Widespread gene duplication and adaptive evolution in the RNA interference pathways of the *Drosophila obscura* group

Danang Crysnanto[1,3*]

Darren J. Obbard[1,2]

Affiliation

1. Institute of Evolutionary Biology, University of Edinburgh, Charlotte Auerbach Road, Edinburgh, United Kingdom

2. Centre for Infection, Evolution and Immunity, University of Edinburgh, Edinburgh, United Kingdom

3. Present address: Animal Genomics, ETH Zurich, Zurich, Switzerland

Email

DC: danang.crysnanto@usys.ethz.ch

DJO: darren.obbard@ed.ac.uk

*Author for correspondence

ABSTRACT

Background: RNA interference (RNAi) related pathways provide defense against viruses and transposable elements, and have been implicated in the suppression of meiotic drive elements. Genes in these pathways often exhibit high levels of adaptive substitution, and over longer timescales show frequent gene duplication and loss—most likely as a consequence of their role in mediating conflict with these parasites. This is particularly striking for *Argonaute 2* (*Ago2*), which is ancestrally the key effector of antiviral RNAi in insects, but has repeatedly formed new testis-specific duplicates in the recent history of the *Drosophila obscura* group.

Results: Here we take advantage of publicly available genomic and transcriptomic data to identify six further RNAi-pathway genes that have duplicated in this clade of *Drosophila*, and examine their evolutionary history. As seen for *Ago2*, we observe high levels of adaptive amino-acid substitution and changes in sex-biased expression in many of the paralogs. However, our phylogenetic analysis suggests that co-duplications of the RNAi machinery were not synchronous, and our expression analysis fails to identify consistent male-specific expression.

Conclusions: These results confirm that RNAi genes, including genes of the antiviral and piRNA pathways, undergo frequent independent duplications and that their history has been particularly labile within the *Drosophila obscura* group. However, they also suggest that the selective pressures driving these changes have not been consistent, implying that more than one selective agent may be responsible.

Keywords: gene duplication, RNAi, RNA interference, adaptive evolution, neofunctionalization

Introduction

Gene duplication is an important process in molecular evolution, providing raw genetic material for evolutionary innovation. The subsequent evolutionary dynamics following gene duplication are often described in terms of two alternative models, 'neofunctionalization' and 'sub-functionalization' [1]. Under neofunctionalization, the functional redundancy following duplication provides relaxed selective constraint, and allows new mutations to accumulate through genetic drift. Most such mutations will reduce the functionality of the gene (resulting in pseudogenization), but some paralogs can be selected for novel or derived functions. Under sub-functionalization, the duplicates independently accumulate mutations that allow them to specialise in a subset of ancestral functions of a pleiotropic gene. Neo-functionalization leads to asymmetrical evolutionary rates among paralogs (with faster evolution in paralogs that gain derived function), whereas equal rates are expected for the latter [2]. It has been suggested that both processes have played an important role in the rapid evolution of RNA interference-related pathways, including the long- and short-term evolutionary history of the Argonautes, the effectors of RNAi [3–5].

The RNAi-related pathways comprise a range of small-RNA mechanisms best known for their roles in mediating the control of gene expression, antiviral responses, and defence against mobile genetic elements (respectively: the miRNA pathway; Dicer-1 and Argonaute-1 in insects [6]; the siRNA pathway; Dcr-2 and Argounate 2 in insects [7]; and the piRNA pathway; piwi-family Argonaute AGO3 and Piwi/Aub in insects [8, 9]). In addition, RNAi-related pathways have been implicated in a variety of biological processes, such as the control of dosage compensation [10–12] and the suppression of genetic drive [13–18], among others. Several genes involved in the defensive piRNA

and siRNA pathways, but not the miRNA pathway, display elevated rates of adaptive protein evolution. This is best studied in *Drosophila* [19–21], but is also detectable in other insects [22]. It has been hypothesised that this is a consequence of parasite-mediated 'arms-race' coevolution [20, 23], either through conflict with parasite-encoded immune suppressors—as widely seen in RNA viruses [24]— or in the case of the piRNA pathway, through selection for 're-tuning' suppression mechanisms [25].

Adaptive evolution of RNAi pathways is also partly reflected in the gain, loss, and functional divergence of Argonaute-family duplications [26]. For example, within the Drosophilidae—an important model for RNAi-related pathways of animals—Piwi has been duplicated in the lineages leading to *Phortica variegata* and *Scaptodrosophila deflexa* [3], and *Ago2* has been duplicated in those leading to *S. deflexa*, *D. willistoni*, *D. melanogaster* (where only one paralog remains – the canonical *Ago2*) and *D. pseudoobscura* [4]. This is particularly striking in the *Drosophila obscura* species group, which has experienced at least 6 independent duplications of *Ago2* over the last 20 million years, with all but one of the resulting duplicates becoming testis-specific, and most displaying evidence of recent and/or ongoing positive selection [4].

Very recently it has been noted that several accessory components of the siRNA and piRNA pathways have also been duplicated in *D. pseudoobscura*, including *armitage*, *asterix*, *cutoff*, *maelstrom*, *tejas* and *vreteno* [22]. In *D. melanogaster*, these proteins are engaged in a number of roles in the piRNA pathway (**Table 1**). Here we use publicly available data to reconfirm the history and expression of *Ago2* in the obscura group, and to test whether duplications in the other genes also show male-specific expression, whether they are contemporaneous with those of *Ago2*, and whether they too show strong signatures of adaptive protein evolution. We find no clear pattern of these duplications being coincident with *Ago2* duplications, but both

101 *asterix* and *cutoff* duplications display increased sexual dimorphism relative to their
102 ancestral copies, through decreased female expression. In addition, several of the
103 gene duplicates show evidence of adaptive protein evolution in *D. pseudoobscura*,
104 including both copies of *cutoff*, the ancestral copy of *asterix*, and the new duplicates
105 of *tejas*, *maelstrom* and *vreteno*.

