# Title: Repeated duplication of Argonaute2 is associated with strong selection and testis specialization in *Drosophila*

Authors: Samuel H. Lewis[1,2]*, Claire L. Webster[1,3], Heli Salmela[4] & Darren J. Obbard[1,5]

Affiliations:

[1]Institute of Evolutionary Biology, University of Edinburgh, Kings Buildings, EH9 3JT, United Kingdom

[2]Present Address: Department of Genetics, University of Cambridge, Downing Street, Cambridge, CB2 3EH

[3]Present Address: Life Sciences, University of Sussex, United Kingdom

[4]Department of Biosciences, Centre of Excellence in Biological Interactions, University of Helsinki, Helsinki, Finland

[5]Centre for Immunity, Infection and Evolution, University of Edinburgh, Kings Buildings, EH9 3JT, United Kingdom

*Author for correspondence: sam.lewis@gen.cam.ac.uk

## Abstract

Argonaute2 (Ago2) is a rapidly evolving nuclease in the *Drosophila melanogaster* RNAi pathway that targets viruses and transposable elements in somatic tissues. Here we reconstruct the history of Ago2 duplications across the *Drosophila obscura* group, and use patterns of gene expression to infer new functional specialization. We show that some duplications are old, shared by the entire species group, and that losses may be common, including previously undetected losses in the lineage leading to *D. pseudoobscura*. We find that while the original (syntenic) gene copy has generally retained the ancestral ubiquitous expression pattern, most of the novel Ago2 paralogues have independently specialised to testis-specific expression. Using population genetic analyses, we show that most testis-specific paralogues have significantly lower genetic diversity than the genome-wide average. This suggests recent positive selection in three different species, and model-based analyses provide strong evidence of recent hard selective sweeps in or near four of the six *D. pseudoobscura* Ago2 paralogues. We speculate that the repeated evolution of testis-specificity in *obscura* group Ago2 genes, combined with their dynamic turnover and strong signatures of adaptive evolution, may be associated with highly derived roles in the suppression of transposable elements or meiotic drive. Our study highlights the lability of RNAi pathways, even within well-studied groups such as *Drosophila*, and suggests that strong selection may act quickly after duplication in RNAi pathways, potentially giving rise to new and unknown RNAi functions in non-model species.

## Introduction

Argonaute genes are found in almost all eukaryotes, where they play a key role in antiviral immune defence, gene regulation and genome stability. They carry out this diverse range of functions through their role in RNA interference (RNAi) mechanisms, an ancient system of nucleic acid manipulation in which small RNA (sRNA) molecules guide Argonaute proteins to nucleic acid targets through base complementarity (reviewed in [1]). Gene duplication has occurred throughout the evolution of the Argonaute gene family, with ancient duplication events characteristic of some lineages – such as three duplications early in plant evolution [2], and multiple expansions and losses throughout the evolution of nematodes (reviewed in [3]) and the Diptera [4]. After duplication, Argonautes have often undergone functional divergence, involving changes in expression patterns and altered small RNA (sRNA) binding partners [5–7]. Duplication early in eukaryotic evolution produced two distinct Argonaute subfamilies, Ago and Piwi, which have since been retained in the vast majority of Metazoa [8]. Members of the Ago subfamily are expressed in both somatic and germline tissue, and variously bind sRNAs derived from host transcripts (miRNAs, endo-siRNAs) or transposable elements (TE endo-siRNAs) and viruses (viRNAs). In contrast, in most vertebrates and arthropods, the Piwi subfamily members are expressed only in association with the germline (reviewed in [9]), and bind sRNAs from TEs and host loci (piRNAs), suggesting that the Piwi subfamily specialised to a germline-specific role on the lineages leading to vertebrates and arthropods.

After the early divergence of the Ago and Piwi subfamilies, subsequent duplications gave rise to three Piwi subfamily members (Ago3, Aubergine (Aub) and Piwi) and two Ago subfamily members (Ago1 & Ago2) in *Drosophila melanogaster*. All three Piwi subfamily genes are associated with the germline and bind Piwi-interacting RNAs (piRNAs) derived from TEs and other repetitive genomic elements: Ago3 and Aub amplify the piRNA signal through the “Ping-Pong” cycle (reviewed in [10]), and Piwi suppresses transposition by directing heterochromatin formation [11]. These functional differences are associated with contrasting selective regimes, with Aub evolving under positive selection [12] and more rapidly than Ago3 and Piwi [13]. In contrast, Ago1 binds microRNAs (miRNAs), and regulates gene expression by inhibiting translation and marking transcripts for degradation (reviewed in [14]). This function imposes strong selective constraint on Ago1, resulting in slow evolution and very few adaptive substitutions [12,13,15]. Finally, Ago2 binds small interfering RNAs (siRNAs) from viruses (viRNAs) and TEs (endo-siRNAs), and functions in gene regulation [16], dosage compensation [17],

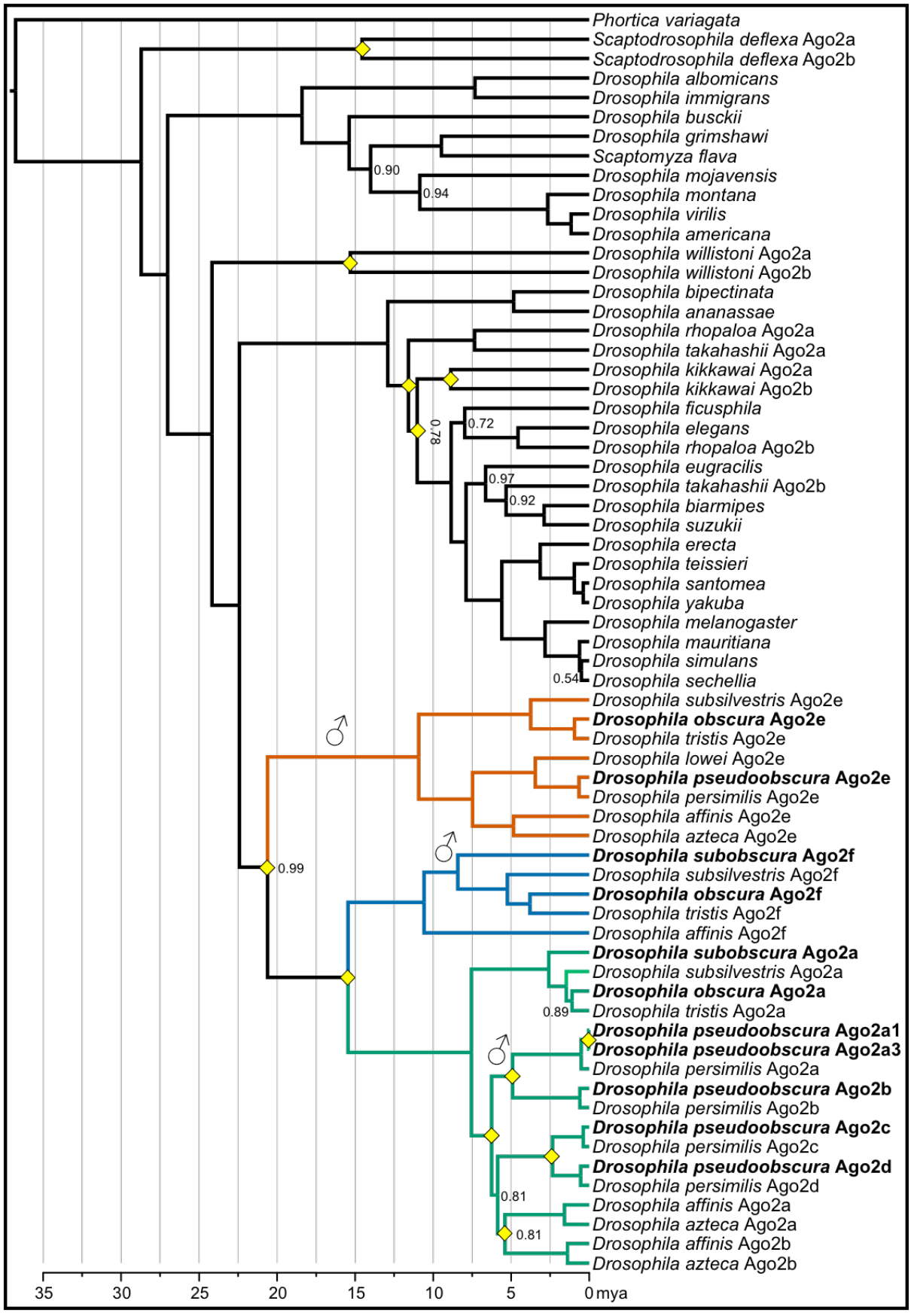
and the ubiquitous suppression of viruses [18,19] and TEs [20,21]. Ago2 also evolves under strong positive selection, with frequent selective sweeps [12,13,15,22,23], possibly driven by an arms race with virus-encoded suppressors of RNAi (VSRs) [15,24,25].

In contrast to *D. melanogaster*, from which most functional knowledge of Ago2 in arthropods is derived, an expansion of Ago2 has been reported in *D. pseudoobscura* [26], providing us with an ideal opportunity to study how the RNAi pathway evolves after duplication. Given the roles of *D. melanogaster* Ago2 in antiviral defence [18,19], TE suppression [20,21], dosage compensation [17], and gene regulation [16], we hypothesized that duplication in *D. pseudoobscura* may have led to subfunctionalization of Ago2 to a subset of these roles, or even to the evolution of entirely new functions. To elucidate the evolution and function of Ago2 paralogues in *D. pseudoobscura* and its relatives, we identified and dated Ago2 duplication events across available *Drosophila* genomes and transcriptomes, tested for divergence in expression patterns between the Ago2 paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*, and quantified the evolutionary rate and positive selection acting on each of these paralogues. We find that testis-specificity of Ago2 paralogues has evolved repeatedly in the *obscura* group, and that the majority of paralogues show evidence of recent positive selection.

