doi:10.1038/nature17150

Sequence-dependent but not sequence-specific piRNA adhesion traps mRNAs to the germ plasm

Anastassios Vourekas[1]*, Panagiotis Alexiou[1]*, Nicholas Vrettos[1], Manolis Maragkakis[1] & Zissimos Mourelatos[1]

The conserved Piwi family of proteins and piwi-interacting RNAs (piRNAs) have a central role in genomic stability, which is inextricably linked to germ-cell formation, by forming Piwi ribonucleoproteins (piRNPs) that silence transposable elements[1]. In *Drosophila melanogaster* and other animals, primordial germ-cell specification in the developing embryo is driven by maternal messenger RNAs and proteins that assemble into specialized messenger ribonucleoproteins (mRNPs) localized in the germ (pole) plasm at the posterior of the oocyte[2,3]. Maternal piRNPs, especially those loaded on the Piwi protein Aubergine (Aub), are transmitted to the germ plasm to initiate transposon silencing in the offspring germ line[4–7]. The transport of mRNAs to the oocyte by midoogenesis is an active, microtubule-dependent process[8]; mRNAs necessary for primordial germ-cell formation are enriched in the germ plasm at late oogenesis via a diffusion and entrapment mechanism, the molecular identity of which remains unknown[8,9]. Aub is a central component of germ granule RNPs, which house mRNAs in the germ plasm[10–12], and interactions between Aub and Tudor are essential for the formation of germ granules[13–16]. Here we show that Aub-loaded piRNAs use partial base-pairing characteristics of Argonaute RNPs to bind mRNAs randomly in *Drosophila*, acting as an adhesive trap that captures mRNAs in the germ plasm, in a Tudor-dependent manner. Notably, germ plasm mRNAs in drosophilids are generally longer and more abundant than other mRNAs, suggesting that they provide more target sites for piRNAs to promote their preferential tethering in germ granules. Thus, complexes containing Tudor, Aub piRNPs and mRNAs couple piRNA inheritance with germline specification. Our findings reveal an unexpected function for piRNP complexes in mRNA trapping that may be generally relevant to the function of animal germ granules.

We performed ultraviolet crosslinking followed by stringent immunoprecipitation (CLIP)[17] for Aub (Fig. 1a) and standard small RNA immunoprecipitation, using a highly specific antibody that we generated (Extended Data Fig. 1a) from wild-type (*yw*) ovaries and from *yw* and Tudor-null (*tud*) *Drosophila* embryos collected up to 2 h after laying (0–2-h embryos); this is before zygotic transcription and degradation of maternal mRNAs. Crosslinked RNA–Aub complexes yielded strong, specific signals that were absent from non-immune serum and no-ultraviolet controls (Fig. 1a). CLIP and immunoprecipitation libraries contained essentially identical 23–29-nucleotide piRNAs (Fig. 1b, Extended Data Figs 1b–g and 2a–f and Extended Data Table 1). We verified minimal changes in the piRNA load of Aub in *tud* versus *yw* ovaries[13] (Extended Data Fig. 2g), and found no changes in the piRNA load of 0–2-h embryos compared to ovaries in both genotypes (Extended Data Fig. 2h, i). Larger CLIP tags (lgCLIPs, ≥36 nucleotides) are present in libraries prepared from larger RNP complexes (Fig. 1a–c, Extended Data Fig. 1d and Supplementary Results).

We observe considerable overlap of retrotransposon lgCLIPs with complementary piRNAs (Extended Data Fig. 3a and Supplementary Table 1) and strong positive correlation of their abundances (Extended Data Fig. 3b, c). Relative distance analysis reveals high occurrence of lgCLIPs with a 10-nucleotide overlap to complementary piRNAs (Fig. 1d, peak at position +9) for all three genotypes. The majority of such lgCLIPs bear an adenine at the tenth position (Fig. 1e), and show prominent 5′–5′ end coincidence with Ago3 piRNAs (Fig. 1f), indicating that they correspond to ping-pong intermediate fragments produced by Aub slicing[1]. Furthermore, a second peak at position −15 (Fig. 1d), which is 25 nucleotides (the median Aub piRNA length) from position +9, represents 5′ ends of fragments of trigger piRNA targets undergoing phased piRNA biogenesis[18]. The above results indicate that CLIP captures piRNA biogenesis, complementary retrotransposon targeting and the transient products of Aub slicing activity (Fig. 1g).

A large percentage (∼50–66%) of lgCLIPs from all CLIP libraries are mRNA-derived (Fig. 1c and Extended Data Fig. 1g). Most Aub-bound mRNAs are not substrates for piRNA processing (Extended Data Fig. 4a). The Aub lgCLIP density is relatively high within 3′ untranslated regions (UTRs) compared to RNA sequencing (RNA-seq) analysis, and overall lgCLIP abundance is not correlated with mRNA abundance (Extended Data Fig. 4b–d), suggesting specific target mRNA recognition. We cross-indexed Aub-bound mRNAs with the mRNA localization categories (compiled in ref. 19). Notably, posterior localization categories are significantly enriched in all three sets of Aub CLIP libraries (embryo: *yw* and *tud*, ovary: *yw*) (Supplementary Table 2). Most importantly, we find 15 posterior and germ-cell localization categories significantly depleted, and ubiquitous mRNAs enriched in *tud* embryo compared to *yw* embryo CLIP libraries (Supplementary Table 3). Posteriorly localized mRNAs appear marginally upregulated compared to other localization categories in *tud* versus *yw* embryo RNA-seq libraries (two-sided *t*-test, $P = 0.01594$), ruling out the possibility that the reduced Aub binding is due to reduced posterior mRNA levels in *tud* embryos. Both Aub (Extended Data Fig. 1a) and germ plasm mRNAs[15,20] are uniformly distributed throughout *tud* embryos; therefore, the observed loss of binding specificity towards posterior mRNAs in the absence of Tudor can only be attributed to the disruption of the germ plasm. Thus, our experimental approach allows the identification of the mRNAs specifically bound by Aub in the germ plasm, irrespective of the function of Aub in the clearance of maternal mRNAs in the somatic part of the embryo[21,22]. To identify the primary mRNA targets of Aub within the germ plasm during the formation of germ cells, we calculated the rank product of the normalized lgCLIP values for mRNAs in the 12 posterior localization categories marked with an asterisk in Supplementary Table 3, from three replicate *yw* embryo libraries ($P < 0.05$). The list contains 220 genes, many of which appear enriched or selectively protected in germ cells[10], and with established roles in germ-cell specification and development such as *cycB*, *nos*, *osk*, *gcl*, *pgc* and *Hsp83* (Supplementary Table 4). Characterization of Aub RNPs from early embryos provides independent support for the association of germ plasm mRNAs with Aub (Supplementary Results

[1]Department of Pathology and Laboratory Medicine, Division of Neuropathology, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine; PENN Genome Frontiers Institute, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA.
*These authors contributed equally to this work.

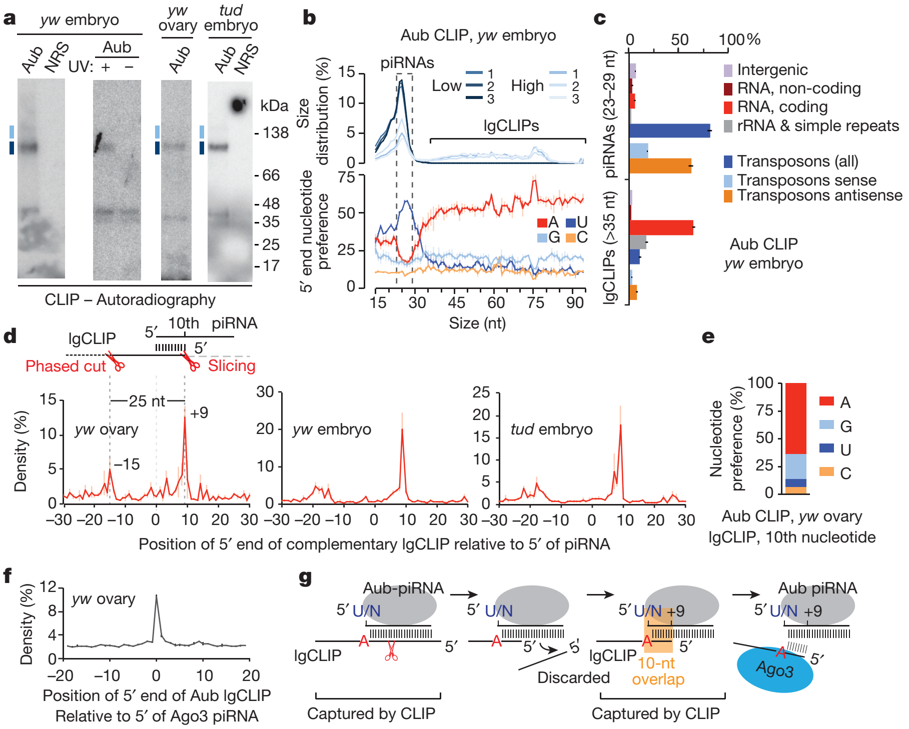


Figure 1 | Transcriptome-wide identification of RNAs bound by Aub and *in vivo* retrotransposon targeting and slicing captured by CLIP. **a**, Aub CLIPs; separate libraries were prepared from RNA extracted from indicated positions. Uncropped gels can be found in Supplementary Fig. 1. kDa, kilodaltons; NRS, non-immune serum; UV, ultraviolet. **b**, Size distribution and 5′ end nucleotide (nt) composition per size of CLIP tag. Error bars represent s.d.; $n = 3$ (biological replicates, the same applies to **c**, **e** and **g**). **c**, Genomic distribution of CLIP tags for three high *yw* embryo (0–2 h) Aub CLIPs. **d**, Position of 5′ ends of retrotransposon lgCLIPs relative to 5′ ends of complementary piRNAs (0, *x* axis). **e**, Nucleotide composition at +9 of retrotransposon-derived lgCLIPs with 10-nucleotide overlap to complementary piRNAs. **f**, *yw* ovary Aub lgCLIP 5′ end positions relative to the 5′ ends of Ago3-loaded piRNAs (0, *x* axis). **g**, Schematic of processing fragments captured by Aub CLIP.

and Extended Data Fig. 5). Four separate analyses provide strong evidence that the extent of the observed Aub binding of mRNAs cannot be explained by piRNA targeting of transposon sequences embedded in mRNAs (Supplementary Results and Extended Data Fig. 6).

To investigate the potential of piRNAs to direct Aub to complementary mRNA sequences further, we analysed chimaeric lgCLIPs[23,24] that each contain an intact piRNA, ligated with a sequence fragment (≥20 nucleotides) that is uniquely aligned on mRNAs (Fig. 2a and Supplementary Table 5). To uncover complementarity patterns, we implemented unweighted local alignment between the piRNA (in reverse complement orientation) and the mRNA fragment, scoring matches (+1), mismatches (−1) and indels (−2), and reporting the best alignment for every chimaeric read. The search was performed within ±100 bases around the midpoint of the mRNA fragment; this allows the identification of the entire complementary sequence that might be missing from the chimaeric fragment, and also provides a reliable estimate of the signal-to-noise ratio. We observed prominent peaks of hundreds of thousands of complementarity events forming around the midpoint and within ±25 nucleotides, in *yw* and *tud* embryo CLIP libraries (Fig. 2b, c). Most events score between 7 and 12; therefore, the

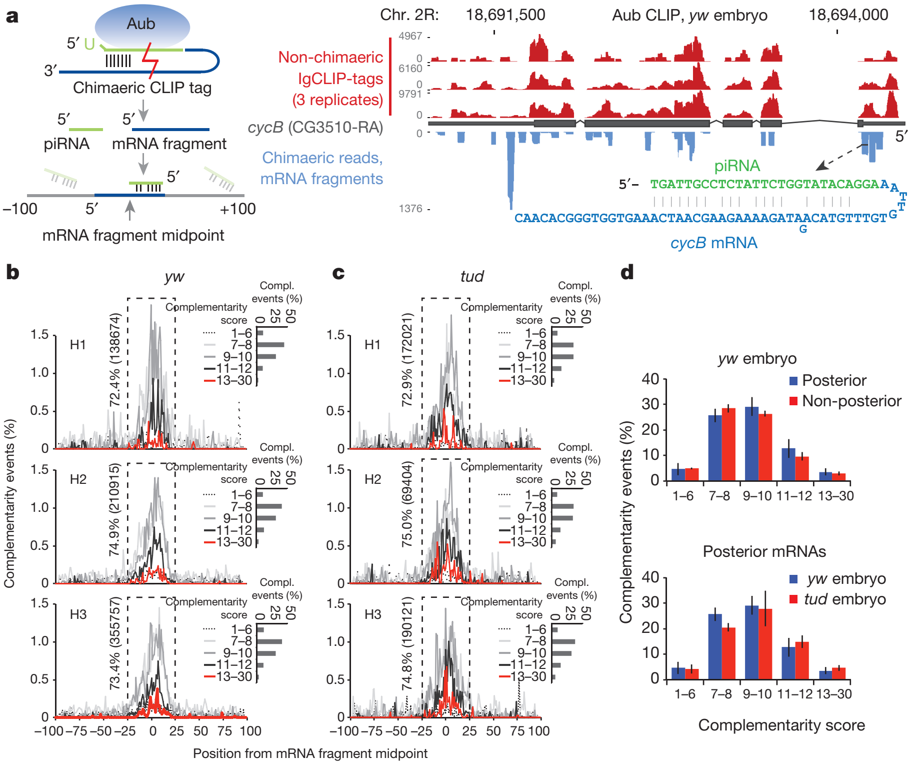


Figure 2 | Complementarity analysis between the piRNA and mRNA parts of chimaeric CLIP tags. **a**, Strategy for chimaeric CLIP tag analysis, and genome browser illustrating Aub lgCLIPs on *cycB*; sequence and base pairing of a chimaeric CLIP tag is shown. Chr, chromosome. **b**, **c**, piRNA–mRNA complementarity events (percentage) within ±100 bases from the midpoint of the mRNA part of the chimaeric read, plotted per alignment score for *yw* (**b**) and *tud* (**c**) embryo Aub CLIPs (biological triplicates). Percentage and number of total events occurring within ±25 bases (dashed rectangles) are shown. Inset, per sample: bar chart of number of complementarity events per score group. **d**, Bar charts of average piRNA–mRNA complementarity events occurring within the ±25-base window of the midpoint of the search area for indicated scores and mRNA localization categories and Aub CLIP libraries. All error bars denote s.d.; $n = 3$.