106 **RESULTS**

107 **Obscura group *Argonaute 2* are duplicated and show male-biased expression**

108 The *Drosophila obscura* group has experienced multiple duplications of *Ago2* and it
109 has previously been shown that these are associated with positive selection and testis-
110 specific expression [4]. Here we reanalyzed the expression patterns and evolutionary
111 history of these genes using publicly available RNAseq and genomic data, additionally
112 including newly available genomic sequences from *D. algonquin*, *D. athabasca*, *D.*
113 *bifasciata* and *D. miranda*. In contrast to the previous qPCR analysis that failed to
114 identify strong expression of the ancestral copy in *D. pseudoobscura* (*Ago2d* [4]), we
115 found that all the *Ago2* homologs in *D. pseudoobscura* were detectable at a high level
116 in RNAseq data, and that that all show significant male-bias (**Figure 1**). The *Ago2d*
117 expression detected here is unlikely to be an artefact of cross-mapping between
118 paralogs as we observed the reads that mapped uniquely across the gene. The male
119 bias was largest for *Ago2e*, where expression in males is approximately 1000-fold
120 higher than females (pMCMC<0.001; **Figure 1**), and smallest in *Ago2c* and the
121 ancestral copy *Ago2d*, consistent with the *ca.* two-fold enrichment of the single copy
122 of *Ago2* in male *D. melanogaster*. We also confirmed that *D. miranda*, a close relative
123 of *D. pseudoobscura* that has not previously been analyzed, displayed a qualitatively
124 similar pattern among those paralogs represented (**Figure 1**). In *D. obscura* and we

125 found the ancestral copy (*Ago2a*) again showed slightly, but marginally significantly,
126 higher expression in males (pMCMC=-0.014), but that other *Ago2* proteins showed a
127 strong male biased expression, with the largest effect for *Ago2f*, where male
128 expression was 2000-fold higher (**Figure 1**; pMCMC<0.001).

129 **Six piRNA pathway genes are duplicated, and *asterix* and *cutoff* duplicates show**
130 **increased male-bias in their expression**

131 Palmer *et al.* [22] recently identified six accessory piRNA pathway genes that have
132 also experienced duplication in the obscura group (**Additional file 1 Table S1**). We
133 could locate the duplicates for all genes in *D. pseudoobscura*, except for *armitage*
134 where we instead identified a duplicate in the affinis subgroup but not in the obscura
135 or subobscura subgroups. Where new chromosomal locations could be determined
136 by synteny in *D. pseudoobscura*, we found that *cutoff*, *maelstrom* and *vreteno* were
137 duplicated from an autosome to the X chromosome, *asterix* duplicated from the X
138 chromosome to an autosome, and *tejas* duplicated between autosomal locations. Two
139 duplicates (*asterix* and *tejas*) lack introns, suggesting they are retro-transcribed copies
140 created through an mRNA intermediate.

141 Using public RNAseq data from *D. pseudoobscura*, *D. miranda*, and *D. obscura*, we
142 found that all of the gene duplicates were expressed (**Figure 2**). *Armitage*, which was
143 not duplicated within the newly examined lineages for which RNAseq data were
144 available, did not show strong sex-biased expression. Similarly, *tejas*, *maelstrom*, and
145 *vreteno* were not strongly differentially expressed between the sexes, and nor were
146 their duplicates in *D. pseudoobscura* and *D. miranda*. In contrast, both *asterix* and
147 *cutoff* duplicates displayed substantially reduced expression in females and slightly
148 increased expression in males (**Figure 2**). For example, as previously reported from

149 qPCR analysis [27] the paralog of *asterix* in *D. pseudoobscura* displays ca. 1000-fold
150 higher expression in males than females. In both *D. pseudoobscura* and *D. miranda*,
151 those genes with overall strongly increased male-biased expression (*Argonaute 2*,
152 *asterix*, *cutoff,* and their paralogs) had the highest expression in testis, and had
153 reduced their expression in ovaries (**Additional file 2 Figure S1**).

154 Adaptive amino-acid substitutions are generally more common in the 155 duplicates

156 Using population genetic data from *D. pseudoobscura*, and *D. miranda a*s an
157 outgroup, we used the McDonald-Kreitman framework and a maximum-likelihood
158 extension to estimate the rate of adaptive substitution in protein sequences, and to
159 test whether this was different between the ancestral and duplicated copies [28, 29].
160 Treating genes individually, we found evidence for positive selection acting on at least
161 one paralog for each of the genes except *asterix* and *Ago2c* ($p < 0.05$; **Additional file
162 3 Table S2**). Among the ancestral copies, only *cutoff* displayed evidence of positive
163 selection. We then tested whether the paralogs generally showed a different pattern
164 of selection to the ancestral copies by dividing the genes into two classes (6 ancestral
165 copies and 8 paralogs) and comparing the likelihood of models that allowed the
166 classes to differ in the adaptive rate α (**Table 2**) [29]. The best supported model
167 allowed α to differ between ancestral and duplicate copies (Akaike weight: 0.81), and
168 the second best supported model was that in which α=0 for the ancestral copies
169 (Akaike weight: 0.19), providing overall evidence that the paralogs have experienced
170 more adaptive protein evolution. In the best-supported model, the α was estimated to
171 be 0.68 for the duplicate group, which more than three times larger than the ancestral
172 group (0.20). In case segregating weakly-deleterious variants had led to a downward
173 bias in estimates of α, we repeated this analysis excluding all alleles with a minor allele

174 frequency <0.125 [30], although this reduced power to the extent that few genes
175 remained individually significant (**Additional file 6 Table S3**). We also repeated the
176 analysis with a larger dataset PRJNA326536 [31] (**Additional file 6 Table S3**), and
177 obtained qualitatively similar results (R^2=0.946 for α estimate between the analyses;
178 the second dataset, while larger, is less suitable for analysis as only the third
179 chromosome is a direct sample from a wild population).