## Results

### Ago2 has undergone numerous ancient and recent duplications in the *obscura* group

Ago2 duplications had previously been noted in *D. pseudoobscura* [26], but their age and distribution in other species was unknown. We used BLAST [27] and PCR to identify 65 Ago2 homologues in 39 species sampled across the Drosophilidae, including 30 homologues in 9 *obscura* group species. To characterize the relationships between Ago2 homologues in the *obscura* group and the other Drosophilidae, and estimate the date of the duplication events that produced them, we carried out a strict clock Bayesian phylogenetic analysis (Figure 1). This showed that there are early diverging Ago2 clades in the *obscura* group: the Ago2e subclade that diverged from other Ago2 paralogues around 21mya (±10 My), and the Ago2a and Ago2f subclades that were produced by a gene duplication event around 16mya (±7 My). Subsequently there have been a series of more recent duplications in the *D. pseudoobscura* subgroup Ago2a-d lineage. Using published genomes,




95

Figure 1: An approximately time-scaled Bayesian gene tree of Ago2 in the Drosophilidae. Duplication events are marked by yellow diamonds, Bayesian posterior support is shown for nodes for which it is less than 100%, and the genes and species that are the focus of the present study are marked in bold. Ago2 has duplicated at least twelve times in the Drosophilidae: seven times in the *obscura* group, twice early in the *melanogaster* group, and once each in the lineages leading to *D. willistoni*, *S. deflexa* and *D. kikkawai*. There has also been a potentially recent duplication of Ago2a on the *D. affinis* / *D. azteca* lineage (~5mya), although the low support for this node may suggest that these paralogues could also nest within the *D. pseudoobscura* / *D. persimilis* expansion, with one paralogue sister to the Ago2a1-Ago2b subclade and the other sister to the Ago2c-Ago2d subclade. After duplication, Ago2 paralogues in the *obscura* group have specialised to the testis three times independently (marked with ♂), and have been retained for an extended period of time (>10 My in the case of Ago2e), suggesting an adaptive basis for testis-specificity. The labelling a-e of paralogous clades corresponds to reference [26], while clade f is newly reported here.

transcriptomes and PCR we were unable to identify Ago2e in *D. subobscura*, Ago2e or Ago2f in *D. lowei*, or Ago2f in *D. pseudoobscura*, *D. persimilis* and *D. azteca*. While some of these losses may reflect incomplete genome assemblies or unexpressed genes in transcriptome surveys, we attempted to validate the losses in *D. pseudoobscura* and *D. subobscura* by extensive PCR, and were again unable to recover these genes.

In release 3.03 of the *D. pseudoobscura* genome Ago2b-Ago2e have confirmed locations, but Ago2a1 and Ago2a3 (the very recent paralogues newly identified here) lie in tandem on an unplaced contig with a third incomplete copy (Ago2a2) between them. We used PCR to confirm the existence, orientation, and relative positioning of these genes, and to identify the location of this contig, which lies in reverse orientation on chromosome XL-group1a (predicted coordinates 3,463,701-3,489,689). We then combined this information with our phylogenetic analysis to reconstruct the positional evolution of *D. pseudoobscura* Ago2 paralogues (Figure 2).

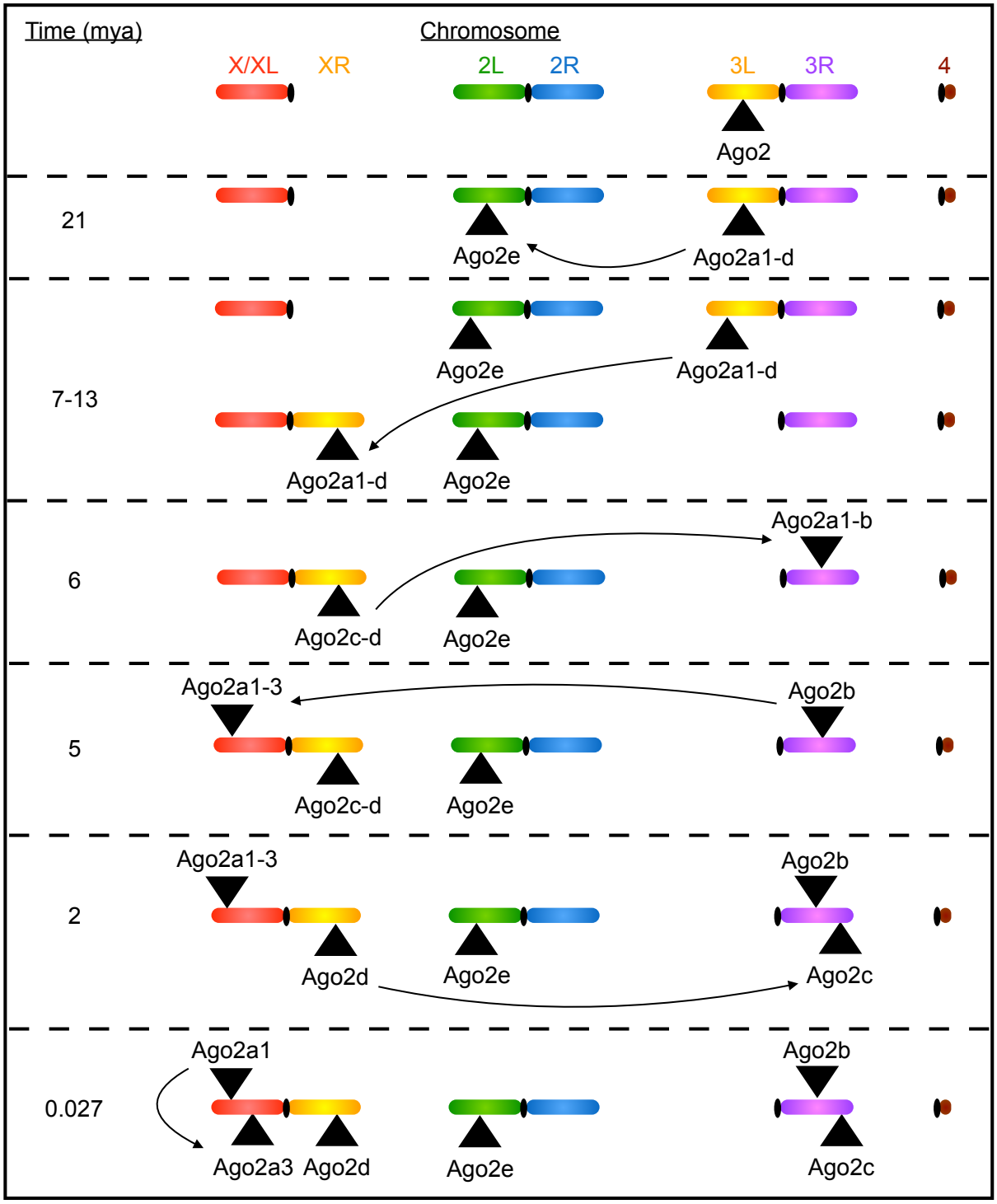



Figure 2: The course of duplications and translocations of Ago2 paralogues in *D. pseudoobscura.* A complex series of duplications and translocations has produced six Ago2 paralogues in *D. pseudoobscura*, located on four different chromosome arms. Chromosome arms correspond to Muller Elements A (X/XL), B (2L), C (2R), D (3L), E (3R) & F (4) (adapted from [91] Fig. 1). Firstly, the Ago2a1-e ancestor duplicated ~21mya to form Ago2a1-d and Ago2e, the latter of which moved onto chromosome 2L. Next, the 3L arm fused with the X chromosome, moving Ago2a1-d onto the X: this happened 7-15mya, after the divergence of the *obscura* group into Palearctic (e.g. *D. subobscura*) and Nearctic (e.g. *D. pseudoobscura*) clades [92]. Ago2a1-d then duplicated ~6mya, forming Ago2c-d and Ago2a1-b, the latter of which moved onto chromosome 2. After this, Ago2a1-b duplicated ~5mya, producing Ago2b and Ago2a1-3, the latter of which moved onto the left arm of the X chromosome. This was followed by a duplication of Ago2c-d ~2mya, forming Ago2d and Ago2c, the latter of which moved onto chromosome 2. Finally, Ago2a1-3 duplicated ~27kya, producing Ago2a1 and Ago2a3 in tandem. Note that due to differences in evolutionary rate between branches, the timings of these events should be treated with caution.

## Ago2 paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura* are probably functional

Our phylogenetic analysis (Figure 1) revealed that the Ago2 paralogues in the *obscura* group have retained coding sequences for millions of generations, showing that they have remained functional for this period. They have also retained PAZ and PIWI domains and a bilobal structure (characteristic of Argonaute proteins across the tree of life), suggesting that they are part of a functional RNAi pathway. In *D. melanogaster*, Ago2 plays a key role in antiviral immunity, but is ubiquitously and highly expressed in both males and females, and is not strongly induced by viral challenge (Figure 3a, [28]). To test whether this expression pattern has been conserved after Ago2 duplication, or whether any Ago2 paralogues have become inducible by viral challenge, we measured the expression of each Ago2 paralogue in female and male *D. subobscura*, *D. obscura* and *D. pseudoobscura* after infection with Drosophila C Virus (DCV). These species are separated by ~10My of evolution, and represent the three major clades within the *obscura* group. Members of the *obscura* group are known to be highly susceptible to DCV, supporting high viral titres and displaying rapid mortality [29]. We found that only one paralogue is expressed in both sexes at a high level in *D. subobscura* (Ago2a), *D. obscura* (Ago2a) and *D. pseudoobscura* (Ago2c). Unexpectedly, and with only one exception, the other Ago2 paralogues in all species were only expressed in males (Figure 3b-d), raising the possibility that they have specialised to a sex-specific role. The one exception was *D. pseudoobscura* Ago2d, which is the ancestral paralogue in this species (inferred by synteny), and for which we could not detect any expression.