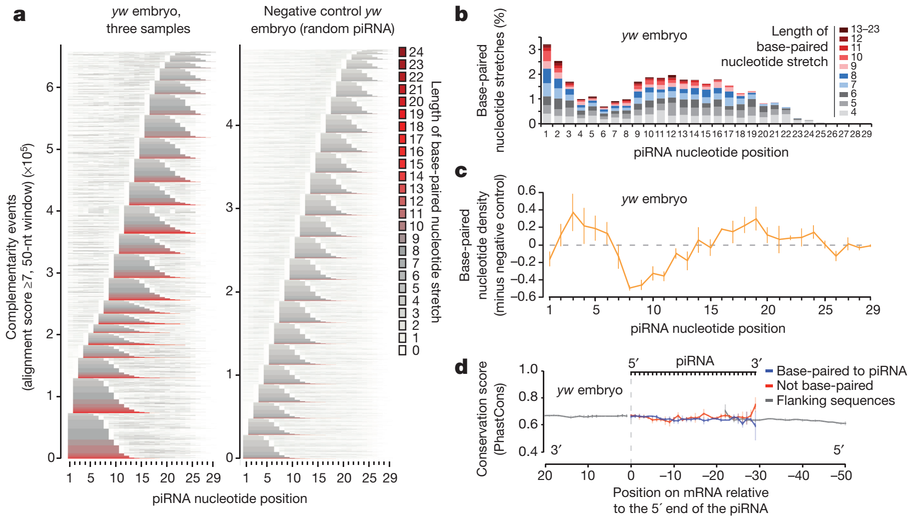


Figure 3 | Characteristics of piRNA base-pairing identified by chimaeric CLIP tag analysis. a, Heat maps showing base-paired nucleotides within the piRNA sequence, for all complementarity events (score ≥7) within ±25-base window, for *yw* embryo and negative control. Stacked piRNAs are sorted (bottom to top) by: starting position and length of the longest stretch and total number of base-paired nucleotides. Every nucleotide position is coloured according to the length of the stretch of consecutively base-paired nucleotides that runs through that position. **b**, Percentage of stretches of consecutive base-paired residues per starting position within the piRNA sequence. **c**, Average base-paired nucleotide density per position minus negative control (random piRNA). **d**, Average mRNA conservation score on and around piRNA–mRNA contact sites. All error bars denote s.d.; $n = 3$.

complementarity is not extensive. The distribution of the complementarity events in the negative control (random piRNA) is completely flat across the search area and has lower scores (Extended Data Fig. 7a), suggesting that the chimaeric reads capture genuine sequence-dependent Aub-piRNA–mRNA contacts.

piRNAs in chimaeric reads are typical Aub piRNAs (Extended Data Fig. 7b–e). piRNA–mRNA complementarities with alignment score ≥7 congregate within a 50-nucleotide window (Fig. 2b–d), so we focused on events that have such scores and locations. piRNA complementarity towards posterior and non-posterior mRNAs is indistinguishable (Fig. 2d and Extended Data Fig. 7f), suggesting that the basis of mRNA binding preference by Aub is not sequence specificity. Chimaeric reads show substantial overlap (Fig. 2a) and the same enrichment in posterior-localized mRNAs with non-chimaeric lgCLIPs (Supplementary Tables 5 and 6), suggesting that they both capture the same RNA binding events.

Base-paired nucleotides for every piRNA from three replicate CLIP libraries are summarized in a comprehensive plot (Fig. 3a and Extended Data Fig. 7g), revealing a bimodal distribution of the complementary regions within the piRNA. Many are found at the 5′ end of the piRNA, starting at positions 1 and 2 (reminiscent of miRNA seed-type binding); additional base-paired stretches start at positions 9–17 (Fig. 3a, b). This pattern is absent from the negative control (Fig. 3a). Net density of base-paired nucleotides reveals a clear preference for piRNAs to use nucleotides at positions 2–6 with additional base pairs in positions 16–24 (Fig. 3c and Extended Data Fig. 7h, i). This profile is markedly similar in *yw* and *tud* libraries, and differs slightly from the miRNA hybridization profile[24] in the less frequent base-pairing in the 2–6 region, suggesting that piRNAs do not use a conserved seed sequence. The periodicity of the graph in Fig. 3c (Extended Data Fig. 7i) evokes the helical conformation and base-pairing availability of the small RNA in the context of an Ago–miRNA–target RNA tripartite complex[25], suggesting that despite the absence of a conserved seed, the mechanics of piRNA complementary binding are analogous to those of microRNAs. Analysis of the evolutionary conservation of paired, unpaired and flanking nucleotides on the mRNA sequence reveals that the piRNA–mRNA contact sites are not preferentially conserved (Fig. 3d).

We used the local alignment approach by which we analysed the chimaeric CLIP tags, to identify potential piRNA target sites in the *D. melanogaster* transcriptome. In 206,400,271 total sites, the vast majority (99.6%) are of scores 7–11 (Fig. 4a). Importantly, the densities of putative piRNA target sites on mRNA regions are essentially identical for mRNAs with or without posterior localization, and very similar to that of the chimaeric mRNA fragments (higher densities in the UTRs compared to the coding sequences; Fig. 4b, c and Extended Data Fig. 8).

mRNAs in the 12 posterior localization categories are significantly longer than non-posterior localized mRNAs[26] (Fig. 4d), and so contain a higher number of piRNA target sites (Fig. 4e); nevertheless, transcript length normalization eliminates this difference (Fig. 4f, g). This holds true when the scores of the predicted sites are accounted for (Fig. 4g), and also when the scores are weighted for the preference of piRNA nucleotides 2–6 and 16–24 to base-pair (not shown). Posterior mRNAs are also more abundant than non-posterior; when factored in, this increases the difference of the target site abundance per transcript for the two localization categories (Fig. 4h). Posterior and non-posterior mRNAs are equally targeted (per kilobase) by each piRNA even when piRNA copy number is accounted for (Extended Data Fig. 9a). Notably, the size differential (and not the absolute length) of posterior and non-posterior mRNAs is conserved among drosophilids: the intra-species size differential always favours posterior mRNAs, although non-posterior mRNAs from one species might be longer than the posterior mRNAs of another (Fig. 4i). Therefore, although piRNAs randomly base pair with non-conserved mRNA sequences, this mechanism is biased towards a specific class of mRNAs for germ plasm anchoring. Additionally, from the two categories of posterior localized mRNAs, 'localized' and 'protected'[10], localized mRNAs have longer 3′ UTRs than protected mRNAs, further supporting the notion that mRNA length positively affects germ plasm enrichment (Extended Data Fig. 9b, c).

The concept of mRNA entrapment at the germ plasm during ooplasmic streaming is well established[8,9,27], but the mechanism at the molecular level has so far been elusive. We propose that germ plasm localized Tud–Aub–piRNA complexes play the role of a non-discriminatory adhesive trap that can form numerous, non-conserved piRNA–mRNA contacts to capture mRNAs and form germ plasm mRNPs (Fig. 4j and Supplementary Discussion). This mechanism probably shows preference for posterior mRNAs because they are

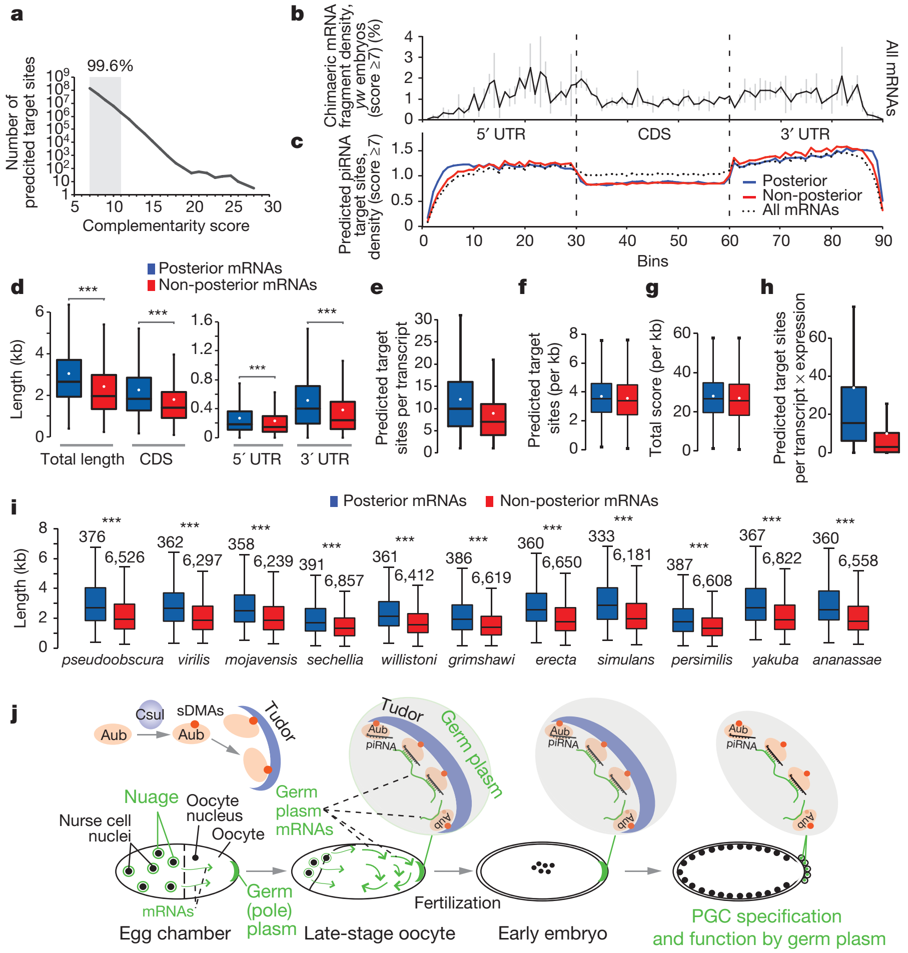


Figure 4 | Transcriptome-wide prediction of piRNA target sites and length differential of posterior-localized mRNAs. a, Number of predicted piRNA complementary sites on mRNAs, per score. **b**, **c**, Average binned density of: chimaeric mRNA fragments (Aub CLIP, *yw* embryo 0–2 h) along the meta-mRNA. CDS, coding sequence. Error bars denote s.d.; $n = 3$ (**b**); predicted piRNA complementary sites within all (14,058), posterior (380), and non-posterior (6,747) localized mRNAs (**c**). **d**–**i**, Box-and-whisker plots of: lengths of mRNAs expressed in *yw* embryos (0–2 h) (**d**); number of predicted piRNA complementary sites per mRNA (**e**); length-normalized number of predicted piRNA complementary sites (**f**); length-normalized total score of predicted piRNA complementary sites (**g**); number of predicted piRNA complementary sites per mRNA multiplied by the abundance of each mRNA RPKM (reads per kilobase per million mapped reads) (**h**); and lengths of orthologous mRNAs in other *Drosophila* species (**i**). Black lines denote median; white dots denote mean. $***P < 0.005$, one-sided *t*-test (**d**); $***P < 1 \times 10^{-16}$, one-sided Wilcoxon exact rank test (**i**). **j**, Aub couples piRNA inheritance with germ-cell specification in *Drosophila*. Aub, carrying symmetrically dimethylated arginine residues (sDMAs) dimethylated by Csul methyltransferase, interacts with Tudor, and both are localized in the germ plasm during mid-stage oogenesis. Ooplasmic streaming at later stages promotes diffusion of mRNPs, facilitating random contacts of mRNAs with the germ plasm. Aub piRNAs form an adhesive trap that captures mRNAs forming numerous low complementarity contacts. mRNAs with posterior functions are longer and more abundant than the rest, form more piRNA-mediated contacts with the germ plasm, and thus their entrapment is enhanced. Tudor–Aub-piRNA–mRNA complexes along with other RNA binding proteins form germ granules that contain both piRNAs and mRNAs that induce primordial germ cell (PGC) specification. Aub and its RNA cargo is incorporated in PGCs, providing the maternal mRNAs that are necessary for PGC function and the maternal piRNAs that will propagate an RNA immune response against transposons.

significantly longer and more abundant[26]. We believe that the above mechanism acts in addition to specific protein–protein, protein–RNA and RNA–RNA interactions that are necessary for mRNA transfer and anchoring to the posterior, and for translational control[10,12,28–30]. The multivalence of Aub–Tudor interactions probably contributes to the formation of multimeric germ granule complexes. We propose that germ-cell specification and function by maternal mRNAs, and piRNA inheritance converge in Aub. Coupling germ-cell specification with piRNA inheritance could be a strategy that increases reproductive fitness by ensuring the propagation of robust transposon silencing mechanisms to germ cells across generations and across the population.

Online Content Methods, along with any additional Extended Data display items and Source Data, are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 7 September 2015; accepted 13 January 2016.
Published online 7 March 2016.

1. Siomi, M. C., Sato, K., Pezic, D. & Aravin, A. A. PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Rev. Mol. Cell Biol.* **12,** 246–258 (2011).
2. Ephrussi, A. & Lehmann, R. Induction of germ cell formation by oskar. *Nature* **358,** 387–392 (1992).
3. Mahowald, A. P. Assembly of the *Drosophila* germ plasm. *Int. Rev. Cytol.* **203,** 187–213 (2001).
4. Brennecke, J. *et al.* An epigenetic role for maternally inherited piRNAs in transposon silencing. *Science* **322,** 1387–1392 (2008).
5. Grentzinger, T. *et al.* piRNA-mediated transgenerational inheritance of an acquired trait. *Genome Res.* **22,** 1877–1888 (2012).
6. Khurana, J. S. *et al.* Adaptation to P element transposon invasion in *Drosophila melanogaster*. *Cell* **147,** 1551–1563 (2011).
7. Bucheton, A. Non-Mendelian female sterility in *Drosophila melanogaster*: influence of aging and thermic treatments. III. Cumulative effects induced by these factors. *Genetics* **93,** 131–142 (1979).