180 Gene duplications were unlikely to be contemporaneous

181 Given the multiple duplications of *Ago2* and the piRNA pathway components in the
182 obscura group, we hypothesized that some duplications may have occurred near-
183 simultaneously, duplicating whole components of a pathway together. We therefore
184 used relaxed-clock phylogenetic methods to estimate the relative timings of each
185 duplication. In agreement with the previous analysis of Ago2 [4], we found that the
186 duplications giving rise to *Ago2e* and *Ago2f* predated the split between the obscura
187 and pseudoobscura subgroups, with a subsequent loss of *Ago2f* from the
188 pseudoobscura subgroup (**Figure 3**). In contrast, we found that duplications in five of
189 the other six genes unambiguously occurred after the obscura/pseudoobscura split,
190 with the timing of duplication in *maelstrom* being uncertain. Briefly, *armitage* displayed
191 a single duplication shared by members of the affinis subgroup, *asterix* and *tejas* a
192 single duplication each in the lineage leading to *D. pseudoobscura* (which were
193 subsequently lost in the affinis subgroup), *cutoff* a single duplication recently in the
194 pseudoobscura subgroup, and *vreteno* a single duplication at the base of the obscura
195 group (**Figure 4**). For *maelstrom*, the maximum clade credibility tree suggests
196 duplication occurred very slightly prior to this split, followed by subsequent loss of one
197 paralog the obscura subgroup (**Figure 4**). However, this was poorly supported, and
198 similar pattern to the others genes may be more parsimonious. We used the posterior

199 distributions of split times, relative to the divergence time of the obscura and
200 pseudoobscura subgroups, to infer whether or not duplications occurred at
201 approximately the same time (**Figure 5**). Although the small amount of information
202 available from single genes made relative timings highly uncertain, it is clear that few
203 of the *Ago2* duplications could have been concurrent with the piRNA-pathway
204 duplications (**Additional file 5 Figure S2**). However, the recent and rapid duplications
205 within the piRNA could have been concurrent, with *vreteno*, *tejas, maelstrom* and
206 *asterix* not differing significantly, all having duplicated very close to the split between
207 *D. obscura* and *D. pseudoobscura* (posterior overlap >0.1 in each case; **Additional**
208 **file 5 Figure S2**).

209 DISCUSSION

210 Although four of the six piRNA pathway duplicates did not display altered tissue
211 specificity compared to the ancestral copy, *asterix* and *cutoff* both became significantly
212 more male biased, as did each of the *Ago2* duplicates [4]; **Figure 1**, **Figure 2**). In each
213 case, this was due to higher (or exclusive) expression in the testis. The duplicated
214 genes also showed higher rates of adaptive amino acid substitution, together and
215 individually, whereas only two (*asterix* and *armitage*) displayed evidence of positive
216 selection when single-copy in *D. melanogaster* (**Additional file 6 Table S3**).

217 This new tissue specificity and the rapid evolution of duplicated copies broadly suggest
218 that gene duplication in these pathways may be associated with functional
219 diversification through neofunctionalization, for example by testis-specific selective
220 pressure. Three main selective pressures seem likely candidates to have driven this.
221 First, given the role of *Ago2* in antiviral defense in *Drosophila* [7], and the role of the
222 piRNA pathway in antiviral defense in mosquitoes [32], it is possible that these

223 duplications have specialized to a virus that is active in the male germline, such as *D.*
224 *obscura* and *D. affinis* Sigmaviruses [33]. Second, given the role of all of these genes
225 in the suppression of transposable elements (TEs), their evolution may have been
226 shaped by the invasion of TEs that are more active in testis, as seen for Penelope [34]
227 and copia [35]. Such a 'duplication arms-race' in response to TE invasion is thought
228 to occur in mammals, where repeated duplications of KRAB-ZNF family are selected
229 following the invasion of novel TEs, and subsequently provide defence [36, 37].
230 Alternatively, duplicates may quantitatively enhance the pre-existing response to TEs,
231 as suggested for another rapidly-evolving piRNA-pathway component, Rhino [38].

232 The third, and arguably most compelling hypothesis, is that selection is mediated by
233 conflict with meiotic drive elements, such as sex ratio distorting X-chromosomes [39,
234 40]. Most directly, meiotic drive elements are common in *Drosophila*, and RNAi-related
235 pathways have been widely implicated in their action and suppression [15, 16, 18]. In
236 addition, sex-chromosome drive is widespread in the *Drosophila obscura* group. X
237 chromosome drive was first described in *D. obscura*, and has also been reported in
238 *pseudoobscura*, *persimilis*, *affinis*, *azteca*, *subobscura* and *athabasca* [39] and is
239 mediated through a testis-specific function (Y-bearing sperm have reduced function).
240 Finally, a testis-specific class of hairpin (endo) siRNAs is required for male fertility in
241 *D. melanogaster* [41], testes-restricted clustered miRNAs show rapid evolutionary
242 turnover and are represented in large numbers in *D. pseudoobscura* [42], and
243 suppression of sex-specific duplicates of S-Lap1 via a small-RNA mechanism has very
244 recently been implicated in the meiotic drive mechanism of *D. pseudoobscura* [17]. In
245 this context, it is also interesting to note that *Ago2* is involved in directing
246 heterochromatin formation in *Drosophila* dosage compensation [10–12], and that in *D.*

247 *melanogaster* the sex-ratio distorting *Spiroplasma* achieves male-killing through the
248 disruption of dosage compensation, although this acts at the embryonic stage [43].

249 Nevertheless, in the absence of mechanistic studies, this remains highly speculative.
250 Testis is generally more permissive to gene expression and testis-specific expression
251 may be a transient state (i.e. the "Out of Testis Hypothesis" [44]). In addition, the
252 application of MK-like analyses to paralogs is inherently flawed [1], as the MK
253 framework implicitly assumes that the selective regime has been consistent across all
254 (group and outgroup) sequences analyzed. If gene duplicates experience an early but
255 transient period of relaxed constraint, a high proportion of the amino-acid fixations may
256 have occurred as a result of genetic drift that is no longer detectable from current
257 patterns of polymorphism.