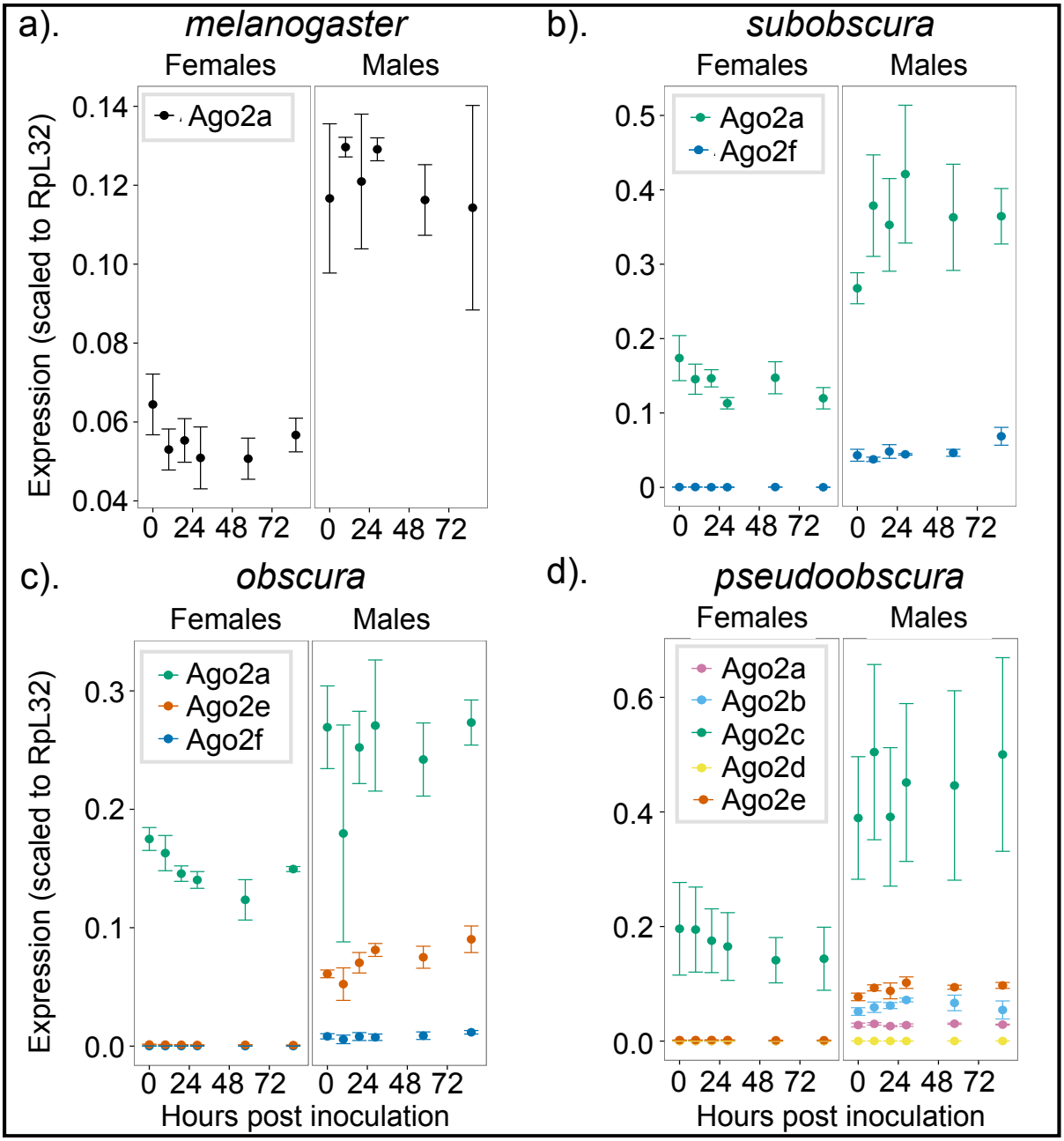



Figure 3: Expression patterns of Ago2 paralogues under challenge with Drosophila C Virus. In each *obscura* group species, only one Ago2 paralogue has retained the ancestral pattern of ubiquitous stable expression in each sex (illustrated by *D. melanogaster*). In contrast, all other paralogues are expressed in males only (apart from *D. pseudoobscura* Ago2d, which is unexpressed in either sex). The high degree of sequence similarity between Ago2a1 and Ago2a3 prevented us from amplifying these genes separately in qPCR, and here they are combined as "Ago2a". Error bars indicate 1 standard error estimated from 2 technical replicates in each of three different genetic backgrounds. Apparent differences in expression between sexes and species should be interpreted with caution, as these may be driven by differences in expression levels of the reference gene (RpL32).

## Ago2 paralogues have repeatedly specialised to the testis

To determine whether the strongly male-biased expression pattern is associated with a testis-specific role, we quantified the tissue-specific expression patterns of Ago2 paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*. In *D. melanogaster* the single copy of Ago2 was expressed in all adult tissues (Figure 4a), and transcripts were present in the embryo (S1 Figure). In *D. subobscura*, *D. obscura* and *D. pseudoobscura*, we found that the Ago2 paralogues exhibited striking differences in their tissue-specific patterns of expression (Figure 4b-d). In each species, one paralogue has retained the ancestral ubiquitous expression pattern in adult tissues. In contrast, every other paralogue was expressed only in the testis, except for the non-expressed *D. pseudoobscura* Ago2d. None of the testis-specific paralogues in *D. pseudoobscura* was detectable in embryos (S1 Figure).

Interestingly, the ubiquitously expressed paralogue in *D. subobscura* and *D. obscura* is the ancestral gene (Ago2a in both cases, as inferred by synteny with *D. melanogaster*), but in *D. pseudoobscura* another paralogue (Ago2c) has evolved the ubiquitous expression pattern, and the ancestral gene (Ago2d) was not expressed at a detectable level in any tissue. When interpreted in the context of the phylogenetic relationships between these paralogues, the most parsimonious explanation is that testis-specificity evolved at least three times: firstly at the base of the Ago2e clade, secondly at the base of the Ago2f clade, and thirdly at the base of the *D. pseudoobscura-D. persimilis* Ago2a-Ago2b subclade (Figure 1).

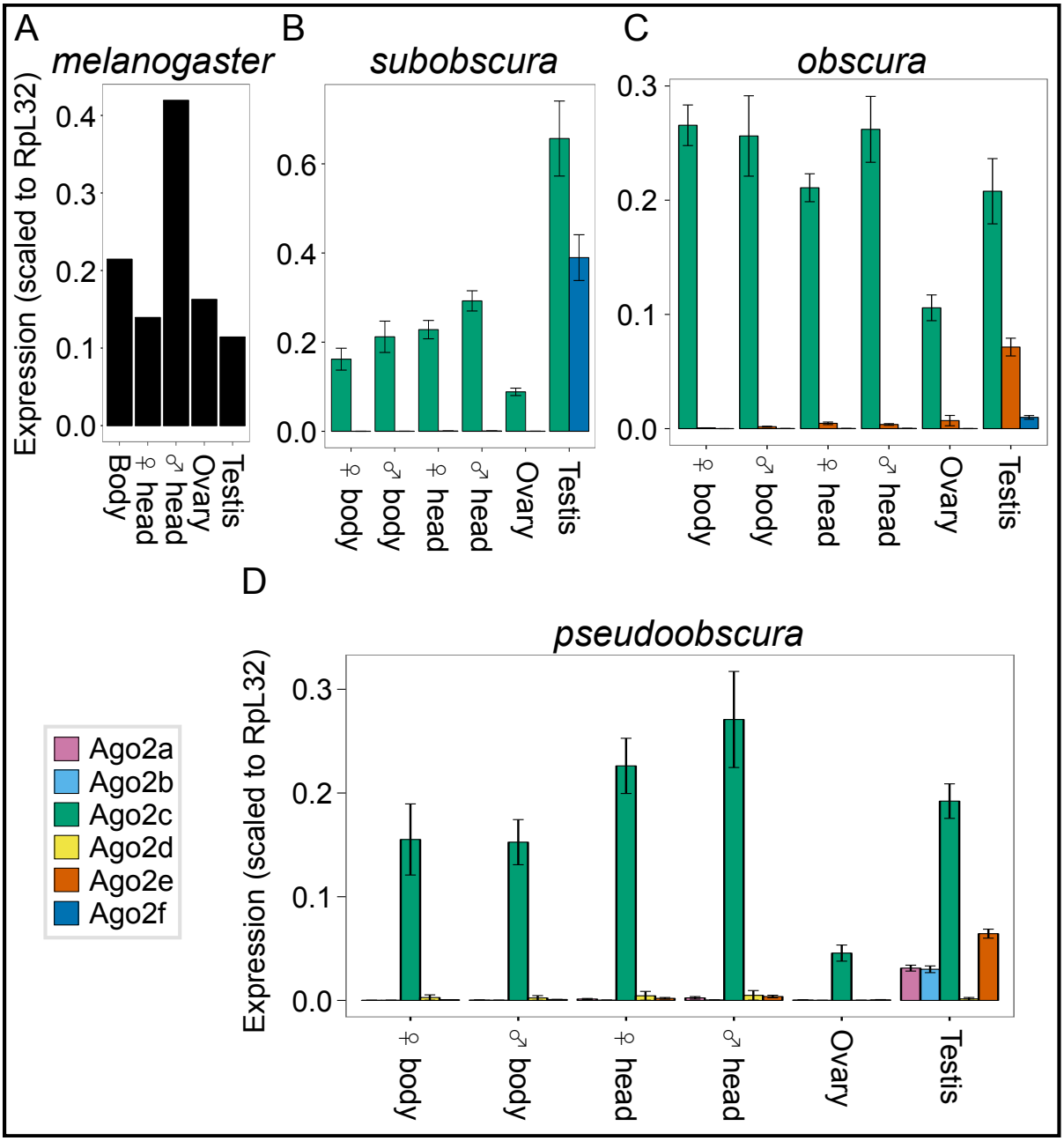



Figure 4: Tissue-specific expression patterns of Ago2 paralogues.
In each of the three *obscura* group species tested, one paralogue has retained the ancestral ubiquitous expression pattern, while the others have specialised to the testis (with the exception of *D. pseudoobscura* Ago2d). The high degree of sequence similarity between Ago2a1 and Ago2a3 prevented us from amplifying these genes separately in qPCR, and here they are combined as "Ago2a". Error bars indicate 1 standard error estimated from 2 technical replicates in each of five different genetic backgrounds. *D. melanogaster* expression levels were taken from a single RNA-seq experiment [71].

150

151 Testis-specificity is associated with faster protein evolution

152 To test for differences in evolutionary rate between testis-specific and ubiquitously expressed Ago2

153 paralogues, we fitted sequence evolution models to the set of drosophilid Ago2 sequences depicted in

Figure 1 using codeml (PAML, Yang 1997). These tests estimate separate dN/dS ratios (ω) and likelihoods for different subclades in the gene tree, providing a test for differential rates of evolution. We found that most support (Akaike weight = 0.99) falls behind a model specifying a different ω for each obscura group Ago2 subclade, and another separate ω for the *D. pseudoobscura-D. persimilis* Ago2a-Ago2b subclade. Under this model, the testis-specific *D. pseudoobscura-D. persimilis* Ago2a-Ago2b subclade has the highest rate of protein evolution (ω=0.32±0.047 SE), followed by the testis-specific Ago2f subclade (ω=0.21±0.014), the ubiquitous Ago2a subclade (ω=0.19±0.012), the testis-specific Ago2e subclade (ω=0.16±0.010), and finally the other Drosophilid Ago2 sequences (ω=0.12±0.002). This shows that the evolution of testis-specificity is generally accompanied by an increase in the rate of protein evolution. We also used the Bayes Empirical Bayes sites test in codeml to identify codons evolving under positive selection across the entire gene tree, and the branch-sites test to identify codons under positive selection in the *obscura* group Ago2 subclade. While we found no positively-selected codons with the sites test, we identified three codons under positive selection (297, 338 & 360) in the *obscura* group Ago2 subclade with the branch-sites test (likelihood ratio test M8 vs M8a, $p<0.005$).