8. Kugler, J. M. & Lasko, P. Localization, anchoring and translational control of oskar, gurken, bicoid and nanos mRNA during *Drosophila* oogenesis. *Fly (Austin)* **3,** 15–28 (2009).
9. Forrest, K. M. & Gavis, E. R. Live imaging of endogenous RNA reveals a diffusion and entrapment mechanism for *nanos* mRNA localization in *Drosophila*. *Curr. Biol.* **13,** 1159–1168 (2003).
10. Rangan, P. *et al.* Temporal and spatial control of germ-plasm RNAs. *Curr. Biol.* **19,** 72–77 (2009).
11. Thomson, T., Liu, N., Arkov, A., Lehmann, R. & Lasko, P. Isolation of new polar granule components in *Drosophila* reveals P body and ER associated proteins. *Mech. Dev.* **125,** 865–873 (2008).
12. Trcek, T. *et al. Drosophila* germ granules are structured and contain homotypic mRNA clusters. *Nature Commun*. **6,** 7962 (2015).
13. Kirino, Y. *et al.* Arginine methylation of Aubergine mediates Tudor binding and germ plasm localization. *RNA* **16,** 70–78 (2010).
14. Liu, H. *et al.* Structural basis for methylarginine-dependent recognition of Aubergine by Tudor. *Genes Dev.* **24,** 1876–1881 (2010).
15. Arkov, A. L., Wang, J.-Y. S., Ramos, A. & Lehmann, R. The role of Tudor domains in germline development and polar granule architecture. *Development* **133,** 4053–4062 (2006).
16. Boswell, R. E. & Mahowald, A. P. *tudor*, a gene required for assembly of the germ plasm in *Drosophila melanogaster*. *Cell* **43,** 97–104 (1985).
17. Vourekas, A. *et al.* Mili and Miwi target RNA repertoire reveals piRNA biogenesis and function of Miwi in spermiogenesis. *Nature Struct. Mol. Biol*. **19,** 773–781 (2012).
18. Mohn, F., Handler, D. & Brennecke, J. piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science* **348,** 812–817 (2015).
19. Lécuyer, E. *et al.* Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* **131,** 174–187 (2007).
20. Thomson, T. & Lasko, P. *Drosophila tudor* is essential for polar granule assembly and pole cell specification, but not for posterior patterning. *Genesis* **40,** 164–170 (2004).
21. Barckmann, B. *et al.* Aubergine iCLIP reveals piRNA-dependent decay of mRNAs involved in germ cell development in the early embryo. *Cell Rep.* **12,** 1205–1216 (2015).
22. Rouget, C. *et al.* Maternal mRNA deadenylation and decay by the piRNA pathway in the early *Drosophila* embryo. *Nature* **467,** 1128–1132 (2010).
23. Moore, M. J. *et al.* miRNA-target chimeras reveal miRNA 3′-end pairing as a major determinant of Argonaute target specificity. *Nature Commun*. **6,** 8864 (2015).
24. Grosswendt, S. *et al.* Unambiguous identification of miRNA: target site interactions by different types of ligation reactions. *Mol. Cell* **54,** 1042–1054 (2014).
25. Schirle, N. T., Sheu-Gruttadauria, J. & MacRae, I. J. Structural basis for microRNA targeting. *Science* **346,** 608–613 (2014).
26. Jambor, H. *et al.* Systematic imaging reveals features and changing localization of mRNAs in *Drosophila* development. *eLife* **4,** e05003 (2015).
27. Sinsimer, K. S., Lee, J. J., Thiberge, S. Y. & Gavis, E. R. Germ plasm anchoring is a dynamic state that requires persistent trafficking. *Cell Rep*. **5,** 1169–1177 (2013).
28. Little, S. C., Sinsimer, K. S., Lee, J. J., Wieschaus, E. F. & Gavis, E. R. Independent and coordinate trafficking of *Drosophila* germ plasm mRNAs. *Nature Cell Biol.* **17,** 558–568 (2015).
29. Ghosh, S., Marchand, V., Gáspár, I. & Ephrussi, A. Control of RNP motility and localization by a splicing-dependent structure in oskar mRNA. *Nature Struct. Mol. Biol*. **19,** 441–449 (2012).
30. Gavis, E. R., Lunsford, L., Bergsten, S. E. & Lehmann, R. A conserved 90 nucleotide element mediates translational repression of nanos RNA. *Development* **122,** 2791–2800 (1996).

Supplementary Information is available in the online version of the paper.

Acknowledgements We thank former and current laboratory members for discussions; M. Siomi for the Tudor antibody; A. Arkov for *tud* flies; G. Dreyfuss for the PABP antibody; and J. Schug for Illumina sequencing. Work was supported by a Brody family fellowship to M.M., and a National Institutes of Health (NIH) grant GM072777 to Z.M.

Author Contributions A.V. and Z.M. conceived, and Z.M. supervised, the study. A.V. and N.V. performed the experiments. P.A. performed bioinformatic analyses with contribution from M.M. and A.V. A.V., P.A., N.V., M.M. and Z.M. interpreted the data. A.V. wrote the manuscript, with contribution from all authors.

Author Information Sequences have been deposited in the Sequence Read Archive (SRA) under the accession number SRP067739. Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to Z.M. (mourelaz@uphs.upenn.edu).

METHODS

***Drosophila* strains, tissue collection.** The following strains and heteroallelic combinations were used: y^1w^{1118} as the wild-type stock (*yw*), $aub^{HN2/QC42}$ (*aub*) and $tud^{1/Df(2R)PurP133}$ (*tud*), for *aub* and *tud* mutants (loss-of-function), respectively[15,31–33]. All flies were grown at 25 °C with 70% relative humidity on a 12-h light–dark cycle. The 2–4-day female flies were crossed to *yw* males for 2 days in standard cornmeal food supplied with yeast paste before ovary dissection. Embryos collected at well-defined time-windows were dechorionated in 50% commercial bleach for 2 min, washed extensively in water and collected in PBS or HBSS or fixation solution, depending on downstream applications.

Antibodies. Antibody against Aubergine (Aub-83) was produced by immunizing rabbits with Aub peptide (HKSEGDPRGSVRGRC, in which terminal cysteine was used to couple to KLH; Genscript) and selected with peptide-affinity purification of sera. Other antibodies that were used in this study: mouse monoclonal anti-PABP (6E2 clone)[34], E7 mouse monoclonal anti-β-tubulin (Developmental Studies Hybridoma Bank) and anti-Tudor mouse monoclonal (gift from M. Siomi).

Immunofluorescence. Fixation and immunohistochemistry of dissected ovaries and embryos was performed according to standard protocols. Primary antibodies against Aub and Tud were used at 1 ng μl^{-1} final concentration. Secondary antibodies conjugated to Alexa 488 and 594 (Life technologies) were used at 1:1,000 dilution. Ovary and embryo samples were imaged on Leica TCS SPE confocal microscope.

Aub CLIP-seq (HITS-CLIP, high-throughput sequencing after crosslinking and immunoprecipitation). CLIP was performed as previously described for Mili, Miwi and MOV10L1 (refs 17, 35, 36). The protocol is described in detail previously[36] and uses stringent buffer conditions to ensure high specificity. The experiment was performed in three biological replicates for each condition (*yw* ovaries, *yw* embryos 0–2 h, *tud* embryos 0–2 h). Approximately 40 mg of *Drosophila* embryos (0–2 h) or ~80 ovaries from 4–6-day females were collected in ice-cold HBSS and ultraviolet-irradiated (3×) at 254 nm (400 mJ cm^{-2}). The tissues were pelleted, washed with PBS and the final tissue pellet was flash-frozen in liquid nitrogen and kept at −80 °C. Ultraviolet-light-treated tissues were lysed in 350 μl PMPG (PBS (no Mg^{2+} and no Ca^{2+}), 2% Empigen) with protease inhibitors and rRNasin (2 U μl^{-1}) and no exogenous RNases; lysates were treated with DNase I (Promega) for 5 min at 37 °C, and then were centrifuged at 100,000*g* for 30 min at 4 °C.

For each immunoprecipitation, approximately 10 μl of our anti-Aub antibody was bound on 150 μl (slurry) of protein A Dynabeads in Ab binding buffer (0.1 M Na-phosphate, pH 8, and 0.1% NP-40) at room temperature for 2 h; antibody-bound beads were washed three times with PMPG. Antibody beads were incubated with lysates (supernatant of 100,000*g*) for 3 h at 4 °C. Low- and high-salt washes of immunoprecipitation beads were performed with 1× and 5× PMPG (5× PBS, 2% Empigen). RNA linkers (RL3 and RL5), as well as 3′ adaptor labelling and ligation to CIP (calf intestinal phosphatase)-treated RNA CLIP tags were performed as previously described[36].

Immunoprecipitation beads were eluted at 70 °C for 12 min using 30 μl of 2× Novex reducing loading buffer. Samples were analysed by NuPAGE (4–12% gradient precast gels, run with MOPS buffer). Cross-linked RNA–protein complexes were transferred onto nitrocellulose (Invitrogen), and the membrane was exposed to film for 1–2 h. Membrane fragments containing the main radioactive signal and fragments up to ~15 kDa higher were excised (Fig. 1a). RNA extraction, 5′ linker ligation, Reverse transcriptase PCR (RT–PCR) and a second PCR step were performed with the DNA primers (DP3 and DP5, DSFP3 and DSFP5) as described previously[36]. Complementary DNA from two PCR steps was resolved on and extracted from 3% Metaphor 1× TAE gels. Size profiles of cDNA libraries prepared from the main radioactive signal and higher molecular mass signal were similar (Fig. 1a). DNA was extracted with QIAquick Gel Extraction kit and submitted for deep sequencing. The cDNA libraries were sequenced with Hi-Seq Illumina at 100 cycles.

Solid-support directional RNA-seq. Solid-support directional RNA-seq was performed as previously described[17], using total RNA (depleted of ribosomal RNA with Ribo-Zero (EpiCentre)) isolated from 0–2-h embryos of appropriate genotypes.

Nycodenz density gradient ultracentrifugation and subsequent analyses. Nycodenz density gradient separation of RNPs was performed as previously described[17] with modifications. A 20–60% (top to bottom) Nycodenz gradient (4.8 ml) in 1× KMH150 (150 mM KCl, 2 mM $MgCl_2$, 20 mM HEPES, pH 7.4, 0.5% NP-40, 0.1 U μl^{-1} rRNasin, and protease inhibitors) was prepared as a step gradient by overlaying five equal parts of Nycodenz solutions and was left to diffuse overnight at 4 °C. 0.2 microlitres of post-nuclear *yw* embryo lysate in 1× KMH150 was laid over the gradient and centrifuged at 150,000*g* for 20 h. We used embryos of stages 4–6, to avoid earlier stages where mRNAs at the soma form distinct mRNPs than the ones formed in the pole plasm PGCs. The gradient was collected in 12 equal fractions. Samples from each fraction were used for protein determination by Bradford and RNA extraction with Trizol LS. Right before RNA extraction, 500 ng of *in vitro* transcript of *Renilla* luciferase mRNA was spiked in each fraction for normalization purposes in subsequent steps.

qRT–PCR. An equal volume of RNA extracted from each fraction was reverse transcribed by Supersript III (Invitrogen 18080-051) in the presence of random hexamers. Equal volume of the cDNA was mixed with primers (*gcl*, *osk*, *Hsp83*, *dhd*, *cycB*: Qiagen QuantiTect Assay; *Renilla* luciferase (rLuc), forward: 5′-CGCTGAAAGTGTAGTAGATGTG-3′ and reverse: 5′-TCCACGAAGAAGTTATTCTCCA-3′) and Power SYBR Green reaction mix (Applied Biosystems 4367659). The reactions were run on a StepOnePlus System (Applied Biosystems) using the default program.

Immunoprecipitation and detection of piRNAs, and preparation of cDNA libraries. Aub immunoprecipitation, 5′ end labelling of piRNAs and cDNA library preparation were carried out as previously described[37,38].

Code availability. We used CLIPSeqTools[39], a bioinformatics suite that we created for analysis of CLIP-seq data sets (accessible at: http://mourelatos.med.upenn.edu/clipseqtools/) and a Perl programming framework that we developed[40]. The latter framework is named GenOO and has been specifically developed for analysis of high-throughput sequencing data. The source code for GenOO has been deposited in GitHub and can be accessed at https://github.com/genoo/.

Statistics. In statistical analyses, we ensured that the assumptions of each statistical test are met and that the statistical test used is appropriate for the analysis. In all analyses the statistical tests and methods used are clearly stated in relevant sections. No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Data. *Drosophila* (assembly dm3) transcript, exon and repeat genomic locations were downloaded from the UCSC genome browser (downloaded 22 March 2011 from http://genome.ucsc.edu). Repeat consensus sequences were downloaded from Flybase (http://flybase.org/ - transposon_sequence_set v9.42). Localization categories for *Drosophila* genes were taken from ref. 19. The localization annotation matrix was downloaded from (http://fly-fish.ccbr.utoronto.ca/annotation_matrix.csv). Transposon categories were as in ref. 31.

Preprocessing. The 3′ end ligated adaptor (GTGTCAGTCACTTCCAGCGGTCGTATGCCGTCTTCTGCTTG) was removed from the sequences using the cutadapt software and a 0.25 acceptable error rate for the alignment of the adaptor on the read. To eliminate reads in which the adaptor was ligated more than one time, adaptor removal was performed three times.