258 Methods

259 Sequence collation and paralog identification

260 The full-length sequences for 7 RNA inteference genes; *armitage* (*armi*), *asterix* (*arx*),
261 *cutoff* (*cuff*), *tejas (tej)*, *vreteno* (*vret*), *maelstrom* (*mael*), and *Argonaute 2* (*Ago2*) from
262 12 obscura group species were identified using tBLASTn (BLAST+ 2.6.0) [45] with a
263 local BLAST database (see below for the details of the construction of local genomic
264 database). Known gene sequences from *D. pseudoobscura* and *D. melanogaster*
265 were used as a query with a stringent e-value threshold (1e-40). Genes were inferred
266 to have been duplicated when BLAST indicated that there were multiple full-length hits
267 located in different genomic regions. The sequences were manually inspected, introns
268 removed and the coding frame identified using Bioedit v 7.2 [46]. Genes in *D.*
269 *pseudoobscura* were classified as ancestral or duplicate copies based on the syntenic

270 orthology with *D. melanogaster* using Flybase Genome Browser [47]. High quality
271 genomes are not available for other members of the obscura groups, and in those
272 cases ancestral/derived status was assigned based on homology with *D.*
273 *pseudoobscura*. To provide a comprehensive overview of the evolution of the RNAi
274 paralogs, we included 24 *Drosophila* species outside of *obscura* group with assembled
275 genomes already available in public databases. The Flybase and NCBI tblastn online
276 portal were used to identify the target genes with queries from *D. melanogaster* or
277 closely related species.

278 Five obscura group species had assembled genomes at the time of this study: *D.*
279 *pseudoobscura* (assembly Dpse_3.0 [48]), *D. miranda* (assembly DroMir_2.2 [49]
280 [50]), *D. persimilis* (assembly dper_caf1 [51]), *D. affinis* (Drosophila affinis Genome
281 Release 1.0 [52]) and *D. lowei* (Drosophila lowei Genome Release 1.0 [52, 53]) and
282 in these cases the genome was directly used for local BLAST database. For four
283 species (*D. obscura*, *D. subobscura*, *D. subsilvestris*, *D. tristis*) we used de novo
284 assembled transcriptomes based on paired RNA-seq reads data from wild-collected
285 males [54] (Accession: PRJNA312496). Assembly was performed using Trinity [55]
286 with '--trimmomatric' and otherwise default parameters, and the assembled
287 transcriptome was searched locally using BLAST. For three other species: *D.*
288 *athabasca*, *D. Algonquin* [56] (Accession: PRJNA274695) and *D. bifasciata*
289 (Accession: PRJDB4817), only unassembled genomic reads were available. For these
290 species we applied a targeted assembly approach as follows: (i) reads that had local
291 similarity with all known duplicated RNAi proteins were identified using Diamond [57]
292 with relaxed e-value of 1; (ii) hits from Diamond were then retained and used for
293 assembly using Spades v3.10.1 [58]; and (iii) scaffolds produced by Spades were then
294 used as references in local BLAST database.

295 ## Phylogenetic analysis and the relative timing of duplications

296 Bayesian relaxed clock trees were used to infer the evolutionary relationship among
297 paralogs. First, the sequences were aligned as translated nucleotide in Clustal W [59]
298 with default parameters. Regions with ambiguous alignment were identified and
299 removed manually by eye. A total of 7 gene trees were then inferred using Beast v1.7.0
300 [60]. Inference used a relaxed clock model with an uncorrelated lognormal distribution
301 among branches, and an HKY substitution model with empirical base frequencies and
302 rate variation among sites was modelled as a gamma distribution with four categories.
303 The site model allowed for third codon position to have different substitution model
304 from the other positions.

305 The trees were scaled by setting the time to most recent common ancestor of the *D.*
306 *obscura* group to have lognormal distribution with a data-scale mean of 1, and a very
307 small standard deviation of 0.01. This had the advantage of scaling all duplications to
308 the same relative timescale, while allowing different genes and different paralogs to
309 vary in their rate. To record the posterior ages of duplication, we specified the ancestral
310 and duplicated genes as a distinct taxon set. The Monte Carlo Markov Chain analysis
311 was run for at least 100 million states and posterior sample was recorded every 10000
312 states. Log files were then inspected in Tracer v1.6 [61] for parameter stationarity, and
313 adequate sampling as indicated by an effective sample size over 200. Finally, 25% of
314 initial trees were discarded as burn-in, and maximum clade credibility trees were
315 summarized using Tree Annotator. Parameter MCMC files were processed using a
316 custom R script [62] to infer the posterior distribution the age of duplication for each
317 gene and to quantify the degree overlapping between these age distributions.

318

319 ## Differential expression analysis of the duplicated RNAi genes

320 For this analysis, we used obscura group transcriptome datasets available in EBI ENA
321 (European Nucleotide Archive, http://www.ebi.ac.uk/ena) and DDBJ (DNA DataBank
322 of Japan, http://www.ddbj.nig.ac.jp/) that included the sex and tissue annotation. The
323 datasets comprised 163, 42 and 34 RNA-seqs datasets of *D. pseudoobscura, D.*
324 *miranda* and *D. obscura* respectively; Bioproject: DRA004463, PRJEB1227 [63],
325 PRJNA226598, PRJNA219224 [64], PRJNA326536 [31], PRJNA74723,
326 PRJNA321079, PRJNA291085 [65], PRJNA268967 [66]. Since our main interest was
327 the comparison expression between sex, but not its absolute expression value, we a
328 performed a simple read-counting analysis. In outline, each RNA-seq dataset was
329 mapped to the full-length CDS using Bowtie2 v2.3.2 [67] with mode '--very sensitive'
330 and otherwise default parameters. The reads mapped to reference were counted
331 using combination of SAMtools view flag -F 4 and SAMtools idxstats v1.4 [68]. The
332 count data were then normalized by gene length and read depth, where it was then
333 scaled relative to the expression of RpL32. To determine the statistical significance of
334 difference gene expression, generalised linear mixed models were fitted using R
335 package MCMCglmm [69] with sex as fixed effect and tissue as a random effect, and
336 log-transformed normalised expression as the response variable. The natural
337 logarithm transformation ($\log_e$) was used to reduce the skewness of the distributions.
338 To allow for zero value for non-expressed genes, the genes with read count 0 was
339 replaced with 1.