McDonald-Kreitman tests identify strong positive selection on *D. pseudoobscura* Ago2e

This increase in evolutionary rate after the evolution of testis-specificity may have occurred as a result of positive selection, or the relaxation of selective constraint. However, unless there are multiple substitutions within single codons, this will be hard to detect using methods such as codeml. Therefore, as a second test for positive selection on Ago2 paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*, we gathered intraspecies polymorphism data for each Ago2 paralogue in these species (S4 Appendix), and performed McDonald-Kreitman (MK) tests (S1 Table). The MK test uses a comparison of the numbers of fixed differences between species at nonsynonymous (Dn) and synonymous (Ds) sites, and polymorphisms within a species at nonsynonymous (Pn) and synonymous (Ps) sites to infer the action of positive selection. If all mutations are either neutral or strongly deleterious, the Dn/Ds ratio should be approximately equal to the Pn/Ps ratio; however, if there is positive selection, an excess of nonsynonymous differences is expected [31]. The majority of MK tests were non-significant (Fisher's exact test, $p>0.1$), despite often displaying relatively high $K_A/K_S$ ratios e.g. *D. pseudoobscura* Ago2a1 ($K_A/K_S$ =0.34), Ago2b ($K_A/K_S$ =0.43) & Ago2d ($K_A/K_S$ =0.36). However, the low diversity at these loci (<10 polymorphic sites in most cases; see below) will

mean that the MK approach has little power, and that estimates of the proportion of substitutions that are adaptive ($\alpha$) are likely to be poor. In contrast to the other loci, we identified strong positive selection acting on *D. pseudoobscura* Ago2e – which has relatively high genetic diversity – with $\alpha$ at 100% ($\alpha$=1.00; Fisher's exact test, p=0.0004). This result is driven by the extreme skew in the proportion of nonsynonymous to synonymous polymorphisms (0 Pn to 17 Ps), despite substantial numbers of fixed differences (77 Dn to 120 Ds), and is robust to the choice of outgroup (S2 Table).

The majority of Ago2 paralogues have extremely low levels of sequence diversity

When strong selection acts to reduce genetic diversity at a locus, it can also reduce diversity at linked loci before recombination can break up linkage [32]. Recent positive selection can therefore be inferred from a reduction in synonymous site diversity compared with other genes. Because MK tests can only detect multiple long-term substitutions, and are hampered by low diversity, diversity-based approaches offer a complementary way to detect very recent strong selection. We therefore compared the synonymous site diversity at each Ago2 paralogue in *D. pseudoobscura* with the distribution of genome-wide synonymous site diversity. We found that all paralogues have unusually low diversity relative to other loci: Ago2a1, Ago2b and Ago2c fall into the lowest percentile, Ago2a3 and Ago2d into the 2nd lowest percentile and Ago2e into the 8th lowest percentile (S3 Figure). A multi-locus extension of the HKA test (ML-HKA [33]) confirmed that the diversity of Ago2a1-Ago2e is significantly lower than the *D. pseudoobscura* genome as a whole (Akaike weight = 0.98). Unfortunately, population-genomic data are not available for *D. subobscura* and *D. obscura*, preventing a similar analysis. However, we found similar results for Ago2a and Ago2e when comparing the diversity of *D. subobscura* and *D. obscura* Ago2 paralogues to levels of diversity inferred from transcriptome data (data from [34]), suggesting that this effect is not limited to *D. pseudoobscura* and these genes may therefore have been recent targets of selection in multiple species. In *D. obscura*, Ago2a and Ago2e fall into the 2nd and 4th lowest diversity percentile respectively, whereas Ago2f falls into the 19th percentile (S3 Figure). In *D. subobscura*, Ago2a falls into the 7th percentile, whereas Ago2f falls into the 16th percentile (S3 Figure). The prevalence of low intraspecific diversity for testis-specific paralogues is consistent with recent selective sweeps, suggesting that positive selection, not merely relaxation of constraint, has contributed to the increased evolutionary rate seen after specialization to the testis.

213 Four out of six *D. pseudoobscura* Ago2 show a strong signature of recent hard selective sweeps

214 The impact of selection on linked diversity (a selective sweep) is expected to leave a characteristic
215 footprint in local genetic diversity around the site of selection, and this forms the basis of explicit
216 model-based approaches to detect the recent action of positive selection [35]. For *D. pseudoobscura*,
217 population genomic data for 11 haplotypes is available from [36], permitting an explicit model-based
218 test for recent hard selective sweeps near to Ago2 paralogues. We therefore combined our Ago2 data
219 with 111kb haplotypes from [36] to analyse the neighbouring region around each paralogue. Ago2a1
220 and Ago2a3 form a tandem repeat, and were therefore analysed together as a single potential sweep.
221 We found strong evidence for recent selective sweeps at or very close to Ago2a1/3, Ago2b and
222 Ago2c, which display sharp troughs in their diversity levels, and large peaks in the composite
223 likelihood of a sweep, which far exceed a significance threshold derived from coalescent simulation
224 ($p<0.01$) (Figure 5). These localised reductions in diversity remain when our own Ago2 haplotype data
225 is removed, showing the results are robust to the fact that our Ago2 sequence data is derived from a
226 different population to the genome-wide data of [36] (S5 Figure; note that sequence data for Ago2
227 paralogues cannot be derived from the data of reference [36], because of their extreme similarity). In
228 addition, there is ambiguous evidence for a sweep at Ago2d, in the form of one significant ($p<0.01$)
229 likelihood peak just upstream of the paralogue, but two other peaks ~1kb and ~3kb further upstream.
230 There is no evidence for a hard sweep at Ago2e, which has no diversity trough or likelihood peak.

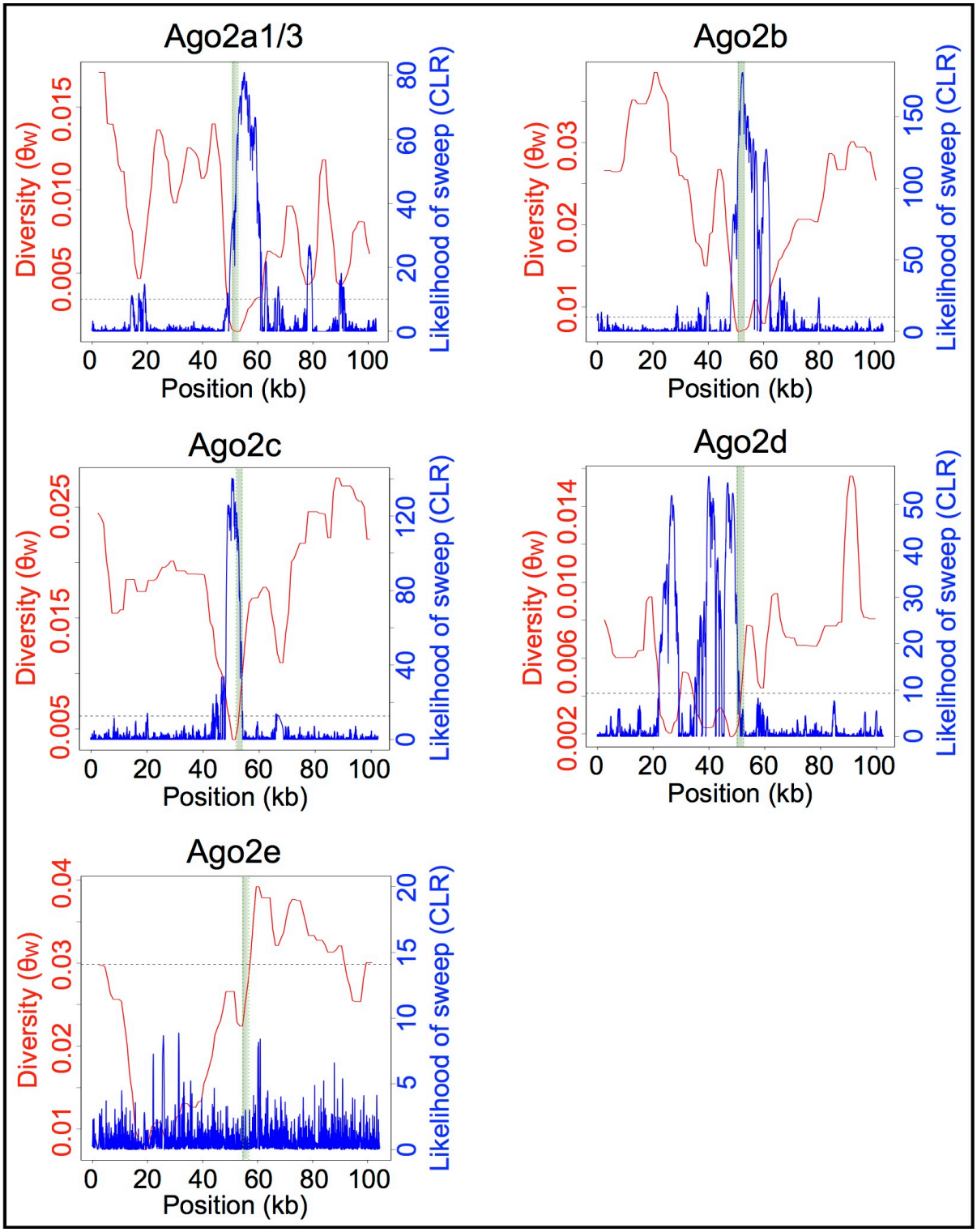



Figure 5: Selective sweeps at *D. pseudoobscura* Ago2 paralogues.
For each paralogue, diversity at all sites (Watterson's θ) is displayed in red, and the likelihood of a sweep centred at that site (composite likelihood ratio, CLR) is displayed in blue. The significance threshold for the CLR is displayed by the horizontal dotted line ($p<0.01$, derived from the 10th-highest CLR out of 1000 coalescent simulations, assuming constant recombination rate and $N_e$). There is strong evidence for sweeps at Ago2a, Ago2b and Ago2c, indicated by troughs in their diversity levels and peaks in the likelihood of a sweep.

231

## Discussion

### Testis-specificity may indicate a loss of antiviral function

We have found that Ago2 paralogues in the *obscura* group have repeatedly evolved divergent expression patterns after duplication, with the majority of paralogues specializing to the testis. This is the first report of testis-specificity for any arthropod Ago2, which is ubiquitously expressed in *D. melanogaster* [37], and provides a strong indication that these paralogues have diverged in function. This testis-specificity (Figure 4), combined with a lack of upregulation on viral challenge (in contrast to virus-responsive genes in the Toll [38] and Jak-STAT [39] signalling pathways in *D. melanogaster*), suggests that these Argonautes are likely to have lost their ancestral ubiquitous antiviral role. In contrast, one paralogue in each species has retained this ubiquitous expression pattern (*D. subobscura* Ago2a, *D. obscura* Ago2a & *D. pseudoobscura* Ago2c, Figure 4), suggesting that these paralogues have retained roles in antiviral defence [18,19], dosage compensation [17] and/or somatic TE suppression [20,21].

### Both ubiquitous and testis-specific Ago2 paralogues show evidence of recent positive selection

We identified selective sweeps at the ubiquitously expressed Ago2 paralogue in *D. pseudoobscura* Ago2c, and very low diversity in the ubiquitously expressed Ago2 paralogues of *D. subobscura* and *D. obscura* (Ago2a), suggesting that all of these genes may have recently experienced strong positive selection. This is consistent with previous findings of strong selection and rapid evolution of Ago2 in *D. melanogaster* [15,22,23] which has also experienced recent sweeps in *D. melanogaster*, *D. simulans*, and *D. yakuba* [23], and across the *Drosophila* more broadly [12]. It has previously been suggested that this is driven by arms-race coevolution with viruses [12,13], some of which encode viral suppressors of RNAi (VSRs) that block Ago2 function [40]. The presence of VSR-encoding viruses, such as Nora virus, in natural *obscura* group populations [34], combined with the host-specificity that VSRs can display [25], suggest that arms-race dynamics may also be driving the rapid evolution of ubiquitously expressed Ago2 paralogues in the *obscura* group.