Alignment. Reads for all samples were aligned against the dm3 *Drosophila melanogaster* genome assembly using the aligner bwa v0.6.2-r126, with the default settings[41]. Reads were also aligned against the Repeat consensus sequences using the same aligner.

Genomic distribution. All mapped reads were divided in the following genomic categories: sense repeat, antisense repeat, non-coding RNA, (protein) coding RNA. The remaining reads were considered to be intergenic reads.

Correlation of replicates. Gene expression was defined as the number of reads that map on each gene and the values were normalized by the upper quartile normalization method[42]. The $\log_2$ gene expression levels of replicates are compared using the Pearson Correlation function in R.

Coincidence with immunoprecipitation. Reads mapping in the same position (same 5′ end mapping) were considered as coinciding. When comparing CLIP with immunoprecipitation libraries, the percentage of piRNA-size CLIP reads that had a coinciding start with any standard immunoprecipitation read were counted as positive.

Significant localization. For each localization category, the quartile-normalized lgCLIP binding level ('mRNA expression level' in each CLIP library) is compared via a two-sided *t*-test between genes that belong to the category versus genes that do not belong to it. To compare two samples, we measure the difference in binding (per gene) between the two conditions ($\log_2$(gene.expr.cond1/gene.expr.cond2)) and then perform a *t*-test of differences in genes belonging to the category versus genes not belonging in the category.

Early embryo posterior localization categories. The following twelve mRNA localization categories[19] were found significantly depleted in *tud* embryo Aub CLIP libraries compared to *yw* embryo libraries, and were used in analyses were 'posterior localized mRNAs' are mentioned: '1:41:RNA islands', '1:42:Pole buds', '1:40:Pole plasm', '3:265:Perinuclear around pole cell nuclei', '4:370:Germ cell localization', '4:403:Germ cell enrichment', '3:348:Pole cell enrichment', '2:141:Pole cell localization', '2:153:Perinuclear around pole cell nuclei', '2:142:Pole cell enrichment', '3:347:Pole cell localization', '1:59:Perinuclear around pole cell nuclei' (http://fly-fish.ccbr.utoronto.ca/). The remaining mRNAs are mentioned as non-posterior localized mRNAs. The following three posterior localization categories were also depleted in *tud* embryo Aub CLIP libraries compared to *yw*: '1:39:Posterior localization', '2:124:Posterior localization', '3:352:Posterior localization'. Almost all of the mRNAs contained in the above twelve categories are also contained in these three, but these three categories also contain some mRNAs that do not actually localize in the pole plasm or the germ cells (that is, with apical localization); therefore, mRNAs belonging in any of these three localization categories but not in any of the above mentioned twelve posterior categories were not considered for the generation of the Supplementary Table 4. Many mRNAs do not have a designated localization pattern, and they are mentioned as 'undetermined localization'. It is worth mentioning that this category contains several mRNAs with clear posterior–pole-plasm localization. Through manual searches of the Berkeley *Drosophila* Genome Project chromogenic ISH database (http://insitu.fruitfly.org/cgi-bin/ex/insitu.pl) we noticed that many Aub-bound mRNAs, the localization of which is not annotated in the Fly-FISH database, are indeed localized in the germ plasm/cells (such as *CG4735/shu*, *CG7070/PyK*, *CG4903/MESR4*, *CG5452/dnk* and *CG9429/Calr*), therefore our analysis is most likely underestimating the true number of Aub-bound mRNAs that are important for germline specification and function. Because of this, mRNAs with 'undetermined localization' were never mixed with 'non-posterior localized' mRNAs in our analyses.

Highly bound genes. To identify highly bound genes, we used the rank product method[43]. Specifically, genes are sorted by expression per sample, and for each gene the product of their ranks is calculated. The probability of this rank product produced by chance is calculated by permutations of all non-zero value genes.

Transcript expression calculation. We calculated the expression for protein-coding transcripts by counting the number of RNA-seq reads that map within the exons of each transcript. The counts were normalized using RPKM and upper quartile normalization, effectively dividing each count by the upper quartile of all counts[42]. The transcript with the highest RPKM score was used ('best transcript') unless otherwise noted.

Transcript Aub-binding calculation. We calculated the expression for protein-coding transcripts by counting the number of CLIP reads that map within the exons of each transcript in the sense orientation. The counts were normalized using reads per million and upper quartile normalization, effectively dividing each count by the upper quartile of all counts[42].

RNA-seq correlation versus CLIP. Upper quartile normalized RPKM for RNA-seq was compared to similarly normalized CLIP binding levels defined as average number of reads per transcript in CLIP replicates. Correlation was calculated using the Pearson Correlation function in R.

Identification of hybrid reads. (1) Identified lgCLIP size reads (read length >35) that did not align to the genome. (2) Made a set of substrings from both ends of reads from (1) of piRNA size ($L=[23,29]$). (3) Identified the substring from (2) to full-length piRNAs ($L=[23,29]$) from corresponding low samples (Extended Data Fig. 1b) (4) The longest aligning piRNAs are retained and coupled with the remainder of the read as piRNA–lgCLIP couples. (5) The piRNA aligning fragment is cut from the read. Very small remainder reads ($L=[<20]$) are discarded. (6) The remainders are aligned to the genome (using bwa default settings). (7) Remainders aligned in one single position that is on a known mRNA are retained.

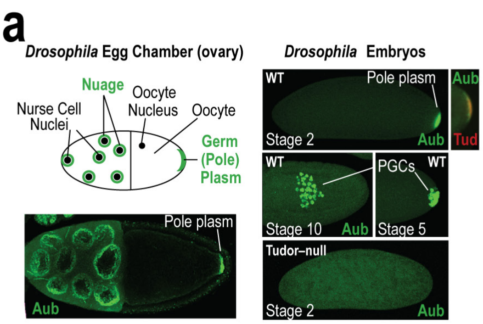
Alignment of piRNAs to regions. (1) Regions of 200-nucleotide length were cut around the midpoint of the genomic alignment region from step 7 of previous routine. Specifically, if ($d=200$ the length of the final region we want and L is the length of the read), a genomic region flanking the read on each side of length $d/2$ was excised from the chromosome sequence. If the alignment was located in the minus strand the sequence was reversed and complemented at this point. This total region has length $d+L$. We discard an equal number of nucleotides from each side to reach a final length of L (specifically we substring starting from int($L/2$) and for d nucleotides. Note, int will always round down). At this point we have a region of length 200 nucleotides centred around the alignment region of the fragment. (2) We use a slightly modified Smith–Waterman[44] alignment method (weights: match $=+1$, mismatch $=-1$, gap $=-2$) to align piRNAs on the 200-nucleotide long regions from (1). Differences of our alignment versus Smith–Waterman: (a) No penalties are given to non-matching nucleotides on the edges of the alignment. (b) If there are multiple optimal alignment scores, one is picked randomly. (c) Alignments in which part of one sequence is outside the boundaries of the other sequence are not considered. (3) The midpoint of the alignment (if k nucleotides matched that is the int($k/2$) nucleotide) is used for graphs of alignment positioning on regions.

mRNA target prediction for the top 2,000 expressed piRNAs. We grouped piRNA sequences into families based on the first 23 nucleotides of each piRNA. Using the alignment algorithm described above we aligned one piRNA (the most abundant) for each of the top 2,000 families to the longest annotated transcript for each protein-coding gene. These 2,000 piRNA families represent ~37% of piRNA reads from low *yw* CLIP libraries. To factor in transcript abundance, we multiplied the RNA-seq (*yw* 0–2-h embryo) RPKM value for each mRNA with the number of predicted piRNA target sites found within the mRNA. This provides a 'targeting potential' of every mRNA species, corrected for its abundance.

We then evaluated the targeting potential of each piRNA–mRNA pair using three different scoring schemes. For the first, we sum the alignment score of all putative piRNA binding sites on the mRNA. For the second, we calculated a weighted alignment score for each putative piRNA binding site and then we sum all scores similar to the previous scheme. The weighted score for each binding site is calculated based on the following formula $\sum_i x_i * A_i$, in which x_i is 1 or 0 based on whether the nucleotide at position i of the piRNA is bound or not, and A_i is the weight for nucleotide i. For the third, we multiplied the total number of predicted complementary sites per piRNA, with the piRNA copy number.

Study of the lengths of *D. melanogaster* orthologous mRNAs in other *Drosophila* species. Transcript sequences (fasta file) for each species were downloaded from Flybase (ftp://ftp.flybase.net/genomes/ on 1 September 2015, current version used for each genome). For each gene (identified as the 'parent' tag in the fasta file header), the longest transcript length was identified. For the analysis of the expressed mRNAs (Fig. 4d), we used our *yw* embryo RNA-seq data to identify the longest transcript with the highest length normalized abundance. Orthologue gene tables were downloaded from Flybase (gene_orthologs_fb_2015_03.tsv.gz) and were used to identify orthologue genes across species. For each species, all genes that mapped to localized and unlocalized *Drosophila melanogaster* genes were used in the comparison and were assigned to the corresponding group as their *D. melanogaster* orthologue. Boxplots were created using the lattice package in R (bwplot) and omitting outliers, P values were calculated using the Wilcoxon exact rank test (wilcox.test in R) one-sided with the hypothesis that localized genes are longer than non-localized.

31. Malone, C. D. *et al.* Specialized piRNA pathways act in germline and somatic tissues of the *Drosophila* ovary. *Cell* **137,** 522–535 (2009).
32. Wilson, J. E., Connell, J. E. & Macdonald, P. M. *aubergine* enhances *oskar* translation in the *Drosophila* ovary. *Development* **122,** 1631–1639 (1996).
33. Schupbach, T. & Wieschaus, E. Female sterile mutations on the second chromosome of *Drosophila melanogaster*. II. Mutations blocking oogenesis or altering egg morphology. *Genetics* **129,** 1119–1136 (1991).
34. Matunis, M. J., Matunis, E. L. & Dreyfuss, G. Isolation of hnRNP complexes from *Drosophila melanogaster*. *J. Cell Biol.* **116,** 245–255 (1992).
35. Vourekas, A. *et al.* The RNA helicase MOV10L1 binds piRNA precursors to initiate piRNA processing. *Genes Dev.* **29,** 617–629 (2015).
36. Vourekas, A. & Mourelatos, Z. HITS-CLIP (CLIP-Seq) for mouse Piwi proteins. *Methods Mol. Biol.* **1093,** 73–95 (2014).
37. Kirino, Y., Vourekas, A., Khandros, E. & Mourelatos, Z. Immunoprecipitation of piRNPs and directional, next generation sequencing of piRNAs. *Methods Mol. Biol.* **725,** 281–293 (2011).
38. Kirino, Y. *et al.* Arginine methylation of Piwi proteins catalysed by dPRMT5 is required for Ago3 and Aub stability. *Nature Cell Biol.* **11,** 652–658 (2009).
39. Maragkakis, M., Alexiou, P., Nakaya, T. & Mourelatos, Z. CLIPSeqTools-a novel bioinformatics CLIP-seq analysis suite. *RNA* **22,** 1–9 (2016).
40. Maragkakis, M., Alexiou, P. & Mourelatos, Z. GenOO: a modern perl framework for high throughput sequencing analysis. Preprint at http://biorxiv.org/content/early/2015/11/03/019265 (2015).
41. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25,** 1754–1760 (2009).
42. Bullard, J. H., Purdom, E., Hansen, K. D. & Dudoit, S. Evaluation of statistical methods for normalization and differential expression in mRNA-seq experiments. *BMC Bioinformatics* **11,** 94 (2010).
43. Breitling, R., Armengaud, P., Amtmann, A. & Herzyk, P. Rank products: A simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett.* **573,** 83–92 (2004).
44. Smith, T. F. & Waterman, M. S. Identification of common molecular subsequences. *J. Mol. Biol.* **147,** 195–197 (1981).