340 $$\mathbf{Y} \sim \boldsymbol{\mu} + \mathbf{sex} + \mathbf{tissue}\ (\mathbf{random}) + \boldsymbol{\varepsilon}$$

341 Where Y is $\log_e$ transformed normalized expression data (response variable), μ is
342 mean of $\log_e$ transformed expression and ε is residual error. The random effects

343 (tissue) and the residual were assumed to be distributed multivariate normal with mean
344 0 and uncorrelated covariance matrix MVN (0, $\mathbf{I}\sigma^2$). Sex was modelled as a factor with
345 2 variables (male-female) and tissue contained 13 variables of different tissue.

346 Population Genetic Analysis of the RNAi Duplicated Genes

347 We used the McDonald-Krietman test [28] to compare the rate of adaptive evolution
348 between ancestral and duplicate genes using polymorphism data from publicly-
349 available sequencing datasets: Pseudobase (12 strains of *pseudoobscura*, Accession
350 list: SRP007802 [53]) and 12 strains *D. miranda* (Bioproject: PRJNA277849 [70]).

351 Genomic reads for each strain were mapped to the genomic reference using Bowtie2
352 with '--very-sensitive' mode and otherwise default parameters and reads mapped to
353 the genes of interest were extracted using SAMtools view (flag -F 4). Duplicate reads
354 were marked using MarkDuplicates (Picard Tools [71]). To reduce the excessive
355 variants surrounding indel, we then applied GATK IndelRealigner [71], which discards
356 the original mapping and performs local-realignment around indel. The output was
357 then sorted and indexed and the BAM file was used for 'mpileup' variant calling
358 (SAMtools v1.4 [72]). The output VCF files were then filtered to only include SNP
359 (GATK SelectVariants [71]), and variants that were covered by less than five reads
360 were masked with 'N' (undetermined bases, --snpmask GATK v3.5 [71]). The variant
361 files were then converted to FASTA format using GATK
362 FastaAlternateReferenceMaker, which replaced genomic reference with variants
363 defined in VCF files [73] and output the heterozygous calls with IUPAC ambiguous
364 code. Finally, FastPHASE [74] was used to generate pseudo-haplotypes, although
365 haplotype information was not utilized by the analysis.

366 MK tests were performed for each gene on *D. pseudoobscura-D miranda.* DNAsp v5.0
367 [75] was used to estimate the statistics for the MK test and Fisher's exact test was
368 used to calculate the statistical significance for single-gene analyses. Genes were
369 then grouped into ancestral and duplicate genes, and a cross-gene analysis was
370 performed using a maximum likelihood extension of the MK test [29]. Five different
371 models were fitted that differed in the constraint of α (proportion of non-synonymous
372 subtitutions estimated to be adaptive), and the relative support between models was
373 compared using Akaike Weights.

374 **List of Abbreviations**

375 **Ago2**: Agronaute 2 **armi**: armitage **arx**: asterix **BAM**: Binary Alignment Map **BLAST**:
376 Basic Alignment Search Tool **cuff**: cutoff **GATK**: Genome Analysis Toolkit **HKY**:
377 Hasegawa, Kishino and Yano model **mael**: maelstrom **MCMC**: Monte Carlo Markov
378 Chain **MK:** McDonald-Kreitman test **RNAi**: RNA interference **SNP**: Single Nucleotide
379 Polymorphism **TE:** Transposable Elements **tej**: tejas **VCF**: Variant Call Format **vret**:
380 vreteno

381 **Declarations**

382 **Ethics approval and consent to participate**

383 Not applicable

384 **Consent to publication**

385 Not applicable

386 **Availability of data and material**

387 Fasta alignment (both for phylogenetic and MK analysis) and raw expression data are
388 available via Figshare (DOI: 10.6084/m9.figshare.7145720).

Competing interests

The authors declare that they have no competing interests.

Funding

DC was financially supported by a Master's Training Scholarship from the Indonesian Endowment Fund for Education (LPDP) and the University of Edinburgh School of Biological Science Bursary for MSc in Quantitative Genetics and Genome Analysis.

Authors' contributions

DJO and DC conceived the study and designed the analysis, DC analyzed the data and wrote the first draft of the manuscript. Both authors read and approved the final manuscript.

Acknowledgements

We thank Billy Palmer for initial discussion on the variant calling, and Billy Palmer and Samuel Lewis for comments on an earlier version of this manuscript. We thank the many people who made their published data publicly available, and Shu Kondo for permission to use unpublished data from *D. bifasciata*.

404 ## References

405 1. Hahn MW. Distinguishing Among Evolutionary Models for the Maintenance of Gene
406 Duplicates. J Hered. 2009;100:605–17. doi:10.1093/jhered/esp047.

407 2. Innan H, Kondrashov F. The evolution of gene duplications: classifying and distinguishing
408 between models. Nat Rev Genet. 2010;11:97–108. doi:10.1038/nrg2689.

409 3. Lewis SH, Salmela H, Obbard DJ. Duplication and diversification of dipteran argonaute
410 genes, and the evolutionary divergence of Piwi and Aubergine. Genome Biol Evol.
411 2016;8:507–18.

412 4. Lewis SH, Webster CL, Salmela H, Obbard DJ. Repeated duplication of Argonaute2 is
413 associated with strong selection and testis specialization in Drosophila. Genetics.
414 2016;204:757–69.

415 5. Singh RK, Gase K, Baldwin IT, Pandey SP. Molecular evolution and diversification of the
416 Argonaute family of proteins in plants. BMC Plant Biol. 2015;15:23. doi:10.1186/s12870-014-
417 0364-6.

418 6. Vidigal JA, Ventura A. The biological functions of miRNAs: lessons from in vivo studies.
419 Trends Cell Biol. 2015;25:137–47. doi:10.1016/j.tcb.2014.11.004.

420 7. Bronkhorst AW, van Rij RP. The long and short of antiviral defense: small RNA-based
421 immunity in insects. Curr Opin Virol. 2014;7:19–28. doi:10.1016/J.COVIRO.2014.03.010.

422 8. Czech B, Hannon GJ. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-
423 Guided Silencing. Trends Biochem Sci. 2016;41:324–37. doi:10.1016/j.tibs.2015.12.008.

424 9. Lewis SH, Quarles KA, Yang Y, Tanguy M, Frézal L, Smith SA, et al. Pan-arthropod
425 analysis reveals somatic piRNAs as an ancestral defence against transposable elements.
426 Nat Ecol Evol. 2018;2:174–81. doi:10.1038/s41559-017-0403-4.