### Potential testis-specific functions

In contrast to their ancestral ubiquitous expression pattern, the dominant fate for Ago2 paralogues in the *obscura* group appears to have been specialization to the testis. Paralogues often undergo a brief period of testis-specificity soon after duplication [41,42], and this has given rise to the 'out-of-the-

testis' hypothesis, in which new paralogues are initially testis-specific before evolving functions in other tissues [43]. However, two lines of evidence suggest an adaptive basis for the testis-specificity observed for the *obscura* group Ago2 paralogues. First, testis-specificity has been retained for more than 10 million years in Ago2e and Ago2f, in contrast to the broadening of expression over time expected under the out-of-the-testis hypothesis [41,43]. Second, all testis-specific Ago2 paralogues in *D. pseudoobscura* show evidence either of long-term positive selection (MK test for the high-diversity Ago2e) or of recent selective sweeps (in low-diversity Ago2a1/3 and Ago2b), and the testis-specific *D. obscura* Ago2e displays a reduction in diversity, potentially driven by selection.

Under a subfunctionalization model for Ago2 testis-specialization, four candidate selective pressures seem likely: testis-specific dosage compensation, antiviral defence, TE suppression, and/or the suppression of meiotic drive. Of these, testis-specific dosage compensation seems the least likely to drive testis-specificity because the male-specific lethal (MSL) complex, which Ago2 directs to X-linked genes to carry out dosage compensation in the soma of *D. melanogaster*, is absent from testis [44]. Testis-specific antiviral defence seems similarly unlikely, as the only known paternally-transmitted *Drosophila* viruses (Sigmaviruses; Rhabdoviridae) pass through both the male and female gametes [45], and so the potential benefits of testis-specificity seem unclear. In contrast, the suppression of TEs or meiotic drive seem more promising candidate selective forces. First, numerous TEs transpose preferentially in the testis, such as *Penelope* in *D. virilis* [46] and *copia* in *D. melanogaster* [47,48], which could impose a selection pressure on Ago2 paralogues to provide a testis-specific TE suppression mechanism. Nevertheless, it should be noted that all members of the canonical anti-TE Piwi subfamily (Ago3, Aub and Piwi) are also expressed in *obscura* group testis (S2 Figure), suggesting that if Ago2 paralogues have specialised to suppress TEs, they are doing so alongside the existing TE suppression mechanism. Second, testis-specificity could have evolved to suppress meiotic drive, which is prevalent (in the form of sex-ratio distortion) in the *obscura* group [49–53], and which is suppressed by RNAi-based mechanisms in other species [54–56]. A high level of meiotic drive in the *obscura* group could therefore impose selection for the evolution of novel suppression mechanisms, leading to the repeated specialization of Ago2 paralogues to the testis.

Prospects for novel functions during the evolution of RNAi

The functional specialization that we observe for *obscura* group Ago2 paralogues raises the prospect of undiscovered derived functions following Argonaute expansions in other lineages. Ago2 has

duplicated frequently across the arthropods, with expansions present in insects (*Drosophila willistoni* (Figure 1) & *Musca domestica* [57]), crustaceans (*Penaeus monodon* [6]) and chelicerates (*Tetranychus urticae*, *Ixodes scapularis*, *Mesobuthus martensii* & *Parasteatoda tepidariorum* [58]). The prevalence of testis-specificity in *obscura* group Ago2 paralogues raises the possibility that specialization to the germline may be more widespread following Argonaute duplication. The expression of Ago2 paralogues has previously been characterized in *P. monodon*, and shows that one paralogue has indeed specialised to the germline of both males and females, but not the testis alone [6]. Publicly available RNAseq data from the head, gonad and carcass of male and female *Musca domestica* [59] suggests that neither Ago2 paralogue has specialised to the testis (S6 Figure). However, public data from the head, thorax and abdomen of male and female *D. willistoni* [60] shows that one Ago2 paralogue (FBgn0212615) is expressed ubiquitously, while the other (FBgn0226485) is expressed only in the male abdomen (S6 Figure), consistent with the evolution of testis-specificity after duplication. This raises the possibility that a testis-specific selection pressure may be driving the retention and specialization of Ago2 paralogues across the arthropods.

In conclusion, we have identified rapid and repeated evolution of testis-specificity after the duplication of Ago2 in the *obscura* group, associated with low genetic diversity and signatures of strong selection. Ago2 and other RNAi genes have undergone frequent expansions in different eukaryotic lineages [4,61], and have been shown to switch between ubiquitous and germline- or ovary-specific functions in isolated species. This study provides evidence for the evolution of a new testis-specific RNAi function, and suggests that positive selection may act on young paralogues to drive the rapid evolution of novel RNAi mechanisms across the eukaryotes.

## Materials and Methods

### Identification of Ago2 homologues in the Drosophilidae

We used tBLASTx to identify Ago2 homologues in transcriptomes and genomes of 39 species of the Drosophilidae, using previously-characterised Ago2 from the closest possible relative to provide the query for each species. If blast returned partial hits, we aligned all hits from the target species to all Argonautes from the query species, and assigned hits to the appropriate Ago lineage based on a neighbour-joining tree. For each query sequence, we then manually curated partial blast hits into

complete genes using Geneious v5.6.2 (http://www.geneious.com [62]) (see Supplementary Materials for sequence accessions).

Additionally, we used degenerate PCR to identify Ago2 paralogues in *D. azteca* and *D. affinis*, and paralogue-specific PCR with a touchdown amplification cycle to validate the Ago2 paralogues identified in *D. subobscura*, *D. obscura* and *D. pseudoobscura*. For each reaction, unincorporated primers were removed with ExonucleaseI (New England Biolabs) and 5' phosphates were removed with Antarctic Phosphatase (NEB), the PCR products were sequenced by Edinburgh Genomics using BigDye V3 reagents on a capillary sequencer (Applied Biosystems), and Sanger sequence reads were trimmed and assembled using Geneious v.5.6.2 (http://www.geneious.com [62]). We also used a combination of PCR and blast searches to locate *D. pseudoobscura* Ago2a1 & Ago2a3, which lie on the unplaced “Unknown_contig_265” in release 3.03 of the *D. pseudoobscura* genome (all PCR primers are detailed in S4 Table).

Phylogenetic analysis of drosophilid Ago2 paralogues

To characterise the evolutionary relationships between Ago2 homologues in the Drosophilidae, we aligned sequences using translational MAFFT [63] with default parameters. We noted that there is a high degree of codon usage bias in *D. pseudoobscura* Ago2e (effective number of codons (ENC)=34.24) and *D. obscura* Ago2e (ENC=40.36), and a lesser degree in *D. subobscura* Ago2f (ENC=45.63) and *D. obscura* Ago2f (ENC=48.39). To reduce the impact of codon usage bias, which disproportionately affects synonymous sites, we stripped all third positions [64]. We then inferred a gene tree using the Bayesian approach implemented in BEAST v1.8.1 [65] under a nucleotide model, assuming a GTR substitution model, variation between sites modelled by a gamma distribution with four categories, and base frequencies estimated from the data. We used the default priors for all parameters, except tree shape (for which we specified a birth-death speciation model) and the date of the *Drosophila*-*Sophophora* split. To estimate a timescale for the tree, we specified a normal distribution for the date of this node using values based on mutation rate estimates in [66], with a mean value of 32mya, standard deviation of 7mya, and lower and upper bounds of 15mya and 50mya respectively. We ran the analysis for 50 million steps, recording samples from the posterior every 1,000 steps, and inferred a maximum clade credibility tree with TreeAnnotator v1.8.1 [65]. Note that precise date estimates are not a primary focus of this study, but that other calibrations [67,68] would lead to more ancient estimates of divergence, and thus stronger evidence for selective maintenance.

350 Domain architecture and structural modelling of Ago2 paralogues in the *obscura* group

351 To infer the location of each domain in each paralogue identified in *D. subobscura*, *D. obscura* and *D.*
352 *pseudoobscura*, we searched the Pfam database [69]. To test for structural differences between the
353 *D. pseudoobscura* paralogues, we built structural models of each paralogue based on the published
354 X-ray crystallographic structure of human Ago2 [70]. We used the MODELER software in the
355 Discovery Studio 4.0 Modeling Environment (Accelrys Software Inc., San Diego, 2013) to calculate
356 ten models, selected the most energetically favourable for each protein, and assessed model quality
357 with the 3D-profile option in the software. To assess variation in selective pressure across the
358 structure of each paralogue, we mapped polymorphic residues onto each structure using PyMol
359 v.1.7.4.1 (Schrödinger, LLC).

360 Quantification of virus-induced expression of Ago2 paralogues

361 We exposed 48-96hr post-eclosion virgin males and females of *D. melanogaster*, *D. subobscura*, *D.*
362 *obscura* and *D. pseudoobscura* to Drosophila C virus (DCV), by puncturing the thorax with a pin
363 contaminated with DCV at a dose of approximately $4x10^7$ $TCID^{50}$ per ml. Infection with DCV using this
364 method has previously been shown to lead to a rapid and ultimately fatal increase in DCV titre in *D.*
365 *melanogaster* and *obscura* group species [29]. All flies were incubated at 18C on a 12L:12D light
366 cycle, with *D. melanogaster* on Lewis medium and *D. subobscura*, *D. obscura* and *D. pseudoobscura*
367 on banana medium. We sampled 4-7 individuals per species at 0, 8, 16, 24, 48 and 72 hours post
368 infection. At each time-point we extracted RNA using TRIzol reagent (Ambion) and a
369 chloroform/isopropanol extraction, treated twice with TURBO DNase (Ambion), and reverse-
370 transcribed using M-MLV reverse transcriptase (Promega) primed with random hexamers. We then
371 quantified the expression of Ago2 paralogues in these samples by qPCR, using Fast Sybr Green
372 (Applied Biosystems) and custom-designed paralogue-specific qPCR primer pairs (see Table S6 for
373 primer sequences). Due to their high level of sequence similarity (99.9% identity), no primer pair could
374 distinguish between *D. pseudoobscura* Ago2a1 and Ago2a3, so these two genes are presented
375 together as "Ago2a". All qPCR reactions for each sample were run in duplicate, and scaled to the
376 internal reference gene Ribosomal Protein L32 (RpL32). To capture the widest possible biological
377 variation, the three biological replicates for each species each used a different wild-type genetic
378 background (see S3 Table for backgrounds used).