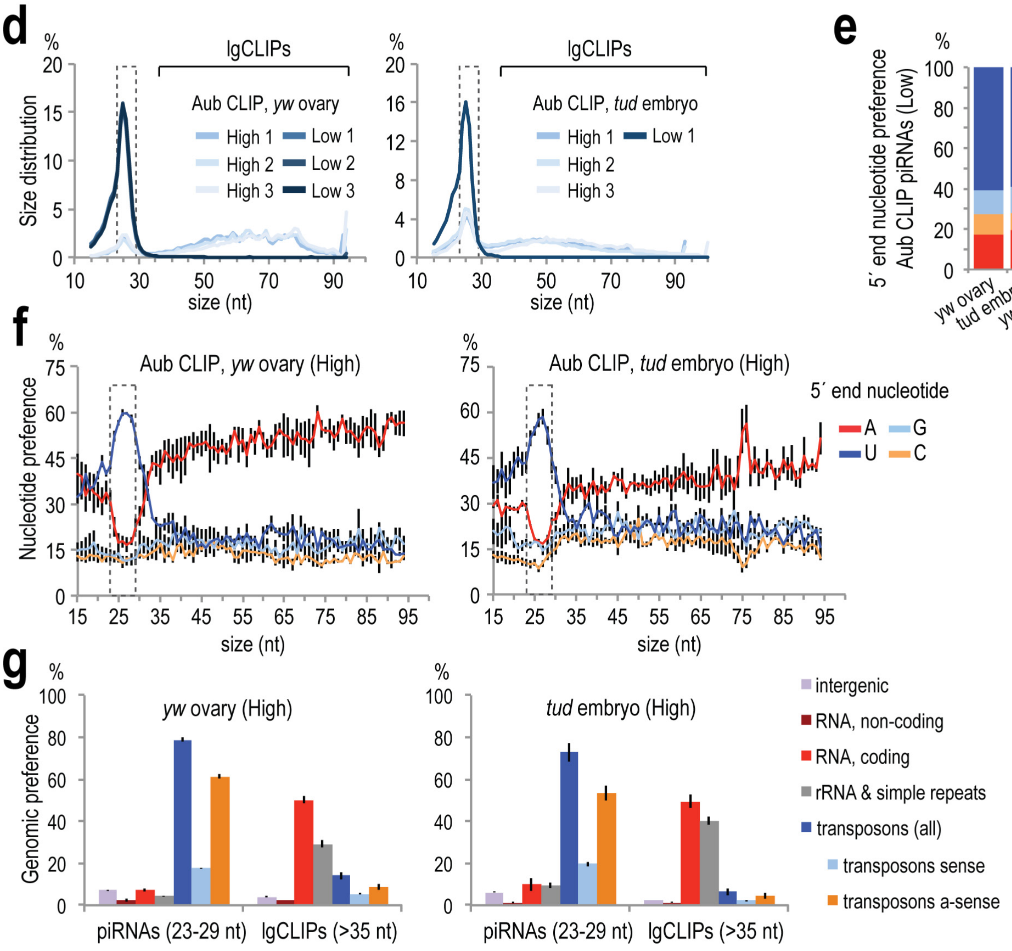


b

CLIP-Seq library	Number of sequenced reads	Number of mapped reads	piRNAs (23-29 nt)	lgCLIPs (>35 nt)
Aub_CLIP_yw_embryo_0_2_H_1	8,818,464	6,510,923	1,470,826	4,052,071
Aub_CLIP_yw_embryo_0_2_H_2	15,616,681	12,398,341	4,035,158	5,597,930
Aub_CLIP_yw_embryo_0_2_H_3	21,455,325	15,409,469	3,022,117	10,454,759
Aub_CLIP_yw_embryo_0_2_L_1	14,062,489	13,826,920	7,624,548	108,605
Aub_CLIP_yw_embryo_0_2_L_2	17,610,892	17,257,652	10,002,213	121,154
Aub_CLIP_yw_embryo_0_2_L_3	18,420,287	18,055,396	10,885,017	140,132
Aub_CLIP_yw_ovary_H_1	19,041,920	15,689,202	1,486,327	13,422,595
Aub_CLIP_yw_ovary_H_2	22,387,813	20,029,273	1,668,627	17,419,663
Aub_CLIP_yw_ovary_H_3	20,862,621	18,425,256	2,059,053	15,309,011
Aub_CLIP_yw_ovary_L_1	24,257,591	23,951,410	16,103,669	227,098
Aub_CLIP_yw_ovary_L_2	18,584,476	18,287,691	12,540,296	134,799
Aub_CLIP_yw_ovary_L_3	52,408,658	45,884,221	31,826,158	678,058
Aub_CLIP_tud_embryo_1_H	43,313,537	36,226,058	7,424,930	23,699,595
Aub_CLIP_tud_embryo_2_H	36,687,776	30,485,898	7,527,602	17,647,540
Aub_CLIP_tud_embryo_3_H	42,050,665	30,959,358	6,438,572	21,434,719
Aub_CLIP_tud_embryo_L	25,284,701	24,773,498	16,404,216	109,985

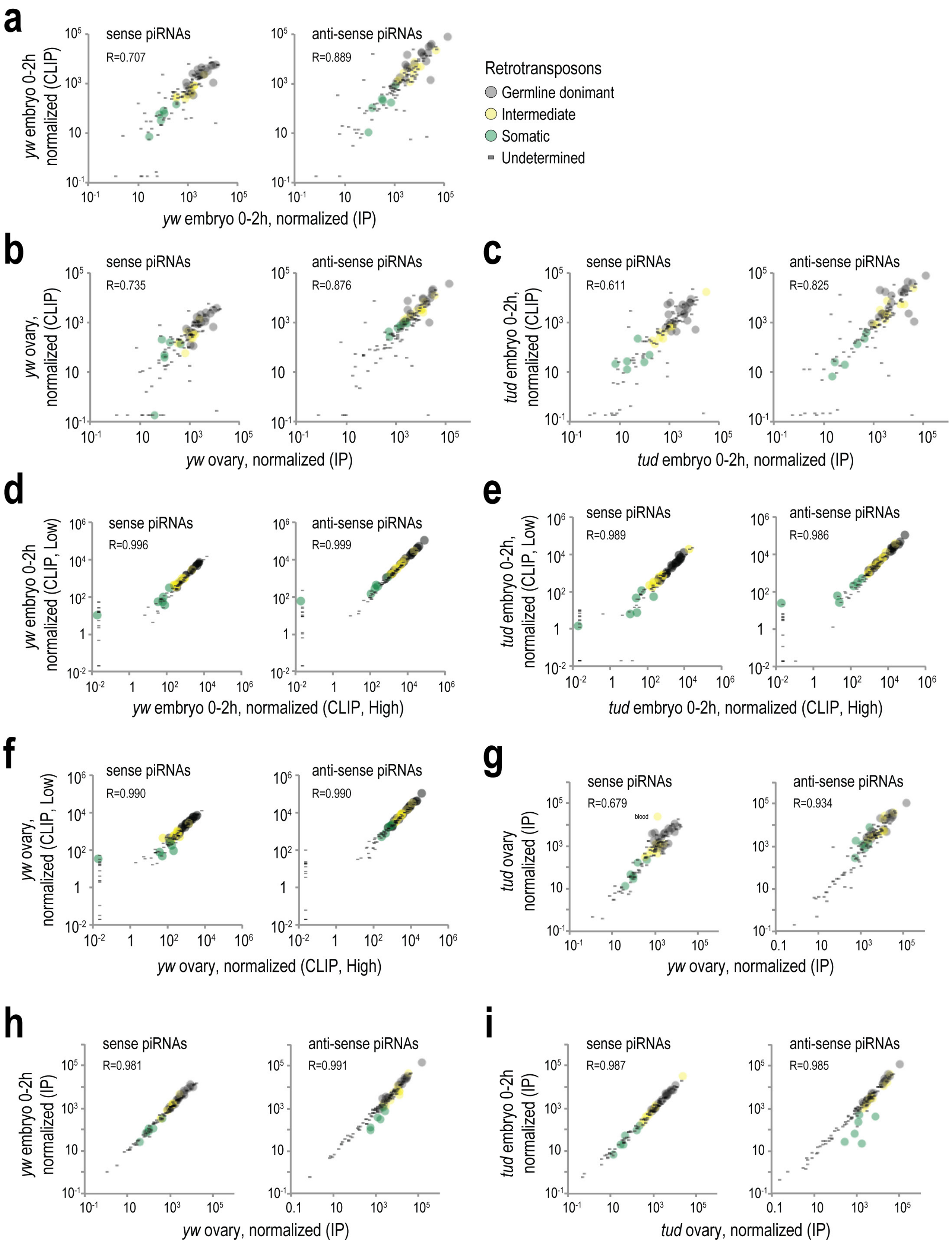
c

Small RNA-Seq library	Number of sequenced reads	Number of mapped reads
Aub_IP_tud_embryo_(0-2h)	35,614,928	35,368,281
Aub_IP_tud_embryo_(stage_4-6)	30,853,326	30,640,207
Aub_IP_tud_ovary	21,050,596	20,905,116
Aub_IP_yw_embryo_(0-2h)	43,405,719	42,958,773
Aub_IP_yw_embryo_(stage_4-6)	43,940,406	43,512,435
Aub_IP_yw_embryo_(stage_7-9)	35,047,292	34,694,141
Aub_IP_yw_ovary	71,829,805	70,295,208



Extended Data Figure 1 | See next page for caption.

Extended Data Figure 1 | Endogenous Aub localization in genotypes used, sequenced and mapped reads of CLIP sequencing and RNA immunoprecipitation libraries used in this study, and general characteristics of *yw* ovary and *tud* embryo (0–2 h) CLIP sequencing libraries. a, Immunofluorescence of ovary and early embryo of indicated genotypes using antibodies against Aub (Aub-83; green) and Tudor (red), and schematic representation of the egg chamber. Aub is localized in the nuage and germ (pole) plasm of wild-type ovaries, in the germ plasm of early wild-type embryos (stage 2) and within PGCs as they form in the posterior pole (stage 5), and as they migrate during gastrulation (stage 10). Tudor colocalizes with Aub in the germ plasm of early embryos but is not detected after PGC formation. In Tudor mutant early embryos, Aub is not concentrated in the posterior but is diffusely present throughout the embryo; PGCs are never specified resulting in agametic adults (see also Extended Data Fig. 9). **b**, Sequenced and mapped reads of CLIP sequencing (CLIP-seq) libraries prepared in this study. **c**, Sequenced and mapped reads of RNA immunoprecipitation deep-sequencing libraries prepared in this study. **d**, Size distribution for the three low (one for *tud*) and three high *yw* ovary and *tud* embryo (0–2 h) Aub CLIP-seq libraries. The size range of piRNAs (23–29 nucleotides) is indicated by a dashed box. **e**, Average 5′ end nucleotide composition for piRNAs (23–29 nucleotides) from three low *yw* ovary, *tud* embryo (0–2 h) (one library) and *yw* embryo (0–2 h) Aub CLIP-seq libraries. **f**, Average 5′ end nucleotide composition of CLIP tags from three high *yw* ovary and *tud* embryo (0–2 h) Aub CLIP-seq libraries. piRNAs (23–29 nucleotides) are indicated by a dashed box. **g**, Genomic distribution of CLIP tags for three high *yw* ovary and *tud* embryo (0–2 h) Aub CLIP-seq libraries. Overlap of piRNAs from CLIP and immunoprecipitation libraries. All error bars denote s.d.; $n = 3$.

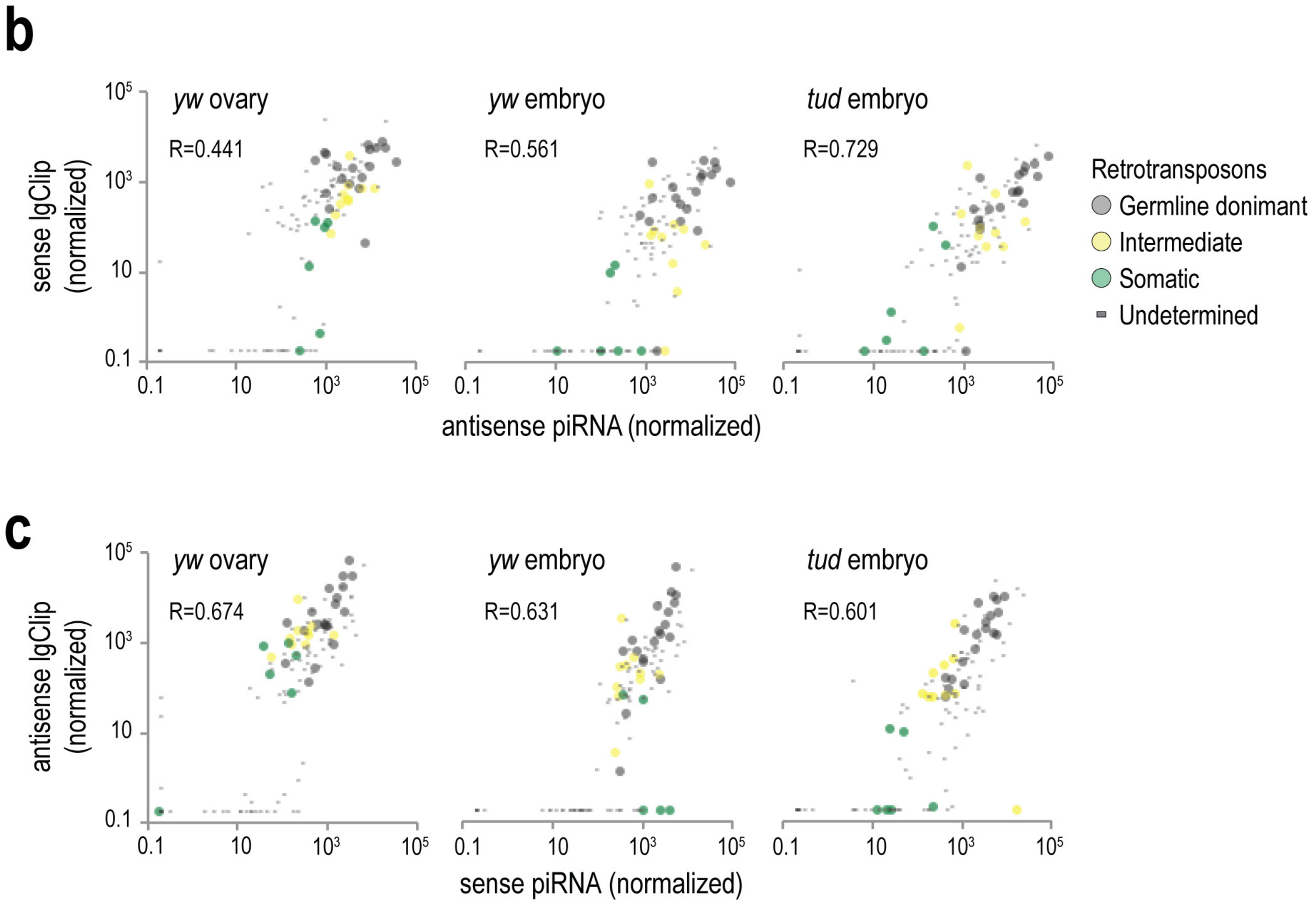


Extended Data Figure 2 | See next page for caption.