427 10. Menon DU, Meller VH. A role for siRNA in X-chromosome dosage compensation in
428 Drosophila melanogaster. Genetics. 2012;191:1023–8. doi:10.1534/genetics.112.140236.

429 11. Tang W, Seth M, Tu S, Shen E-Z, Li Q, Shirayama M, et al. A Sex Chromosome piRNA
430 Promotes Robust Dosage Compensation and Sex Determination in C. elegans. Dev Cell.
431 2018;44:762–770.e3. doi:10.1016/j.devcel.2018.01.025.

432 12. Deshpande N, Meller VH. Chromatin That Guides Dosage Compensation Is Modulated
433 by the siRNA Pathway in Drosophila melanogaster. Genetics. 2018;209:1085–97.
434 doi:10.1534/genetics.118.301173.

435 13. Tao Y, Masly JP, Araripe L, Ke Y, Hartl DL. A sex-ratio Meiotic Drive System in
436 Drosophila simulans. I: An Autosomal Suppressor. PLoS Biol. 2007;5:e292.
437 doi:10.1371/journal.pbio.0050292.

438 14. Tao Y, Araripe L, Kingan SB, Ke Y, Xiao H, Hartl DL. A sex-ratio Meiotic Drive System in
439 Drosophila simulans. II: An X-linked Distorter. PLoS Biol. 2007;5:e293.
440 doi:10.1371/journal.pbio.0050293.

441 15. Gell SL, Reenan RA. Mutations to the piRNA pathway component aubergine enhance
442 meiotic drive of segregation distorter in Drosophila melanogaster. Genetics. 2013;193:771–
443 84. doi:10.1534/genetics.112.147561.

444 16. Aravin AA, Klenov MS, Vagin V V, Bantignies F, Cavalli G, Gvozdev VA. Dissection of a
445 natural RNA silencing process in the Drosophila melanogaster germ line. Mol Cell Biol.
446 2004;24:6742–50. doi:10.1128/MCB.24.15.6742-6750.2004.

447 17. Ellison C, Leonard C, Landeen E, Gibilisco L, Phadnis N, Bachtrog D. Rampant cryptic
448 sex chromosome drive in Drosophila. doi:10.1101/324368.

449 18. Lin C-J, Hu F, Dubruille R, Smibert P, Loppin B, Correspondence ECL. The
450 hpRNA/RNAi Pathway Is Essential to Resolve Intragenomic Conflict in the Drosophila Male
451 Germline. 2018. doi:10.1016/j.devcel.2018.07.004.

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

454 20. Obbard DJ, Gordon KHJ, Buck AH, Jiggins FM. The evolution of RNAi as a defence
455 against viruses and transposable elements. Philos Trans R Soc Lond B Biol Sci.
456 2009;364:99–115. doi:10.1098/rstb.2008.0168.

457 21. Kolaczkowski B, Hupalo DN, Kern AD. Recurrent Adaptation in RNA Interference Genes
458 Across the Drosophila Phylogeny. Mol Biol Evol. 2011;28:1033–42.
459 doi:10.1093/molbev/msq284.

460 22. Palmer WH, Hadfield JD, Obbard DJ. RNA-Interference Pathways Display High Rates of
461 Adaptive Protein Evolution in Multiple Invertebrates. Genetics. 2018;208:1585–99.
462 doi:10.1534/genetics.117.300567.

463 23. Marques JT, Carthew RW. A call to arms: coevolution of animal viruses and host innate
464 immune responses. Trends Genet. 2007;23:359–64. doi:10.1016/J.TIG.2007.04.004.

465 24. van Mierlo JT, Overheul GJ, Obadia B, van Cleef KWR, Webster CL, Saleh M-C, et al.
466 Novel Drosophila Viruses Encode Host-Specific Suppressors of RNAi. PLoS Pathog.
467 2014;10:e1004256. doi:10.1371/journal.ppat.1004256.

468 25. Blumenstiel JP, Erwin AA, Hemmer LW. What drives positive selection in the Drosophila
469 piRNA machinery? The genomic autoimmunity hypothesis. Yale J Biol Med. 2016;89:499–
470 512. http://www.ncbi.nlm.nih.gov/pubmed/28018141. Accessed 24 Jul 2017.

471 26. Dowling D, Pauli T, Donath A, Meusemann K, Podsiadlowski L, Petersen M, et al.
472 Phylogenetic Origin and Diversification of RNAi Pathway Genes in Insects. Genome Biol
473 Evol. 2017;1:evw281. doi:10.1093/gbe/evw281.

474 27. Meisel RP, Hilldorfer BB, Koch JL, Lockton S, Schaeffer SW. Adaptive Evolution of
475 Genes Duplicated from the Drosophila pseudoobscura neo-X Chromosome. Mol Biol Evol.
476 2010;27:1963–78. doi:10.1093/molbev/msq085.

477 28. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila.
478 Nature. 1991;351:652–654. doi:10.1038/350055a0.

479 29. Welch JJ. Estimating the genomewide rate of adaptive protein evolution in drosophila.
480 Genetics. 2006;173:821–37. doi:10.1534/genetics.106.056911.

481 30. Charlesworth J, Eyre-Walker A. The McDonald-Kreitman Test and Slightly Deleterious
482 Mutations. Mol Biol Evol. 2008;25:1007–15. doi:10.1093/molbev/msn005.

483 31. Fuller ZL, Haynes GD, Richards S, Schaeffer SW. Genomics of Natural Populations:
484 How Differentially Expressed Genes Shape the Evolution of Chromosomal Inversions in
485 Drosophila pseudoobscura. Genetics. 2016;204:287–301. doi:10.1534/genetics.116.191429.

486 32. Campbell CL, Black WC, Hess AM, Foy BD. Comparative genomics of small RNA
487 regulatory pathway components in vector mosquitoes. BMC Genomics. 2008;9:425.
488 doi:10.1186/1471-2164-9-425.