379 Quantification of Ago2 paralogue expression in different tissues and life stages

380 For *D. subobscura*, *D. obscura* and *D. pseudoobscura*, we extracted RNA from the head,
381 testis/ovaries and carcass of 48-96hr post-eclosion virgin adults, with males and females extracted
382 separately. Each sample consisted of 8-15 individuals in *D. subobscura*, 10 individuals in *D. obscura*
383 and 15 individuals in *D. pseudoobscura*. We then used qPCR to quantify the expression of each Ago2
384 paralogue in each tissue, with two technical replicates per sample (reagents, primers and cycling
385 conditions as above). We carried out five replicates per species, each using a different wild-type
386 background (see S3 Table for details of backgrounds used). To provide an informal comparison with
387 the expression pattern of Ago2 before duplication (an "ancestral" expression pattern), we used the
388 BPKM (bases per kilobase of gene model per million mapped bases) values for Ago2 calculated from
389 RNA-seq data from the body (carcass and digestive system), head, ovary and testis of 4 day old *D.*
390 *melanogaster* adults by [71], scaling each BPKM value to the value for RpL32 in each tissue. Due to
391 the design of that experiment, the body data are derived from pooled samples of males and females
392 [71].

393 To quantify expression of Ago2 paralogues in *D. pseudoobscura* embryos, we collected eggs within
394 30 minutes of laying, and used qPCR to measure the expression of each Ago2 paralogue (reagents
395 and primers as above) in two separate wild-type genetic backgrounds (MV8 and MV10). As above,
396 we estimated an ancestral expression pattern of Ago2 before duplication from the BPKM values for
397 Ago2 in 0-2hr old *D. melanogaster* embryos according to [71], scaled to the BPKM value for RpL32 in
398 embryos. To determine any changes in the expression of other *D. pseudoobscura* Argonautes (Ago1,
399 Ago3, Aub & Piwi) that are associated with Ago2 duplication, we measured their expression in adult
400 tissues and embryos as detailed above, and compared this with the expression of the Argonautes in
401 *D. melanogaster* as measured by [71].

402 Testing for evolutionary rate changes associated with tissue-specificity of Ago2

403 We used codeml (PAML, Yang 1997) to fit variants of the M0 model (a single dn/ds ratio, $\omega$) to the 65
404 drosophilid Ago2 homologues shown in Figure 1. In contrast to the tree topology, which was based on
405 1$^{st}$ and 2$^{nd}$ positions only, the alignment for the codeml analysis included all positions. To compare the
406 evolutionary rates of ubiquitously expressed and testis-specific Ago2 paralogues, we fitted a model
407 specifying one $\omega$ for the Ago2 paralogues that were shown to be testis-specific by qPCR, and another

ω for the rest of the tree. We also fitted two models to account for rate variation between the *obscura* group Ago2 subclades. The first model specified a separate ω for the Ago2a subclade, the Ago2e subclade, the Ago2f subclade and the rest of the tree. The second model additionally incorporated an extra ω specified for the *D. pseudoobscura-D. persimilis* Ago2a-Ago2b subclade (which is testis-specific, in contrast with the rest of the *obscura* group Ago2a subclade). We used Akaike weights to assess which model provided the best fit to the data, given the number of parameters.

Sequencing of Ago2 paralogue haplotypes from *D. subobscura*, *D. obscura* and *D. pseudoobscura*

To gain genotype data for the Ago2 paralogues in *D. subobscura*, *D. obscura* and *D. pseudoobscura*, we sequenced the Ago2 paralogues from six males and six females of each species, each from a different wild-collected line (detailed in S3 Table, sequence polymorphism data in S4 Appendix). We extracted genomic DNA from each individual using the DNeasy Blood and Tissue kit (Qiagen), and amplified and Sanger sequenced each Ago2 paralogue from each individual (reagents and PCR primers as above, sequencing primers detailed in S5 Table). We trimmed and assembled Sanger sequence reads using Geneious v.5.6.2 (http://www.geneious.com [62]), and identified polymorphic sites by eye. After sequencing Ago2a (annotated as a single gene in the *D. pseudoobscura* genome), we discovered two very recent Ago2a paralogues (Ago2a1 & Ago2a3), both of which had been cross-amplified. For each *D. pseudoobscura* individual we therefore re-sequenced Ago2a3 using one primer targeted to its neighbouring locus GA22965, and used this sequence to resolve polymorphic sites in the Ago2a1/Ago2a3 composite sequence, thereby gaining both genotypes for each individual. For each Ago2 paralogue, we inferred haplotypes from these sequence data using PHASE [72], apart from the X-linked paralogues (Ago2a1, Ago2a3 & Ago2d) in *D. pseudoobscura* males, for which phase was obtained directly from the sequence data. The hemizygous haploid X-linked sequenced were used in phase inference, and should substantially improve the inferred phasing of female genotypes.

To quantify differences between paralogues in their population genetic characteristics, we aligned haplotypes using translational MAFFT [63], and used DnaSP v.5.10.01 [73] to calculate the following summary statistics for each Ago2 paralogue: π (pairwise diversity, with Jukes-Cantor correction as described in [74]) at nonsynonymous ($\pi_a$) and synonymous ($\pi_s$) sites, Tajima's *D* [75] and the effective number of codons (ENC) [76]. To compare the ENC for each gene with the genome as a whole, we used codonW v1.4.2 [77] to calculate the ENC for the longest ORF from each gene or

transcript in the genomes of *D. subobscura*, *D. obscura* and *D. pseudoobscura* (ORF sets detailed below). In each species, we then compared the ENC values of each Ago2 paralogue with this genome-wide ENC distribution.

Testing for positive selection on Ago2 paralogues in the *obscura* group

We used McDonald-Kreitman (MK) tests [31] to test for positive selection on each Ago2 paralogue. For each paralogue, we chose an outgroup with divergence at synonymous sites ($K_S$) in the range 0.1-0.2 where possible. However, the prevalence of duplications and losses of Ago2 paralogues in the *obscura* group meant that for some tests a suitably divergent extant outgroup sequence did not exist. In these cases, we reconstructed hypothetical ancestral sequences using the M0 model in PAML [30]. To assess the effect of these outgroup choices on our results, we repeated each test with another outgroup, and found no effect of outgroup choice on the significance of any tests, and only marginal differences in estimates of α and $\omega_\alpha$ (results of tests using primary and alternative outgroups are detailed in S1 & S2 Tables).

A complementary approach to identifying positive selection is to test for reduced diversity at a locus compared with the genome as a whole. To compare the diversity of each *D. pseudoobscura* Ago2 paralogue with the genome-wide distribution of synonymous site diversity, we used genomic data for 12 lines generated by [36]. We mapped short reads to the longest ORF for each gene in the R3.2 gene set using Bowtie2 v2.1.0 [78], and estimated synonymous site diversity ($\theta_W$ based on fourfold synonymous sites) at each ORF using PoPoolation [79]. We then plotted the distribution of synonymous site diversity, limited to genes in the size range of 0.75kb - 3kb for comparability with the Ago2 paralogues, and compared the fourfold synonymous site diversity levels of each *D. pseudoobscura* Ago2 paralogue with this distribution. Some *D. pseudoobscura* paralogues are located on autosomes (Ago2b, Ago2c & Ago2e) and some on the X chromosome (Ago2a1, Ago2a3 & Ago2d). Therefore, because of the different population genetic expectations for autosomal and X-linked genes [80], we examined separate distributions for autosomal and X-linked genes. To provide an additional test for reduced diversity at *D. pseudoobscura* Ago2 paralogues, we performed maximum-likelihood Hudson-Kreitman-Aguadé tests [33], using divergence from *D. affinis* and intraspecific polymorphism data for 84 *D. pseudoobscura* loci generated by [81]. We performed 63 tests to encompass all one, two, three, four, five and six-way combinations of the paralogues, and

467 calculated Akaike weights from the resulting likelihood estimates to provide an estimate of the level of
468 support for each combination.

469 To infer a genome-wide distribution of synonymous site diversity for *D. obscura* and *D. subobscura*,
470 for which genomic data are unavailable, we used pooled transcriptome data from wild-collected adult
471 male flies that had previously been generated for surveys of RNA viruses [25,34]. To generate a *de*
472 *novo* transcriptome for each species, we assembled short reads with Trinity r20140717 [82]. For each
473 species, we mapped short reads from the pooled sample to the longest ORF for each transcript,
474 estimated synonymous site diversity at each locus using PoPoolation [79], and plotted the distribution
475 of diversity (as described above for *D. pseudoobscura*). The presence of heterozygous sites in males
476 (identified by Sanger sequencing) confirmed that all Ago2 paralogues in *D. subobscura* and *D.*
477 *obscura* are autosomal: we therefore compared the synonymous site diversity for these paralogues
478 with the autosomal distribution, and do not show the distributions for putatively X-linked genes. Our
479 use of transcriptome data for *D. obscura* and *D. subobscura* will bias the resulting diversity
480 distributions in three ways. First, variation in expression level will cause individuals displaying high
481 levels of expression to be over-represented among reads, downwardly biasing diversity. Second,
482 highly expressed genes are easier to assemble, and highly expressed genes tend to display lower
483 genetic diversity [83,84]. Third, high-diversity genes are harder to assemble, *per se*. However, as all
484 three biases will tend to artefactually reduce diversity in the genome-wide dataset relative to Ago2,
485 this makes our finding that Ago2 paralogues display unusually low diversity conservative.

486 Identifying selective sweeps in Ago2 paralogues of *D. pseudoobscura*

487 To test whether the unusually low diversity seen in the *D. pseudoobscura* Ago2 paralogues is due to
488 recent selection or generally reduced diversity in that region of the genome, we compared diversity at
489 each paralogue to diversity in their neighbouring regions. We obtained sequence data for the 50kb
490 either side of each of these paralogues from the 11 whole genomes detailed in [36]. Note that the very
491 high similarity of these Ago2 paralogues means that they cannot be accurately assembled from short
492 read data, and are not present in the data from [36]. For each genome, we therefore replaced the
493 poorly-assembled region corresponding to the paralogue with one of our own Sanger-sequenced
494 haplotypes, making a set of 11 ca. 102kb sequences for each paralogue. We aligned these
495 sequences using PRANK [85] with default settings, and calculated Watterson's θ at all sites in a
496 sliding window across each alignment, with a window size of 5kb and a step of 1kb. For Ago2a1 and

Ago2a3, which are located in tandem, we analysed the same genomic region. Since our Ago2 haplotypes were sampled from a different North American population of *D. pseudoobscura* to those of [36], an apparent reduction in local diversity might result from differences in diversity between the two populations. Therefore, we also repeated these analyses on a dataset in which our Sanger sequenced haplotypes were removed, leaving missing data.

To test explicitly for selective sweeps at each region, we used Sweepfinder [86] to calculate the likelihood and location of a sweep in or near each Ago2 paralogue. We specified a grid size of 20,000, a folded frequency spectrum for all sites, and included invariant sites. To infer the significance of any observed peaks in the composite likelihood ratio, we used ms [87] to generate 1000 samples of 11 sequences under a neutral coalescent model. We generated separate samples for each region surrounding an Ago2 paralogue, conditioning on the number of polymorphic sites observed in that region, the sequence length equal to the alignment length, and an effective population size at $10^6$ (based on a previous estimate for *D. melanogaster* by [88]). We specified the recombination rate at 5cM/Mb, a conservative value based on previous estimates for *D. pseudoobscura* [36], which will lead to larger segregating linkage groups and therefore a more stringent significance threshold.

