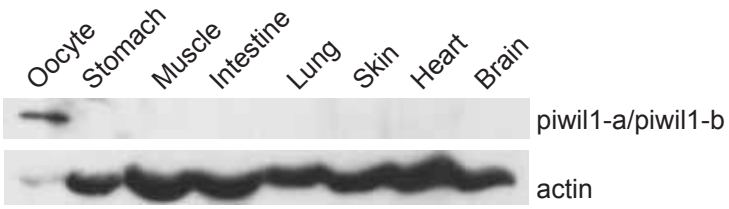
Supplemental Figure 6. (A) piRNA expression throughout oogenesis in *Xenopus tropicalis*. 150 oocyte were used for each experiment. EDC was used to crosslink small RNAs to the membrane prior to hybridization. Expression of miR-148a was assessed to compare with the expression of piRNA. (B) β-elimination was performed to assay for 2′O-methyl-modified 3′ nucleotides of egg piRNAs. 5S rRNA was used as a loading control.

Supplemental Figure 7. (A) Overhang length = top end - bottom start AND bottom end - top start. Values will range from -24 to +24. (B) Overhang lengths were determined for all 20-24nt reads from oocyte libraries.

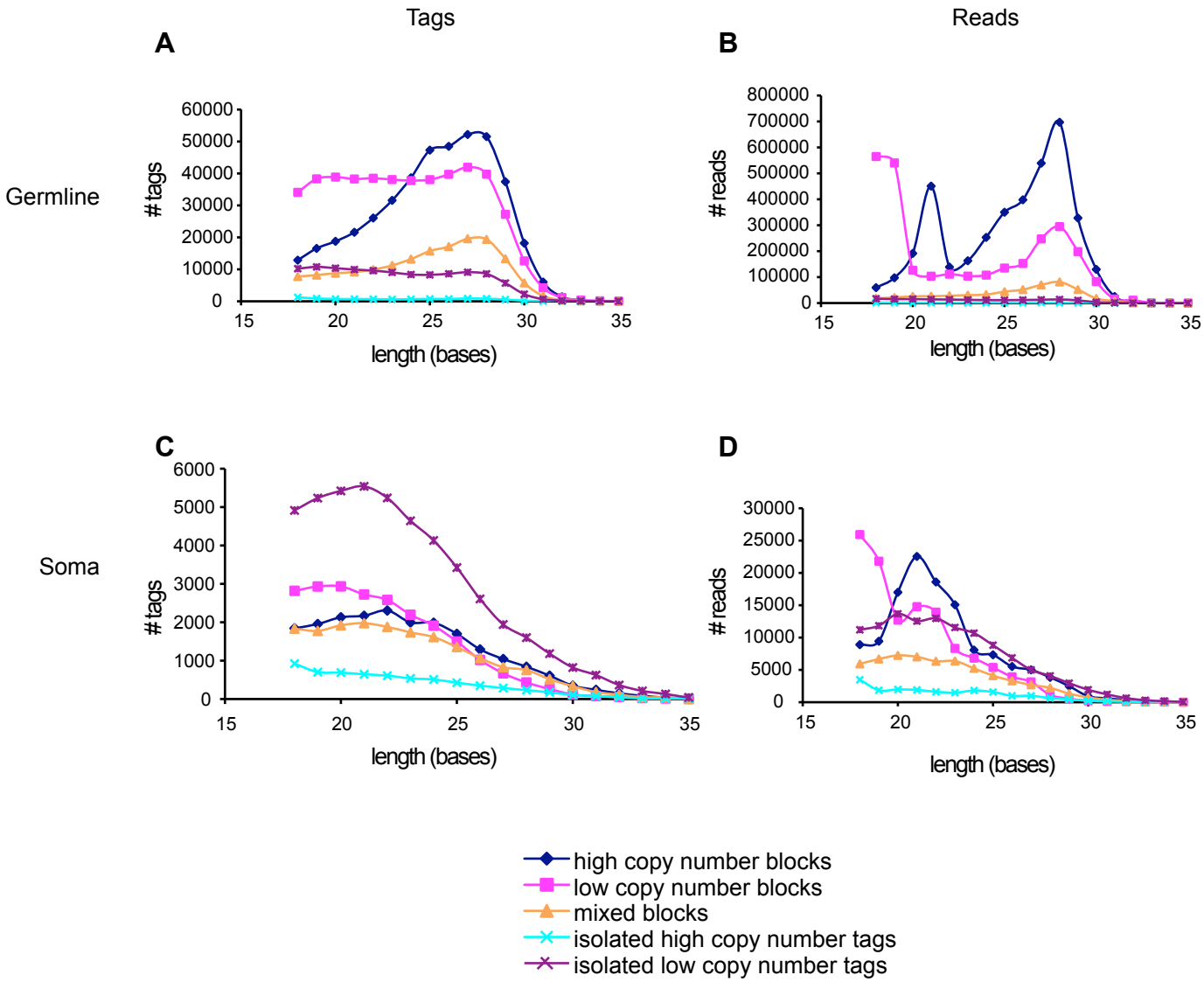