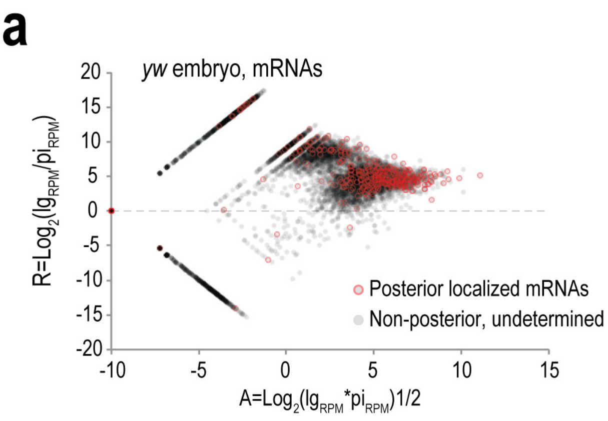
Extended Data Figure 2 | Pairwise comparisons of transposon piRNA populations from various libraries. a–**c**, Scatterplot comparison of normalized abundance of piRNAs mapped on consensus retrotransposon sequences (sense and antisense), from *yw* embryo (0–2 h) standard Aub immunoprecipitation and Aub CLIP libraries (**a**); from *yw* ovary libraries (**b**); and from *tud* embryo (0–2 h) libraries (**c**). Pearson correlation is shown for all elements in each plot. Retrotransposon categories are set as in ref. 31. **d**–**f**, Scatterplot comparison of normalized abundance of transposon-derived piRNAs in Aub CLIP libraries prepared from higher molecular mass signals (high; Fig. 1a, marked with a light blue line), with the piRNAs found in the libraries prepared from the main radioactive signal (low; Fig. 1a, marked with a dark blue line) from *yw* embryo (0–2 h) (**d**); from *yw* ovary Aub CLIP 'high' and 'low' libraries (**e**); and from *tud* embryo (0–2 h) Aub CLIP high and low libraries (**f**). These comparisons indicate that the piRNA loads in low and high CLIP libraries are essentially identical. **g**, Scatterplot comparison of normalized abundance of transposon-derived piRNAs for *yw* ovary and *tud* ovary Aub immunoprecipitation libraries, to evaluate changes of piRNA load in the absence of Tudor. While antisense-derived piRNAs are largely unchanged, a few sense-derived piRNAs are changed (blood retrotransposon is indicated). **h**, **i**, Scatterplot comparison of normalized abundance of transposon-derived piRNAs for *yw* ovary and *yw* embryo (0–2 h) Aub immunoprecipitation libraries (**h**); and for *tud* ovary and *tud* embryo (0–2 h) libraries (**i**).

a

Library	total lgClips mapped within retrotransposons (consensus)	lgClips with overlapping antisense piRNAs	lgClips with overlapping antisense piRNAs (%)	average (%)
dme_Aub_CLIP_tud_embryo_0-2h_H1	3866282	785302	20.31155513	27.18524074
dme_Aub_CLIP_tud_embryo_0-2h_H2	3832702	1130647	29.49999765	
dme_Aub_CLIP_tud_embryo_0-2h_H3	3601118	1143145	31.74416945	
dme_Aub_CLIP_yw_embryo_0-2h_H1	886069	66353	7.488468731	15.62595067
dme_Aub_CLIP_yw_embryo_0-2h_H2	2558390	471642	18.43510958	
dme_Aub_CLIP_yw_embryo_0-2h_H3	1784247	373876	20.95427371	
dme_Aub_CLIP_yw_ovary_H1	858940	138301	16.10135749	21.03149503
dme_Aub_CLIP_yw_ovary_H2	999022	151961	15.21097633	
dme_Aub_CLIP_yw_ovary_H3	1231405	391367	31.78215128	

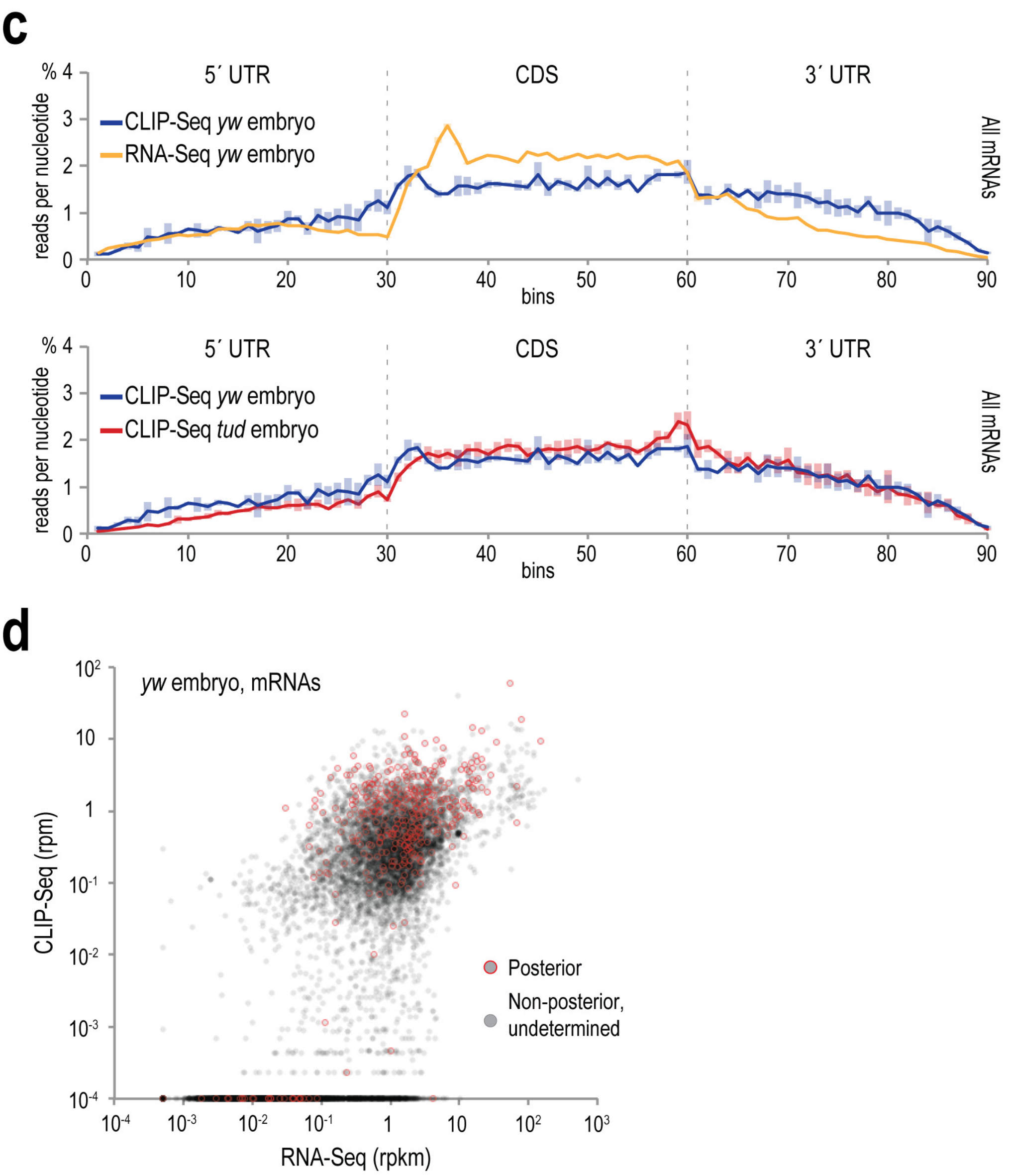


Extended Data Figure 3 | Retrotransposon targeting by complementary piRNAs identified by Aub CLIP. a, Overlap of lgCLIPs with complementary piRNAs from CLIP libraries, mapping on retrotransposons. **b**, **c**, Scatterplots of normalized abundance of antisense piRNAs and sense lgCLIPs (**b**) and for sense piRNAs and antisense lgCLIPs (**c**) mapped on retrotransposons for the indicated Aub CLIP libraries. Pearson correlation is shown for all elements in every plot. Retrotransposon categories are set as in ref. 31.



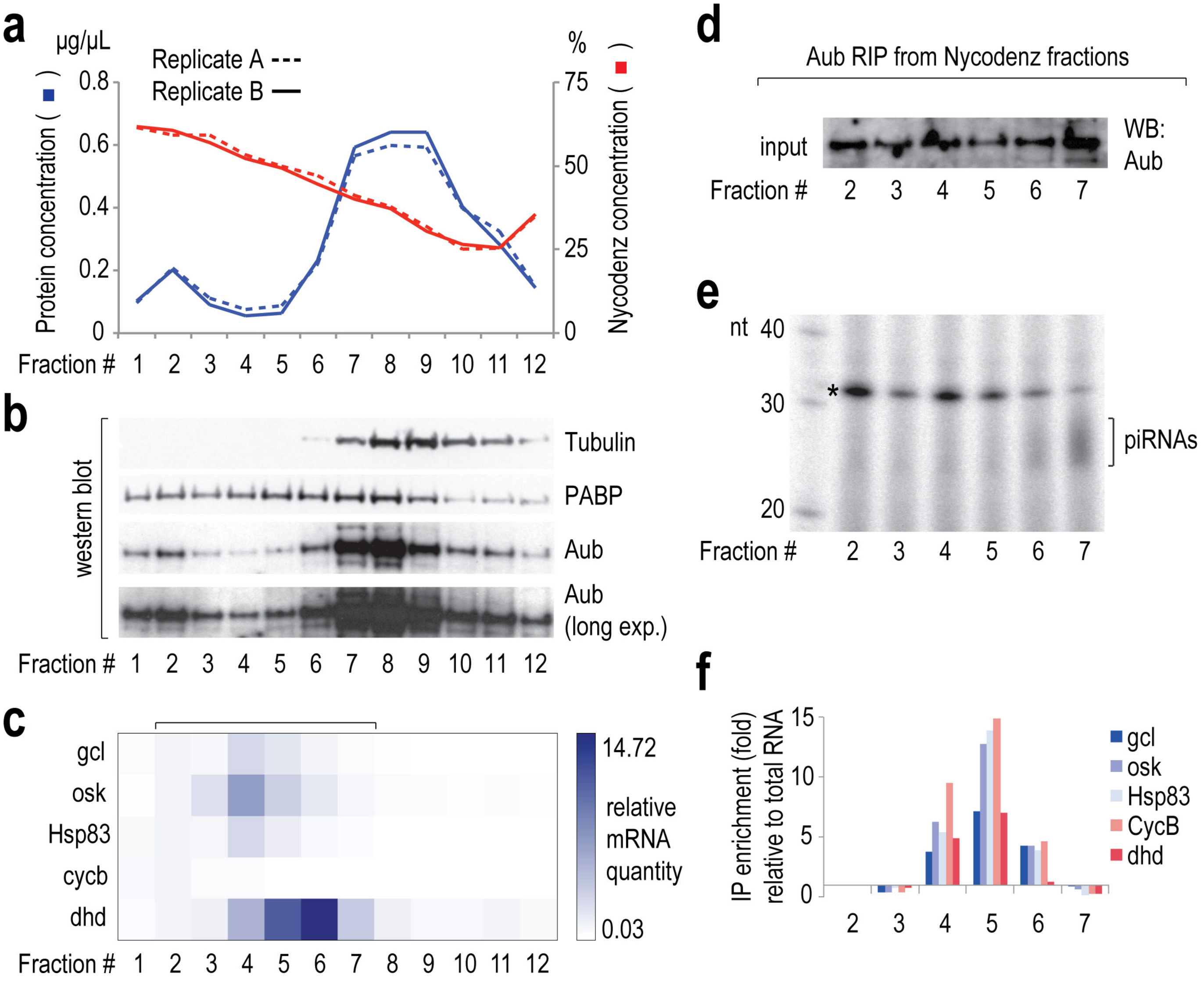
b

RNA-Seq library	Number of sequenced reads	Number of mapped reads
yw_embryo_0-2h_1	50,017,107	36,677,726
yw_embryo_0-2h_2	24,011,351	18,364,321
tud_embryo_0-2h_1	10,098,572	9,283,256
tud_embryo_0-2h_2	11,677,772	10,217,987
aub_embryo_0-2h_1	43,856,239	30,767,884
aub_embryo_0-2h_2	42,343,127	26,912,631



Extended Data Figure 4 | See next page for caption.