489 33. Longdon B, Wilfert L, Obbard DJ, Jiggins FM. Rhabdoviruses in two species of
490 drosophila: Vertical transmission and a recent sweep. Genetics. 2011;188:141–50.
491 doi:10.1534/genetics.111.127696.

492 34. Rozhkov N V., Aravin AA, Zelentsova ES, Schostak NG, Sachidanandam R, McCombie
493 WR, et al. Small RNA-based silencing strategies for transposons in the process of invading
494 Drosophila species. RNA. 2010;16:1634–45. doi:10.1261/rna.2217810.

495 35. Pasyukova, S. Nuzhdin, W. Li E, Nuzhdin S, Li W, Flavell AJ. Germ line transposition of
496 the copia retrotransposon in Drosophila melanogaster is restricted to males by tissue-
497 specific control of copia RNA levels. Mol Gen Genet MGG. 1997;255:115–24.
498 doi:10.1007/s004380050479.

499 36. Thomas JH, Schneider S. Coevolution of retroelements and tandem zinc finger genes.
500 Genome Res. 2011;21:1800–12. doi:10.1101/gr.121749.111.

501 37. Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, et al. An
502 evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1
503 retrotransposons. Nature. 2014;516:242–5. doi:10.1038/nature13760.

504 38. Levine MT, Wende Vander HM, Hsieh E, Baker EP, Malik HS. Recurrent gene
505 duplication diversifies genome defense repertoire in Drosophila. Mol Biol. 2016;33:1–13.

506 39. Jaenike J. Sex Chromosome Meiotic Drive. Annu Rev Ecol Syst. 2001;32:25–49.
507 doi:10.1146/annurev.ecolsys.32.081501.113958.

508 40. Lindholm AK, Dyer KA, Firman RC, Fishman L, Forstmeier W, Holman L, et al. The
509 Ecology and Evolutionary Dynamics of Meiotic Drive. Trends Ecol Evol. 2016;31:315–26.
510 doi:10.1016/J.TREE.2016.02.001.

511 41. Wen J, Duan H, Bejarano F, Okamura K, Fabian L, Brill JA, et al. Adaptive regulation of
512 testis gene expression and control of male fertility by the Drosophila hairpin RNA pathway.
513 [Corrected]. Mol Cell. 2015;57:165–78. doi:10.1016/j.molcel.2014.11.025.

514 42. Mohammed J, Flynt AS, Panzarino AM, Mondal MMH, DeCruz M, Siepel A, et al. Deep
515 experimental profiling of microRNA diversity, deployment, and evolution across the
516 Drosophila genus. Genome Res. 2018;28:52–65. doi:10.1101/gr.226068.117.

517 43. Harumoto T, Anbutsu H, Lemaitre B, Fukatsu T. Male-killing symbiont damages host's
518 dosage-compensated sex chromosome to induce embryonic apoptosis. Nat Commun.
519 2016;7:12781. doi:10.1038/ncomms12781.

520 44. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res.
521 2010;20:1313–26. doi:10.1101/gr.101386.109.

522 45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool.
523 J Mol Biol. 1990;215:403–10. doi:10.1016/S0022-2836(05)80360-2.

524 46. Hall Thomas. BioEdit: a user-firendly biological sequence alignment editor and analysis
525 program for Windows 95/95/NT. Oxford Univ. 1999;41:95–8. doi:citeulike-article-id:691774.

526 47. St. Pierre SE, Ponting L, Stefancsik R, McQuilton P. FlyBase 102—advanced
527 approaches to interrogating FlyBase. Nucleic Acids Res. 2014;42:D780–8.
528 doi:10.1093/nar/gkt1092.

529 48. English AC, Richards S, Han Y, Wang M, Vee V, Qu J, et al. Mind the Gap: Upgrading
530 Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. PLoS One.
531 2012;7:e47768. doi:10.1371/journal.pone.0047768.

532 49. Alekseyenko AA, Ellison CE, Gorchakov AA, Zhou Q, Kaiser VB, Toda N, et al.
533 Conservation and de novo acquisition of dosage compensation on newly evolved sex
534 chromosomes in Drosophila. Genes Dev. 2013;27:853–8. doi:10.1101/gad.215426.113.

535 50. Zhou Q, Bachtrog D. Sex-specific adaptation drives early sex chromosome evolution in
536 Drosophila. Science (80-). 2012;337:341–5. doi:10.1126/science.1225385.

537 51. Drosophila 12 Genomes Consortium AG, Clark AG, Eisen MB, Smith DR, Bergman CM,
538 Oliver B, et al. Evolution of genes and genomes on the Drosophila phylogeny. Nature.
539 2007;450:203–18. doi:10.1038/nature06341.

540 52. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. Elife.
541 2014;3:e01311. doi:10.7554/eLife.01311.

542 53. McGaugh SE, Heil CSS, Manzano-Winkler B, Loewe L, Goldstein S, Himmel TL, et al.
543 Recombination Modulates How Selection Affects Linked Sites in Drosophila. PLoS Biol.
544 2012;10:e1001422. doi:10.1371/journal.pbio.1001422.

545 54. Webster CL, Waldron FM, Robertson S, Crowson D, Ferrari G, Quintana JF, et al. The
546 Discovery, Distribution, and Evolution of Viruses Associated with Drosophila melanogaster.
547 PLOS Biol. 2015;13:e1002210. doi:10.1371/journal.pbio.1002210.

548 55. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al. Full-length
549 transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol.
550 2011;29:644–52.

551 56. Wong Miller KM, Bracewell RR, Eisen MB, Bachtrog D. Patterns of Genome-Wide
552 Diversity and Population Structure in the Drosophila athabasca Species Complex. Mol Biol
553 Evol. 2017. doi:10.1093/molbev/msx134.

554 57. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat
555 Methods. 2014;12:59–60. doi:10.1038/nmeth.3176.

556 58. Nurk S, Bankevich A, Antipov D, Gurevich A, Korobeynikov A, Lapidus A, et al.
557 Assembling genomes and mini-metagenomes from highly chimeric reads. In: Lecture Notes
558 in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture
559 Notes in Bioinformatics). Springer, Berlin, Heidelberg; 2013. p. 158–70. doi:10.1007/978-3-
560 642-37195-0_13.