Supplemental Figures



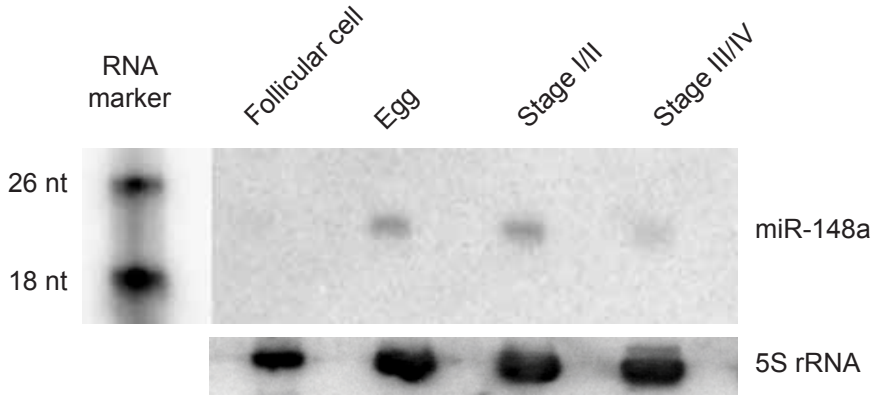
Supplemental Figure 1



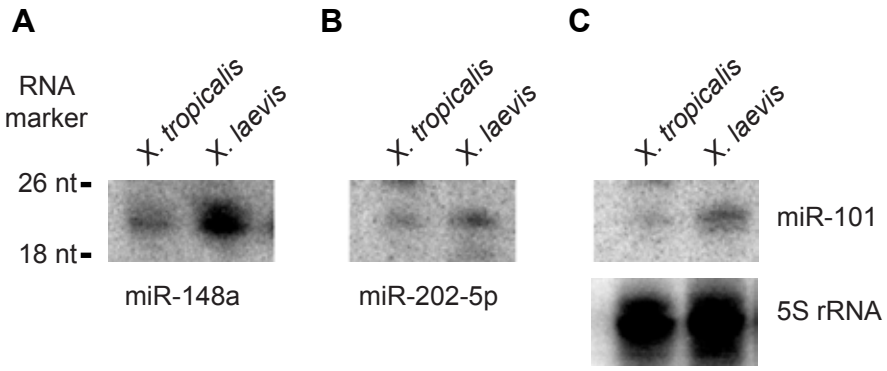
Supplemental Figure 2



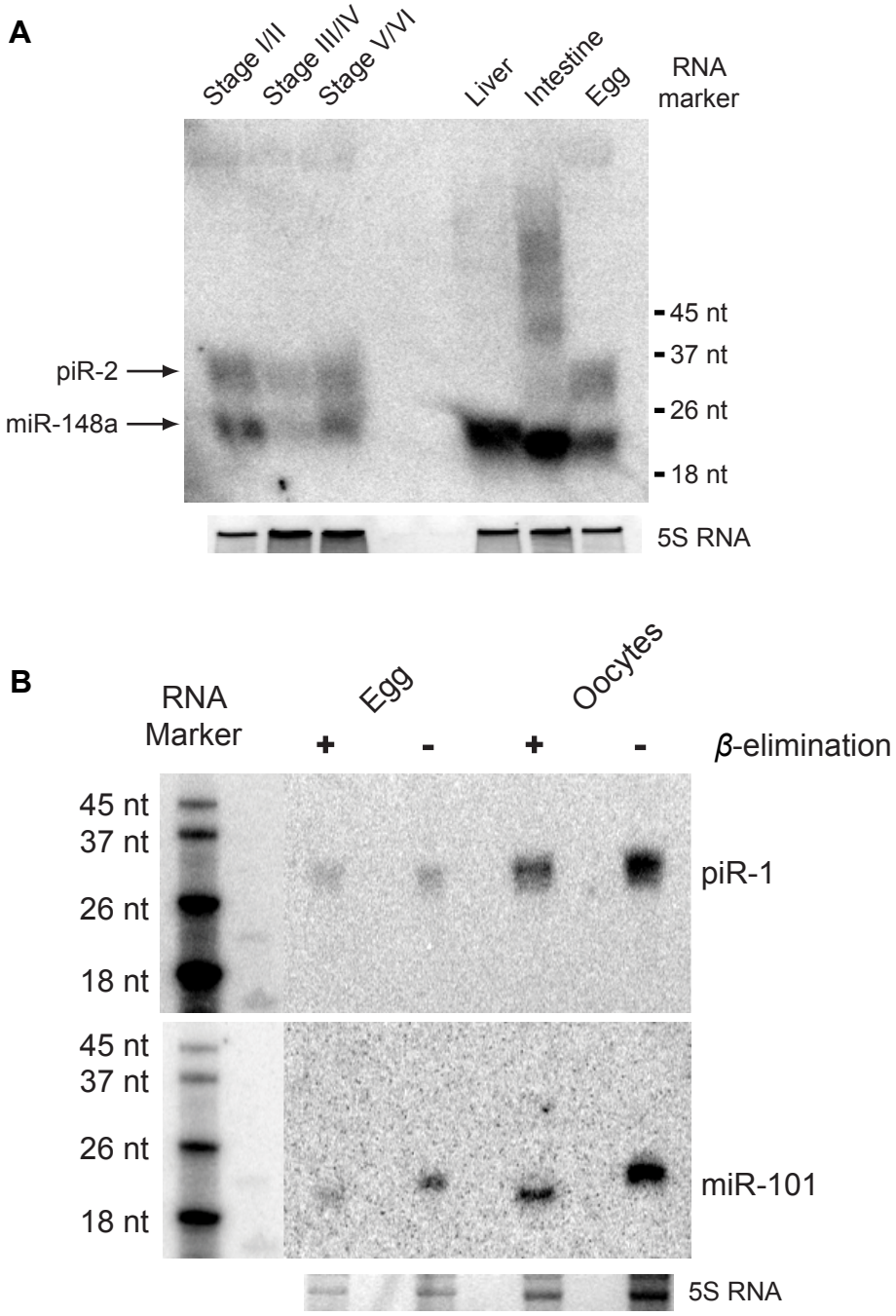
Supplemental Figure 3



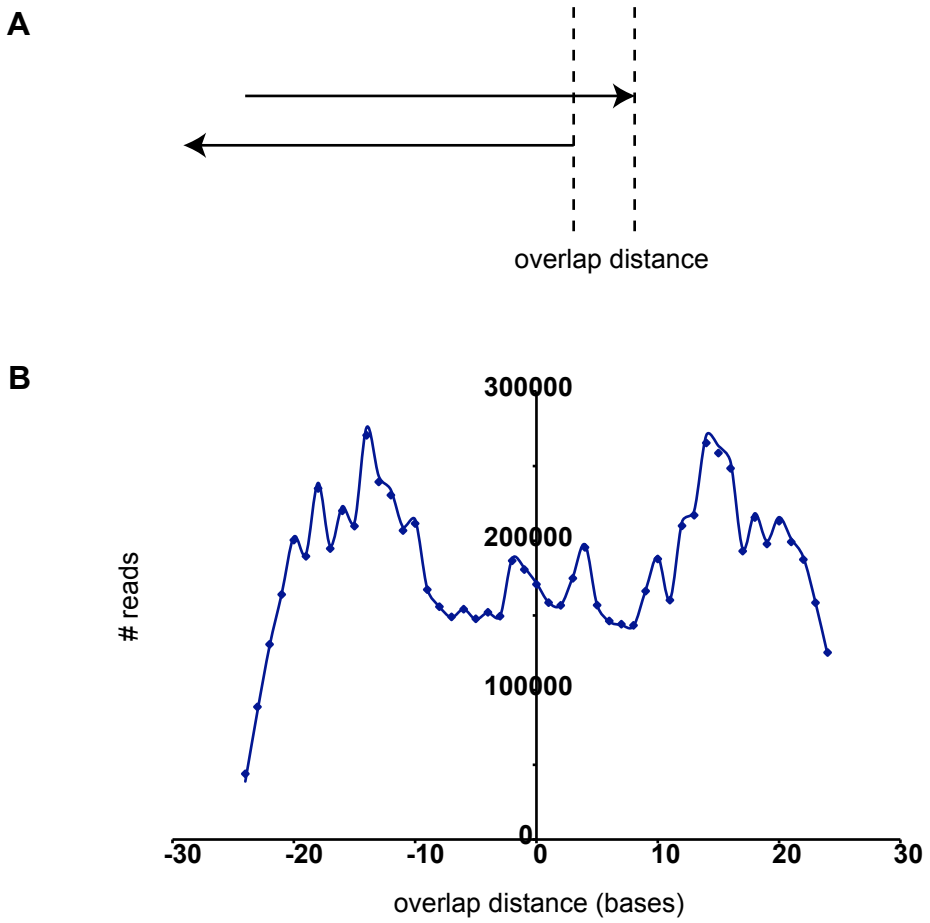
Supplemental Figure 4



Supplemental Figure 5



Supplemental Figure 6



Supplemental Figure 7