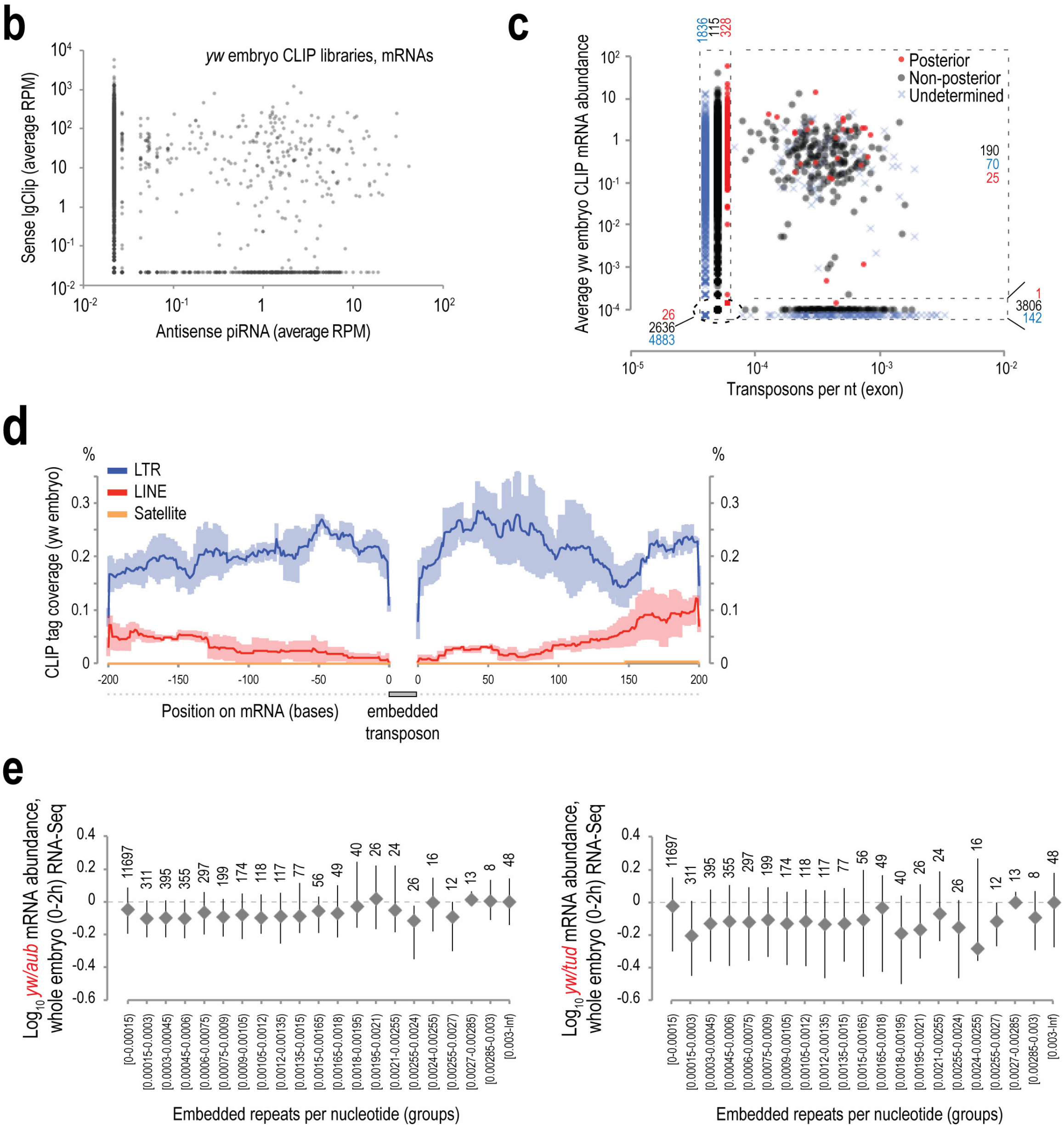
Extended Data Figure 4 | CLIP identifies extensive mRNA binding by Aub. a, Ratio average plot of normalized (reads per million, RPM) Aub CLIP tag (pi, piRNA; lg, lgCLIP) abundance (*A* value) versus lgCLIPs over piRNA abundance (*R* value), for all mRNAs. Outlined circles (red) correspond to genes that belong in the 12 posterior localization categories depleted in *tud* versus *yw* Aub CLIP libraries. Zero values are substituted with a small (smallest than the minimum) value so that log calculations are possible. This graph strongly suggests that mRNA binding by Aub as captured by CLIP is not for piRNA biogenesis purposes. **b**, Sequenced and mapped reads of RNA-seq libraries prepared in this study. **c**, Density of Aub CLIP-seq tags (*yw* embryo, and bottom panel: *tud* embryo) and RNA-seq reads (top panel: *yw* embryo) within the UTRs and coding sequences of the meta-mRNA. Each mRNA region is divided in 30 bins, and the number of the chimaeric mRNA fragments (genomic coordinate of the mRNA fragment midpoint) mapped within each bin is counted. Error bars indicate one s.d., $n = 3$ for CLIP-seq; minimum and maximum values for the two RNA-seq replicate libraries. **d**, Scatterplot of average normalized mRNA abundance for *yw* embryo RNA-seq (RPKM) and Aub CLIP-seq (RPM). Aub highly bound mRNAs with posterior localizations (Supplementary Table 4) are marked with a red circle. Zero values are substituted with a small (smallest than the minimum) value so that log calculations are possible. CLIP-seq identifies mRNAs that span the whole expression range of RNA-seq libraries, indicating that Aub CLIP does not capture transcripts simply based on abundance.



Extended Data Figure 5 | Partial purification of Aub RNPs from early embryo supports binding of germ plasm mRNAs by Aub. **a**, Fractionation of isopycnic Nycodenz density gradients of post-nuclear *yw* embryo lysate. Protein and Nycodenz concentration for every fraction is plotted. **b**, Western blot detection of indicated proteins in gradient fractions. A short and a long exposure (exp.) for Aub is shown. Uncropped gels for **b**, **d** and **e** can be found in Supplementary Fig. 1. **c**, Heat map of levels of indicated germ plasm mRNAs determined by quantitative RT–PCR (qRT–PCR), normalized to spiked luciferase RNA, and with fraction 2 as a reference. **d**, Western blot detection of Aub in indicated diluted Nycodenz fractions used for Aub RNA immunoprecipitation. **e**, Electrophoretic analysis on denaturing polyacrylamide gel of ^{32}P-labelled small RNAs immunoprecipitated with Aub from indicated gradient fractions. A bracket denotes piRNAs, detected primarily in fractions 6 and 7 (asterisk denotes 2S rRNA). **f**, Bar chart showing fold enrichment (over fraction-extracted total RNA) of indicated germ plasm mRNAs in Aub immunoprecipitations from gradient fractions, measured by qRT–PCR. Luciferase mRNA was used as a spike.

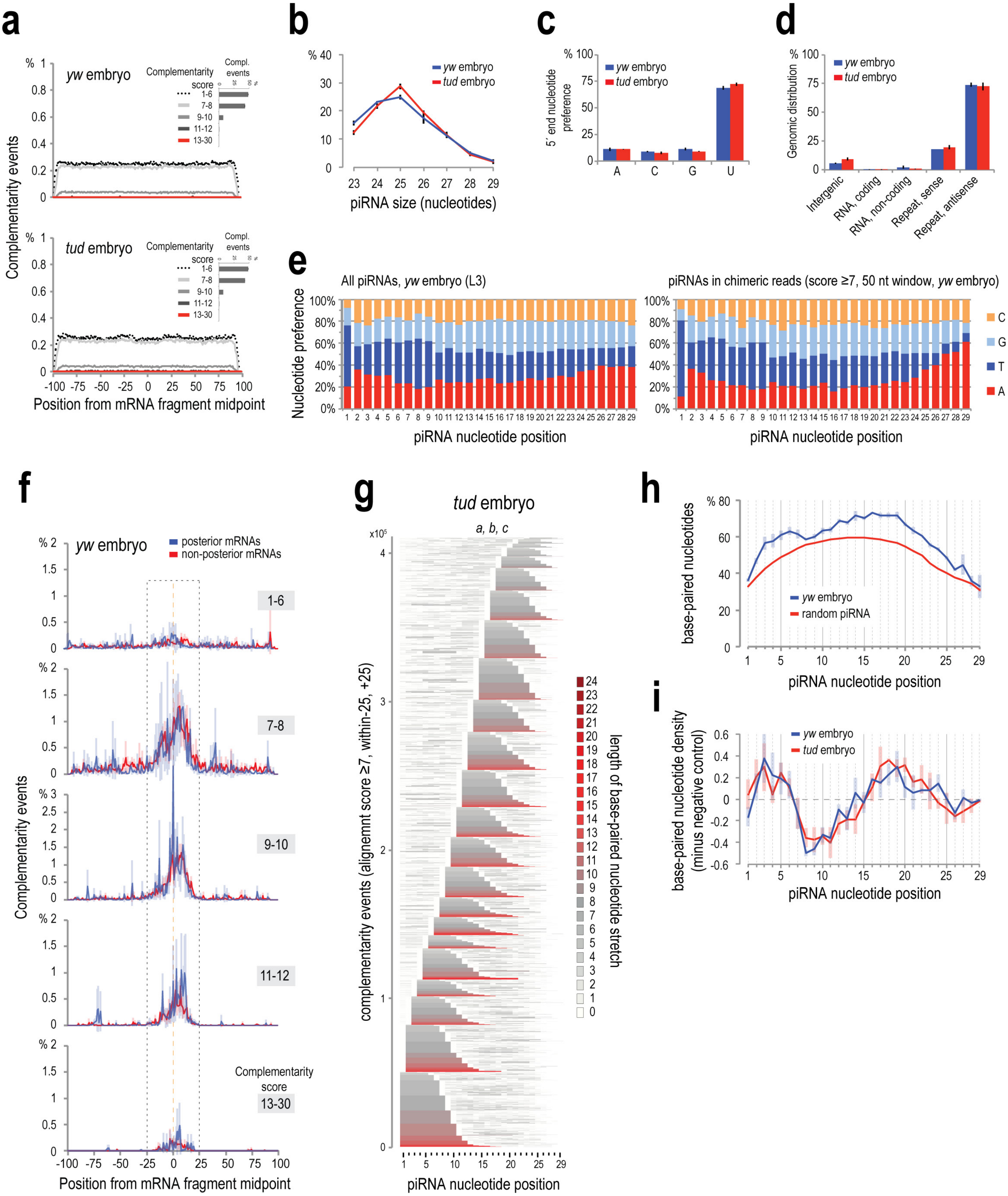
a

Library	total lgClips mapped within mRNAs	lgClips with overlapping antisense piRNAs	lgClips with overlapping antisense piRNAs (%)	average
dme_Aub_CLIP_tud_embryo_0-2h_H1	1000054	447	0.0%	0.1%
dme_Aub_CLIP_tud_embryo_0-2h_H2	909433	1300	0.1%	
dme_Aub_CLIP_tud_embryo_0-2h_H3	416111	941	0.2%	
dme_Aub_CLIP_yw_embryo_0-2h_H1	131081	141	0.1%	0.2%
dme_Aub_CLIP_yw_embryo_0-2h_H2	264642	369	0.1%	
dme_Aub_CLIP_yw_embryo_0-2h_H3	290583	1096	0.3%	
dme_Aub_CLIP_yw_ovary_H1	151394	281	0.2%	0.3%
dme_Aub_CLIP_yw_ovary_H2	148381	179	0.1%	
dme_Aub_CLIP_yw_ovary_H3	175321	1171	0.6%	



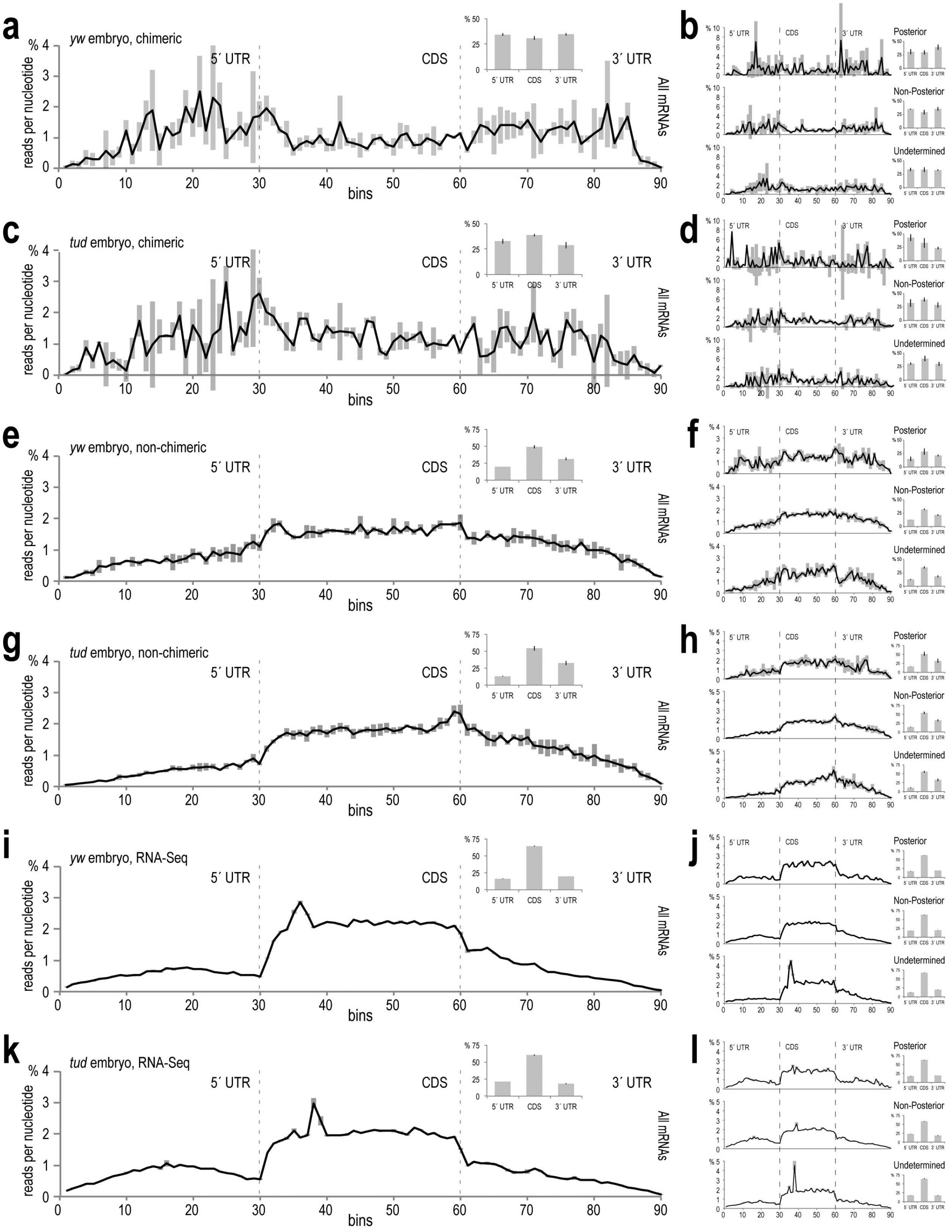
Extended Data Figure 6 | See next page for caption.