Acknowledgements

This work was supported by a Natural Environment Research Council Doctoral Training Grant (NERC DG NE/J500021/1 to SHL), the Academy of Finland (265971 to HS), a University of Edinburgh Chancellor's Fellowship and a Wellcome Trust Research Career Development Fellowship (WT085064 to DJO), and a Wellcome Trust strategic award to the Centre for Immunity, Infection and Evolution (WT095831 to the CIIE). We thank Ben Longdon and Brian Charlesworth for providing us with strains of *D. obscura* and *D. pseudoobscura* respectively, and Francis Jiggins for providing us with DCV.

525 References

526 1. Meister G. Argonaute proteins: functional insights and emerging roles. Nat Rev Genet.
527 2013;14: 447–59.

528 2. Singh RK, Gase K, Baldwin IT, Pandey SP. Molecular evolution and diversification of the
529 Argonaute family of proteins in plants. BMC Plant Biol. 2015;15: 1–16.

530 3. Buck AH, Blaxter M. Functional diversification of Argonautes in nematodes: an expanding
531 universe. 2013;41: 881–6.

532 4. Lewis SH, Salmela H, Obbard DJ. Duplication and diversification of Dipteran Argonaute genes,
533 and the evolutionary divergence of Piwi and Aubergine. Genome Biol Evol. 2016; Advance Ac:
534 1–30.

535 5. Lu H-L, Tanguy S, Rispe C, Gauthier J-P, Walsh T, Gordon K, et al. Expansion of genes
536 encoding piRNA-associated argonaute proteins in the pea aphid: diversification of expression
537 profiles in different plastic morphs. PLoS One. 2011;6: e28051.

538 6. Leebonoi W, Sukthaworn S, Panyim S, Udomkit A. A novel gonad-specific Argonaute 4 serves
539 as a defense against transposons in the black tiger shrimp Penaeus monodon. Fish Shellfish
540 Immunol. 2015;42: 280–288.

541 7. Miesen P, Girardi E, van Rij RP. Distinct sets of PIWI proteins produce arbovirus and
542 transposon-derived piRNAs in Aedes aegypti mosquito cells. Nucleic Acids Res. 2015;43:
543 6545–56.

544 8. Cerutti H, Casas-Mollano JA. On the origin and functions of RNA-mediated silencing: from
545 protists to man. Curr Genet. 2006;50: 81–99.

546 9. Ross RJ, Weiner MM, Lin H. PIWI proteins and PIWI-interacting RNAs in the soma. Nature.
547 2014;505: 353–9.

548 10. Luteijn MJ, Ketting RF. PIWI-interacting RNAs: from generation to transgenerational
549 epigenetics. Nat Rev Genet. 2013;14: 523–34.

550 11. Sienski G, Dönertas D, Brennecke J. Transcriptional silencing of transposons by Piwi and
551 maelstrom and its impact on chromatin state and gene expression. Cell. 2012;151: 964–980.

552 12. Kolaczkowski B, Hupalo DN, Kern AD. Recurrent adaptation in RNA interference genes across

553 the Drosophila phylogeny. Mol Biol Evol. 2011;28: 1033–1042.

554 13. Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. The evolution of RNAi as a defence against
555 viruses and transposable elements. Philos Trans R Soc London Biol Sci. 2009;364: 99–115.

556 14. Eulalio A, Huntzinger E, Izaurralde E. Getting to the Root of miRNA-Mediated Gene Silencing.
557 Cell. 2008;132: 9–14.

558 15. Obbard DJ, Jiggins FM, Halligan DL, Little TJ. Natural selection drives extremely rapid
559 evolution in antiviral RNAi genes. Curr Biol. 2006;16: 580–5.

560 16. Wen J, Duan H, Bejarano F, Okamura K, Fabian L, Brill JA, et al. Adaptive Regulation of
561 Testis Gene Expression and Control of Male Fertility by the Drosophila Harpin RNA Pathway.
562 Mol Cell. 2015;57: 165–178.

563 17. Menon DU, Meller VH. A role for siRNA in X-chromosome dosage compensation in Drosophila
564 melanogaster. Genetics. 2012;191: 1023–8.

565 18. Li H, Li WX, Ding SW. Induction and suppression of RNA silencing by an animal virus.
566 Science. 2002;296: 1319–1321.

567 19. van Rij RP, Saleh M-C, Berry B, Foo C, Houk A, Antoniewski C, et al. The RNA silencing
568 endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster.
569 Genes Dev. 2006;20: 2985–95.

570 20. Czech B, Malone CD, Zhou R, Stark A, Schlingeheyde C, Dus M, et al. An endogenous small
571 interfering RNA pathway in Drosophila. Nature. 2008;453: 798–802.

572 21. Chung W-J, Okamura K, Martin R, Lai EC. Endogenous RNA interference provides a somatic
573 defense against Drosophila transposons. Curr Biol. 2008;18: 795–802.

574 22. Obbard DJ, Welch JJ, Kim K-W, Jiggins FM. Quantifying adaptive evolution in the Drosophila
575 immune system. PLoS Genet. 2009;5: e1000698.

576 23. Obbard DJ, Jiggins FM, Bradshaw NJ, Little TJ. Recent and recurrent selective sweeps of the
577 antiviral RNAi gene Argonaute-2 in three species of Drosophila. Mol Biol Evol. 2011;28: 1043–
578 56.

579 24. Marques JT, Carthew RW. A call to arms: coevolution of animal viruses and host innate
580 immune responses. Trends Genet. 2007;23: 359–364.

581 25. van Mierlo JT, Overheul GJ, Obadia B, van Cleef KWR, Webster CL, Saleh M-C, et al. Novel
582 Drosophila Viruses Encode Host-Specific Suppressors of RNAi. PLoS Pathog. 2014;10:
583 e1004256.

584 26. Hain D, Bettencourt BR, Okamura K, Csorba T, Meyer W, Jin Z, et al. Natural variation of the
585 amino-terminal glutamine-rich domain in Drosophila argonaute2 is not associated with
586 developmental defects. PLoS One. 2010;5: e15264.

587 27. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and
588 PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res.
589 1997;25: 3389–3402.

590 28. Aliyari R, Wu Q, Li H-W, Wang X-H, Li F, Green LD, et al. Mechanism of induction and
591 suppression of antiviral immunity directed by virus-derived small RNAs in Drosophila. Cell Host
592 Microbe. 2008;4: 387–97.

593 29. Longdon B, Hadfield JD, Day JP, Smith SCL, McGonigle JE, Cogni R, et al. The causes and
594 consequences of changes in virulence following pathogen host shifts. PLoS Pathog. 2015;11:
595 e1004728.

596 30. Yang Z. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput
597 Appl Biosci. 1997;13: 555–6.

598 31. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature.
599 1991;351: 652–654.

600 32. Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genet Res. 1974;23:
601 23–35.

602 33. Wright SI, Charlesworth B. The HKA test revisited: a maximum-likelihood-ratio test of the
603 standard neutral model. Genetics. 2004;168: 1071–6.

604 34. Webster CL, Longdon B, Lewis SH, Obbard DJ. Twenty five new viruses associated with the
605 Drosophilidae (Diptera); 2016. Preprint. Available: http://dx.doi.org/10.1101/041665. Accessed
606 21 March 2016.

607 35. Nielsen R, Bustamante C, Clark AG, Glanowski S, Sackton TB, Hubisz MJ, et al. A scan for
608 positively selected genes in the genomes of humans and chimpanzees. PLoS Biol. 2005;3:

609 e170.

610 36. McGaugh SE, Heil CSS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, et al.
611 Recombination modulates how selection affects linked sites in Drosophila. PLoS Biol. 2012;10:
612 e1001422.

613 37. Celniker SE, Dillon LAL, Gerstein MB, Gunsalus KC, Henikoff S, Karpen GH, et al. Unlocking
614 the secrets of the genome. Nature. 2009;459: 927–930.

615 38. Zambon RA, Nandakumar M, Vakharia VN, Wu LP. The Toll pathway is important for an
616 antiviral response in Drosophila. Proc Natl Acad Sci. 2005;102: 7257–62.

617 39. Dostert C, Jouanguy E, Irving P, Troxler L, Galiana-Arnoux D, Hetru C, et al. The Jak-STAT
618 signaling pathway is required but not sufficient for the antiviral response of drosophila. Nat
619 Immunol. 2005;6: 946–953.

620 40. Bronkhorst AW, van Rij RP. The long and short of antiviral defense: small RNA-based
621 immunity in insects. Curr Opin Virol. 2014;7C: 19–28.

622 41. Assis R, Bachtrog D. Neofunctionalization of young duplicate genes in Drosophila. Proc Natl
623 Acad Sci. 2013;110: 17409–14.

624 42. Assis R, Bachtrog D. Rapid divergence and diversification of mammalian duplicate gene
625 functions. BMC Evol Biol. 2015;15: 138.

626 43. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res.
627 2010;20: 1313–26.

628 44. Conrad T, Akhtar A. Dosage compensation in Drosophila melanogaster: epigenetic fine-tuning
629 of chromosome-wide transcription. Nat Rev Genet. 2012;13: 123–134.

630 45. Longdon B, Jiggins FM. Vertically transmitted viral endosymbionts of insects: do sigma viruses
631 walk alone? Proc R Soc B. 2012;279: 3889–3898.

632 46. Rozhkov N V, Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie WR, et
633 al. Small RNA-based silencing strategies for transposons in the process of invading Drosophila
634 species. RNA. 2010;16: 1634–45.

635 47. Pasyukova E, Nuzhdin S, Li W, Flavell AJ. Germ line transposition of the copia
636 retrotransposon in Drosophila melanogaster is restricted to males by tissue-specific control of

637 copia RNA levels. Mol Gen Genet. 1997;255: 115–124.

638 48. Morozova T V, Tsybulko EA, Pasyukova EG. Regularory elements of the copia
639 retrotransposon determine different levels of expression in different organs of males and
640 females of Drosophila melanogaster. Genetika. 2009;45: 169–177.

641 49. Gershenson S. A New Sex-Ratio Abnormality in Drosophila obscura. Genetics. 1928;13: 488–
642 507.

643 50. Sturtevant AH, Dobzhansky T. Geographical Distribution and Cytology of “Sex Ratio” in
644 Drosophila Pseudoobscura and Related Species. Genetics. 1936;21: 473–490.

645 51. Wu CI, Beckenbach AT. Evidence for extensive genetic differentiation between the sex-ratio
646 and the standard arrangement of Drosophila pseudobscura and D. persimilis and identification
647 of hybrid sterility factors. Genetics. 1983;105: 71–86.

648 52. Jaenike J. Sex chromosome meiotic drive. Annu Rev Ecol Syst. 2001;32: 25–49.

649 53. Unckless RL, Larracuente AM, Clark AG. Sex-ratio meiotic drive and Y-linked resistance in
650 Drosophila affinis. Genetics. 2015;199: 831–40.

651 54. Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, Hartl DL. A sex-ratio meiotic drive system in
652 Drosophila simulans. II: An X-linked distorter. PLoS Biol. 2007;5: 2576–2588.

653 55. Kotelnikov RN, Klenov MS, Rozovsky YM, Olenina L V., Kibanov M V., Gvozdev V a.
654 Peculiarities of piRNA-mediated post-transcriptional silencing of Stellate repeats in testes of
655 Drosophila melanogaster. Nucleic Acids Res. 2009;37: 3254–3263.

656 56. Gell SL, Reenan RA. Mutations to the piRNA pathway component aubergine enhance meiotic
657 drive of segregation distorter in Drosophila melanogaster. Genetics. 2013;193: 771–784.

658 57. Scott JG, Warren WC, Beukeboom LW, Bopp D, Clark AG, Giers SD, et al. Genome of the
659 house fly, Musca domestica L., a global vector of diseases with adaptations to a septic
660 environment. Genome Biol. 2014;15: 466–482.

661 58. Palmer WJ, Jiggins FM. Comparative Genomics Reveals the Origins and Diversity of
662 Arthropod Immune Systems. Mol Biol Evol. 2015;32: 2111–2129.

663 59. Meisel RP, Scott JG, Clark AG. Transcriptome Differences between Alternative Sex
664 Determining Genotypes in the House Fly, Musca domestica. Genome Biol Evol. 2015;7: 2051–

2061.

60. Meisel RP, Malone JH, Clark AG. Disentangling the relationship between sex-biased gene expression and X-linkage. Genome Res. 2012;22: 1255–1265.

61. Mukherjee K, Campos H, Kolaczkowski B. Evolution of animal and plant dicers: early parallel duplications and recurrent adaptation of antiviral RNA binding in plants. Mol Biol Evol. 2013;30: 627–41.

62. Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 2012;28: 1647–1649.

63. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT : a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30: 3059–3066.

64. Behura SK, Severson DW. Codon usage bias: causative factors, quantification methods and genome-wide patterns: with emphasis on insect genomes. Biol Rev. 2013;88: 49–61.

65. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol. 2012;29: 1969–1973.

66. Obbard DJ, MacLennan J, Kim KW, Rambaut A, O'Grady PM, Jiggins FM. Estimating divergence dates and substitution rates in the Drosophila phylogeny. Mol Biol Evol. 2012;29: 3459–3473.

67. Russo C a, Takezaki N, Nei M. Molecular phylogeny and divergence times of Drosophilid species. Mol Biol Evol. 1995;12: 391–404.

68. Tamura K. Temporal Patterns of Fruit Fly (Drosophila) Evolution Revealed by Mutation Clocks. Mol Biol Evol. 2004;21: 36–44.

69. Finn RD, Mistry J, Tate J, Coggill P, Heger a., Pollington JE, et al. The Pfam protein families database. Nucleic Acids Res. 2009;38: D211–D222.

70. Schirle NT, Macrae IJ. The Crystal Structure of Human Argonaute2. Science. 2012;336: 1037–1040.

71. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, et al. Diversity and dynamics of the Drosophila transcriptome. Nature. 2014;512: 393–399.

693 72. Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from
694 population data. Am J Hum Genet. 2001;68: 978–989.

695 73. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA polymorphism
696 data. Bioinformatics. 2009;25: 1451–2.

697 74. Lynch M, Crease TJ. The analysis of population survey data on DNA sequence variation. Mol
698 Biol Evol. 1990;7: 377–394.

699 75. Tajima F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.
700 Genetics. 1989;123: 585–595.

701 76. Wright F. The “effective number of codons” used in a gene. Gene. 1990;87: 23–29.

702 77. Peden J. Analysis of codon usage bias. PhD Thesis, The University of Nottingham. 1995.
703 Available:
704 http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.186.1796&rep=rep1&type=pdf

705 78. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of
706 short DNA sequences to the human genome. Genome Biol. 2009;10: R25.

707 79. Kofler R, Orozco-terWengel P, De Maio N, Pandey RV, Nolte V, Futschik A, et al. PoPoolation:
708 a toolbox for population genetic analysis of next generation sequencing data from pooled
709 individuals. PLoS One. 2011;6: e15925.

710 80. Vicoso B, Charlesworth B. Evolution on the X chromosome: unusual patterns and processes.
711 Nat Rev Genet. 2006;7: 645–653.

712 81. Haddrill PR, Loewe L, Charlesworth B. Estimating the parameters of selection on
713 nonsynonymous mutations in Drosophila pseudoobscura and D. miranda. Genetics. 2010;185:
714 1381–96.

715 82. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
716 transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol.
717 2011;29: 644–52.

718 83. Pal C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. Genetics. 2001;158:
719 927–931.

720 84. Lemos B, Bettencourt BR, Meiklejohn CD, Hartl DL. Evolution of proteins and gene expression

levels are coupled in Drosophila and are independently associated with mRNA abundance, protein length, and number of protein-protein interactions. Mol Biol Evol. 2005;22: 1345–1354.

85. Löytynoja A, Goldman N. An algorithm for progressive multiple alignment of sequences with insertions. Proc Natl Acad Sci. 2005;102: 10557–10562.

86. Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome Res. 2005;15: 1566–75.

87. Hudson RR. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 2002;18: 337–338.

88. Li H, Stephan W. Inferring the demographic history and rate of adaptive substitution in Drosophila. PLoS Genet. 2006;2: 1580–1589.

89. Anders S, Pyl PT, Huber W. HTSeq - A Python framework to work with high-throughput sequencing data. Bioinformatics. 2015;2: 166–169.

90. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5: 621–628.

91. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, et al. Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred from genetic and physical maps. Genetics. 2008;179: 1601–55.

92. Segarra C, Aguadé M. Molecular organization of the X chromosome in different species of the obscura group of Drosophila. Genetics. 1992;130: 513–521.

748 Supporting Information Captions

749 **S1 Figure: The expression of *D. pseudoobscura* Ago2 paralogues in embryos.** Error bars
750 indicate 1 standard error estimated from 2 technical replicates in each of two different genetic
751 backgrounds. *D. melanogaster* expression levels were taken from a single publicly-available RNA-seq
752 experiment [71]. Ago2c is highly expressed in embryos, but none of the testis-specific Ago2
753 paralogues (Ago2a, Ago2b & Ago2e) are expressed.

754 **S2 Figure: The tissue-specific expression patterns of other members of the Argonaute gene**
755 **family (Ago1, Ago3, Aub & Piwi) in *D. melanogaster* and *D. pseudoobscura*.** For *D.*
756 *pseudoobscura* embryo, error bars indicate 1 standard error estimated from 2 technical replicates in
757 each of two different genetic backgrounds. For all other *D. pseudoobscura* tissues, error bars indicate
758 1 standard error estimated from 2 technical replicates in each of five different genetic backgrounds. *D.*
759 *melanogaster* expression levels were taken from a single RNA-seq experiment [71]. In *D.*
760 *pseudoobscura*, Ago1 is expressed in all tissues, but the other genes are only expressed in the
761 embryo and germline.

762 **S3 Figure: The distribution of synonymous site diversity across genes, derived from genome**
763 **(*D. pseudoobscura*) or transcriptome (*D. subobscura* & *D. obscura*) data.** The percentile of the
764 distribution into which each paralogue falls is indicated in brackets under the paralogue name. In each
765 species, members of the Ago2a and Ago2e subclades have very low diversity compared with the
766 genome as a whole.

767 **S4 Figure: The distribution of codon usage bias, derived from genome (*D. pseudoobscura*) or**
768 **transcriptome (*D. subobscura* & *D. obscura*) data.** The percentile of the distribution into which
769 each paralogue falls is indicated in brackets under the paralogue name. Ago2e has a very low
770 effective number of codons (ENC) compared with the genome as a whole, indicating a high degree of
771 codon usage bias.

772 **S5 Figure: Genetic diversity in the regions surrounding each *D. pseudoobscura* Ago2**
773 **paralogue, with Ago2 paralogue haplotype sequences removed.** After specifying Ago2 paralogue
774 sequence data as missing information, sharp troughs in diversity remain at Ago2a, Ago2b and Ago2c,
775 indicating a selective sweep.

776 **S6 Figure: The tissue-specific expression patterns of the Argonaute gene family in *D. willistoni***

777 **and *M domestica*.** Transcriptome data for *D. willistoni* were taken from [60], and transcriptome data
778 for *M. domestica* were taken from [59]. For both species, we mapped reads to coding sequences
779 using Bowtie 2.1 [78], counted reads mapping to each coding sequence using HTSeq [89], and
780 converted counts to reads per kilobase per million reads (RPKM [90]) to account for coding sequence
781 length and sequencing depth. For *M. domestica*, error bars indicate two biological replicates, each in
782 a different genetic background.

783 **S1 Table: McDonald-Kreitman test results.** Pn & Ps are the number of within-species
784 polymorphisms after singletons have been removed. All values are displayed to 2dp, except $\omega_a$ which
785 is displayed to 4dp.

786 **S2 Table: McDonald-Kreitman test results with alternative outgroups.** Pn & Ps are the number of
787 within-species polymorphisms after singletons have been removed. All values are displayed to 2dp,
788 except $\omega_a$ which is displayed to 4dp.

789 **S3 Table: Genetic backgrounds used in each experiment.** Line refers to an individual isofemale
790 line, and Origin refers to the geographic location where the female who founded that line was caught.

791 **S4 Table: Primers used for PCR and qPCR amplification of Ago2 paralogues.** All primers are
792 displayed in the 5' to 3' direction.

793 **S5 Table: Primers used for Sanger sequencing of Ago2 paralogue haplotypes.** All primers are
794 displayed in the 5' to 3' direction.

795 **S1 Appendix: Sequence alignment of drosophilid Ago2 homologues.** This alignment has had all
796 3$^{rd}$ positions stripped, and was used for time-scaled phylogenetic analysis of drosophilid Ago2
797 evolution.

798 **S2 Appendix: Sequence alignment of drosophilid Ago2 homologues.** This alignment has had all
799 3$^{rd}$ positions stripped, and was used for model-based analysis of differential evolutionary rate and
800 codon-specific positive selection.

801 **S3 Appendix: Sequence metadata for drosophilid Ago2 homologues.**

802 **S4 Appendix: Sequence polymorphism data for *D. subobscura*, *D. obscura* and *D.***
803 ***pseudoobscura* Ago2 paralogues**

804 **S5 Appendix: Raw data used to plot Figures 3, 4, 5, S1, S2, S3, S4 & S6.**