Extended Data Figure 6 | Analysis of Aub CLIP tags mapping to mRNAs with regard to the presence of mRNA embedded transposons. a, Overlap of lgCLIPs with complementary piRNAs from CLIP libraries, mapping on mRNAs. **b**, Scatterplot of *yw* embryo Aub lgCLIPs mapped in the sense orientation on mRNAs, with piRNAs mapped in the antisense orientation. Zero values are substituted with a small (smallest than the minimum) value so that log calculations are possible. Contrary to retrotransposons (Extended Data Fig. 3), there is no correlation, suggesting that extensive piRNA complementarity cannot explain the widespread mRNA binding shown by mRNA lgCLIPs. **c**, Scatterplot of *yw* embryo Aub lgCLIPs mapped in the sense orientation on mRNAs with per base (nucleotide) mRNA embedded retrotransposons (LINE, long terminal repeat (LTR), satellite). Posterior, non-posterior and undetermined localizations are marked as indicated. The graph is separated into four quadrants: clockwise from bottom left corner: 0 embedded repeats, 0 CLIP tags; 0 embedded repeats, >0 CLIP tags; >0 embedded repeats, >0 CLIP tags, >0 embedded repeats, 0 CLIP tags. The number of genes in the four quadrants is indicated. Zero values are substituted with a small (smallest than the minimum) value (different small value for every localization category was used for clarity) so that log calculations are possible. This graph suggests that there is no correlation between the numbers of CLIP tags and embedded repeats within the mRNAs. **d**, Aub lgCLIPs density surrounding (±200 bases) mRNA-embedded retrotransposons (LINE, LTR, satellite as indicated). This analysis shows that there is no increase in the lgCLIP density in the areas flanking embedded repeats, suggesting that repeat sequences are not used as enriched target areas for mRNA binding by Aub. Error bars denote s.d.; $n = 3$. **e**, Analysis of mRNA expression level in relation to the number of embedded repeats. The number of embedded repeats per nucleotide of exon was plotted with the ratio ($\log_{10}$) of mRNA expression in *yw* embryo (0–2 h) versus *aub*$^{HN2/QC42}$ embryo (0–2 h) (left), and *yw* embryo (0–2 h) versus *tud* embryo (0–2 h) (right). The mRNAs are divided into groups based on the number of embedded repeats. The number above each data point denotes the number of mRNAs in each group. The graphs suggest that there is no proportional or consistent abundance change, decrease or increase, with the number of embedded repeats.



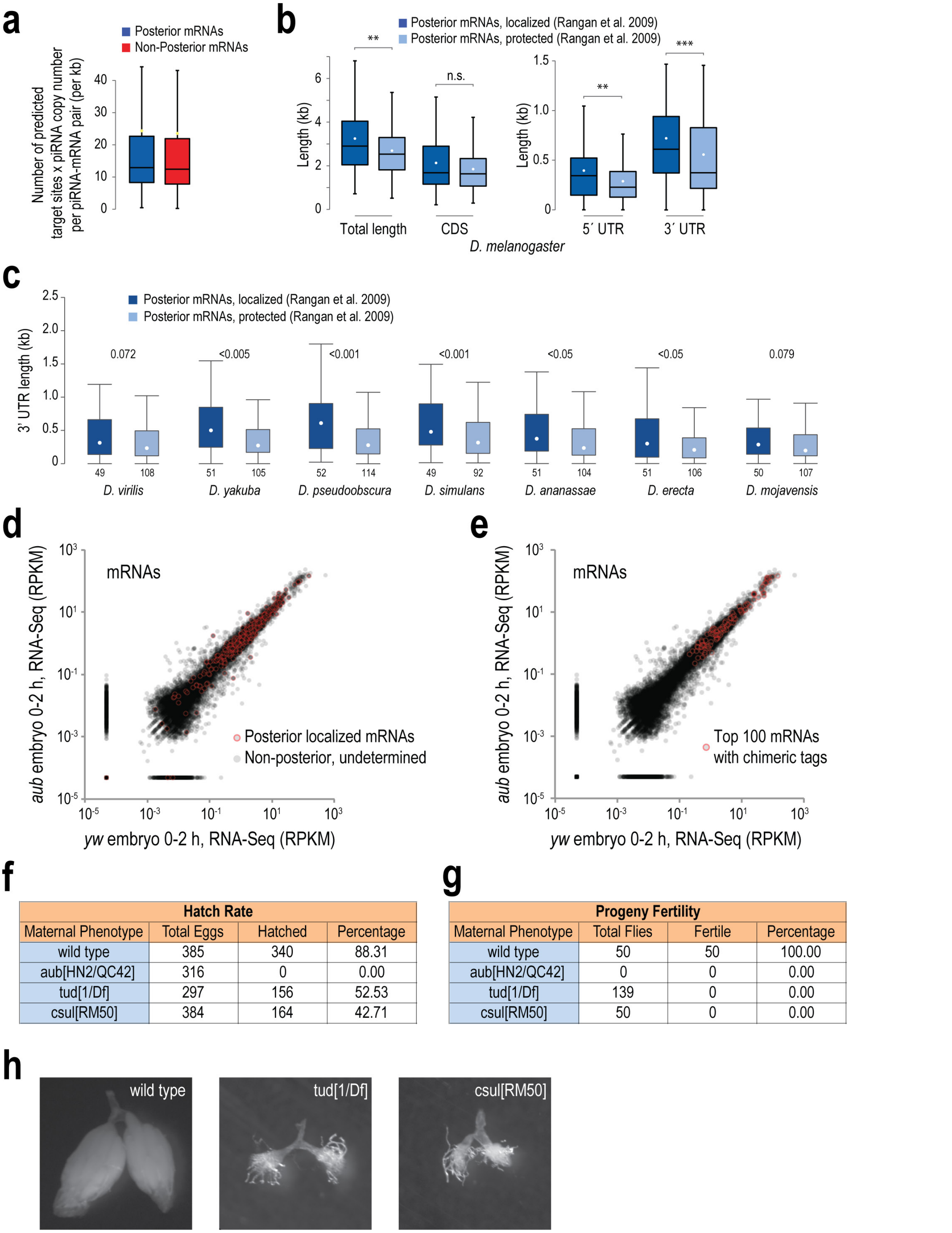
Extended Data Figure 7 | See next page for caption.

Extended Data Figure 7 | Characteristics of piRNAs and piRNA base-pairing with complementary target sites identified from analysis of chimaeric CLIP tags. a, piRNA–mRNA complementarity events for a random piRNA (negative control, average of three *yw* (top) and *tud* (bottom) embryo (0–2 h) samples), within ±100 bases from the midpoint of the mRNA part of the chimaeric read. Complementarity events are plotted per alignment score group as indicated, for clarity. Inset (per sample): bar chart of average complementarity events per score group. **b**, Size distribution of the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo CLIP libraries. Only the piRNAs implicated in the complementarity events occurring within ±25 nucleotides from the midpoint of the mRNA fragment and with score ≥7 are analysed in this graph, and the graphs in **c**–**e**, **g**–**i**. **c**, 5′ end nucleotide preference for the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo Aub CLIP libraries. **d**, Genomic distribution for the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo Aub CLIP libraries. **e**, Per position nucleotide preference for all piRNAs in Aub *yw* embryo (0–2 h) CLIP library L3 (left), and for the piRNAs identified within chimaeric CLIP tags, for *yw* and *tud* embryo Aub CLIP libraries. **f**, Complementarity events between piRNAs and mRNA fragments of chimaeric reads, for posterior and non-posterior localized mRNAs (*yw* embryo). The plots are separated per score group. **g**, Heat maps showing base-paired nucleotides of piRNAs for all complementarity events identified within chimaeric CLIP tags (events occurring within ±25 nucleotides from mRNA fragment midpoint, score ≥7) for *tud* embryo. Colour is according to the length of the consecutive stretch of base-paired nucleotides that runs over every position (colour code shown on the right). Stacked piRNAs are aligned at their 5′ ends and sorted (bottom to top) following these rules: (a) starting position of the longest stretch of consecutive base paired nucleotides, relative to the piRNA end; (b) length of longest base-paired stretch; (c) total number of base-paired nucleotides. **h**, Base-pairing frequency along the piRNA length for *yw* embryo libraries (blue) and their negative control (red). **i**, Net base-pairing frequency along the piRNA length (red) and net density of base paired nucleotides (grey) in mRNAs from chimaeric CLIP tags from *tud* embryo libraries. All error bars denote s.d.; $n = 3$.



Extended Data Figure 8 | See next page for caption.

Extended Data Figure 8 | Non-chimaeric Aub CLIP tag (lgCLIP), chimaeric mRNA fragment and RNA-seq read density along the untranslated and coding sequences of mRNAs. a, Average density of chimaeric mRNA fragments (Aub CLIP, *yw* 0–2-h embryo) along the three parts of the meta-mRNA. Each mRNA region is divided in 30 bins and the number of the chimaeric mRNA fragments (genomic coordinate of the mRNA fragment midpoint) mapped within each bin is counted. Inset: bar plot showing cumulative density in each mRNA region. **b**, Average density of the chimaeric mRNA fragments on mRNA regions; mRNAs are separated into three localization groups as indicated: posterior localized (12 categories; Supplementary Table 3), non-posterior and undetermined localization. Inset: bar plot showing cumulative density in each mRNA region. **c**, As in **a** for chimaeric mRNA fragments from Aub CLIP libraries, *tud* embryo (0–2 h). **d**, As in **b** for chimaeric mRNA fragments from Aub CLIP libraries, *tud* embryo (0–2 h). **e**, As in **a** for non-chimaeric lgCLIPs from Aub CLIP libraries, *yw* embryo (0–2 h). **f**, As in **b** for non-chimaeric lgCLIPs from Aub CLIP libraries, *yw* embryo (0–2 h). **g**, As in **a** for non-chimaeric lgCLIPs from Aub CLIP libraries, *tud* embryo (0–2 h). **h**, As in **b** for non-chimaeric lgCLIPs from Aub CLIP libraries, *tud* embryo (0–2 h). **i**, As in **a** for RNA-seq reads, *yw* embryo (0–2 h). **j**, As in **b** for RNA-seq reads, *yw* embryo (0–2 h). **k**, As in **a** for RNA-seq reads, *tud* embryo (0–2 h). **l**, As in **b** for RNA-seq reads, *tud* embryo (0–2 h). Error bars denote s.d.; $n = 3$.



Hatch Rate			
Maternal Phenotype	Total Eggs	Hatched	Percentage
wild type	385	340	88.31
aub[HN2/QC42]	316	0	0.00
tud[1/Df]	297	156	52.53
csul[RM50]	384	164	42.71

Progeny Fertility			
Maternal Phenotype	Total Flies	Fertile	Percentage
wild type	50	50	100.00
aub[HN2/QC42]	0	0	0.00
tud[1/Df]	139	0	0.00
csul[RM50]	50	0	0.00

Extended Data Figure 9 | See next page for caption.

Extended Data Figure 9 | Lengths of posterior localized mRNAs in *Drosophila* species; characteristics of embryos used in our studies. **a**, Box-and-whisker plot of the number of predicted piRNA target sites (per kilobase of mRNA sequence) for every mRNA–piRNA pair, multiplied by the piRNA copy number. Posterior and non-posterior mRNAs are as indicated. Black lines denote the median. This graph indicates that the 'targeting potential' (number of predicted complementary sites multiplied by the piRNA copy number) of each piRNA against each mRNA is the same for the two localization categories, suggesting that the piRNA copy number is not a contributing factor for the observed preference of posterior localized mRNAs for piRNA adhesion. **b**, Box-and-whisker plot of the lengths of *D. melanogaster* mRNAs (and their 5′ UTR, coding sequences and 3′ UTR parts) that are found in the enriched and protected categories, as defined previously[10]. Black lines denote the median; white dots denote the mean. n.s., not significant ($P > 0.05$); $^{**}P < 0.01$; $^{***}P < 0.001$; one-sided Wilcoxon rank sum test. **c**, Box-and-whisker plot of the lengths of the 3′ UTRs of mRNAs from the indicated *Drosophila* species that are orthologous to the *D. melanogaster* mRNAs found in the localized and protected categories, as defined previously[10]. Incomplete annotation did not allow us to perform this analysis for all the species shown in Fig. 4i.White dots denote the mean. *P* values of the statistical test (one-sided Wilcoxon test) of whether the lengths of the localized versus protected mRNAs are different, are shown for each species. **d**, **e**, RNA-seq scatterplots from 0–2-h wild-type (*yw*) and 0–2-h Aub-null (*aub*) embryos. Shown in red are posterior localized mRNAs (**d**) or the top 100 mRNAs identified from Aub CLIP piRNA–mRNA chimaeric reads (**e**). There is no change in mRNA levels between wild-type and *aub* mutant 0–2-h embryos. **f**, **g**, Hatch rates (**f**) and fertility of progeny (**g**) of embryos from indicated genotypes. Note that, unlike Tudor and Csul, the absence of Aub ($aub^{HN2/QC42}$) leads to complete embryo lethality. **h**, Gross ovary appearance of wild-type (*yw*), Tudor mutant (*tud[1/Df])* and Csul mutant ($csul^{RM50}$) adult flies. Note complete absence of germline ovarian tissue in adult flies lacking Tudor or Csul; embryos from these flies develop into agametic adults because PGCs are never specified.

Extended Data Table 1 | Overlap of piRNAs from CLIP and immunoprecipitation libraries

Library 1	Library 2	unique piRNA sequences in library 1	unique piRNA sequences in library 2	common	percent1	percent2	average percent 2
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_H1	6913438	348812	150654	2.179147336	43.19060124	42.22806
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_H2	6913438	838891	333876	4.829377222	39.79968792	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_H3	6913438	694458	284532	4.115636822	40.97180823	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_H1	9938639	560082	286627	2.883966306	51.17589924	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_H3	9938639	293375	156976	1.579451673	53.50694504	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_H2	9938639	332484	176012	1.770986953	52.93848727	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_L1	5147948	1257672	458182	8.900284152	36.43096133	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_H2	5147948	1104187	460392	8.943213879	41.69511143	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_H3	5147948	2567880	948630	18.42734231	36.94214683	
Aub_IP_tud_embryo_0-2h	Aub_CLIP_tud_embryo_0-2h_H1	5147948	1040626	379030	7.362739484	36.4232683	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_L1	9938639	1850192	874693	8.800933407	47.27579624	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_L2	9938639	2407082	1108175	11.15016855	46.03810755	
Aub_IP_yw_ovary	Aub_CLIP_yw_ovary_L3	9938639	3082922	1367516	13.75959022	44.35778784	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_L1	6913438	2012094	722743	10.45417634	35.91994211	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_L2	6913438	2161685	769241	11.12675054	35.58524947	
Aub_IP_yw_embryo_0-2h	Aub_CLIP_yw_embryo_0-2h_L3	6913438	2701578	902250	13.0506703	33.39714789	

Comparisons of piRNA sequences found in CLIP and immunoprecipitation libraries from same tissues.