561 59. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: Improving the sensitivity of
562 progressive multiple sequence alignment through sequence weighting, position-specific gap
563 penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
564 doi:10.1093/nar/22.22.4673.

565 60. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees.
566 BMC Evol Biol. 2007;7:214. doi:10.1186/1471-2148-7-214.

567 61. Rambaut A. Tracer v1.6. http://tree.bio.ed.ac.uk/software/tracer/. 2013.

568 62. R Core Team. R: A language and environment for statistical computing. R Foundation
569 for Statistical Computing, Vienna, Austria. 2016. https://www.r-project.org/.

570 63. Chen Z-X, Sturgill D, Qu J, Jiang H, Park S, Boley N, et al. Comparative validation of the
571 D. melanogaster modENCODE transcriptome annotation. Genome Res. 2014;24:1209–23.
572 doi:10.1101/gr.159384.113.

573 64. VanKuren NW, Vibranovski MD. A novel dataset for identifying sex-biased genes in
574 Drosophila. J genomics. 2014;2:64–7. doi:10.7150/jgen.7955.

575 65. Nyberg KG, Machado CA. Comparative Expression Dynamics of Intergenic Long
576 Noncoding RNAs in the Genus *Drosophila*. Genome Biol Evol. 2016;8:1839–58.
577 doi:10.1093/gbe/evw116.

578 66. Gomes S, Civetta A. Hybrid male sterility and genome-wide misexpression of male
579 reproductive proteases. Sci Rep. 2015;5:11976. doi:10.1038/srep11976.

580 67. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment
581 of short DNA sequences to the human genome. Genome Biol. 2009;10:R25. doi:10.1186/gb-

582 2009-10-3-r25.

583 68. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence
584 Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.
585 doi:10.1093/bioinformatics/btp352.

586 69. Hadfield JD. MCMC Methods for Multi-Response Generalized Linear Mixed Models: The
587 MCMCglmm R Package. J Stat Softw. 2010;33:1–22. doi:10.18637/jss.v033.i02.

588 70. Smukowski Heil CS, Ellison C, Dubin M, Noor MAF. Recombining without Hotspots: A
589 Comprehensive Evolutionary Portrait of Recombination in Two Closely Related Species of
590 Drosophila. Genome Biol Evol. 2015;7:2829–42. doi:10.1093/gbe/evv182.

591 71. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et
592 al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best
593 practices pipeline. Curr Protoc Bioinforma. 2013; SUPL.43:11.10.1-11.10.33.
594 doi:10.1002/0471250953.bi1110s43.

595 72. Li H. A statistical framework for SNP calling, mutation discovery, association mapping
596 and population genetical parameter estimation from sequencing data. Bioinformatics.
597 2011;27:2987–93. doi:10.1093/bioinformatics/btr509.

598 73. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant
599 call format and VCFtools. Bioinformatics. 2011;27:2156–8.
600 doi:10.1093/bioinformatics/btr330.

601 74. Scheet P, Stephens M. A Fast and Flexible Statistical Model for Large-Scale Population
602 Genotype Data: Applications to Inferring Missing Genotypes and Haplotypic Phase. Am J
603 Hum Genet Am J Hum Genet. 2006;7878:629–44. www.ajhg.org. Accessed 17 Jul 2017.

604 75. Librado P, Rozas J. DnaSP v5: a software for comprehensive analysis of DNA
605 polymorphism data. Bioinformatics. 2009;25:1451–2. doi:10.1093/bioinformatics/btp187.

606 76. Leader DP, Krause SA, Pandit A, Davies SA, Dow JAT. FlyAtlas 2: a new version of the
607 Drosophila melanogaster expression atlas with RNA-Seq, miRNA-Seq and sex-specific data.
608 Nucleic Acids Res. 2018;46:D809–15. doi:10.1093/nar/gkx976.

609 77. Mohn F, Sienski G, Handler D, Brennecke J. The Rhino-Deadlock-Cutoff Complex
610 Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. Cell.
611 2014;157:1364–79. doi:10.1016/j.cell.2014.04.031.

612 78. Zamparini AL, Davis MY, Malone CD, Vieira E, Zavadil J, Sachidanandam R, et al.
613 Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA
614 biogenesis in Drosophila. Development. 2011;138:4039–50. doi:10.1242/dev.069187.

615 79. Saito K, Ishizu H, Komai M, Kotani H, Kawamura Y, Nishida KM, et al. Roles for the Yb
616 body components Armitage and Yb in primary piRNA biogenesis in Drosophila. Genes Dev.
617 2010;24:2493–8. doi:10.1101/gad.1989510.

618 80. Vourekas A, Zheng K, Fu Q, Maragkakis M, Alexiou P, Ma J, et al. The RNA helicase
619 MOV10L1 binds piRNA precursors to initiate piRNA processing. Genes Dev. 2015;29:617–
620 29. doi:10.1101/gad.254631.114.

621 81. Patil VS, Kai T. Repression of Retroelements in Drosophila Germline via piRNA Pathway
622 by the Tudor Domain Protein Tejas. Curr Biol. 2010;20:724–30.
623 doi:10.1016/j.cub.2010.02.046.

624 82. Ohtani H, Iwasaki YW, Shibuya A, Siomi H, Siomi MC, Saito K. DmGTSF1 is necessary
625 for Piwi-piRISC-mediated transcriptional transposon silencing in the Drosophila ovary.
626 Genes Dev. 2013;27:1656–61. doi:10.1101/gad.221515.113.

627 83. Dönertas D, Sienski G, Brennecke J. Drosophila Gtsf1 is an essential component of the
628 Piwi-mediated transcriptional silencing complex. Genes Dev. 2013;27:1693–705.
629 doi:10.1101/gad.221150.113.

630 84. Sato K, Siomi MC. Functional and structural insights into the piRNA factor Maelstrom.
631 FEBS Lett. 2015;589:1688–93. doi:10.1016/j.febslet.2015.03.023.

632 85. Li WH. Unbiased estimation of the rates of synonymous and nonsynonymous
633 substitution. Journal of Molecular Evolution. 1993;36:96–9. doi:10.1007/BF02407308.

634

Figures

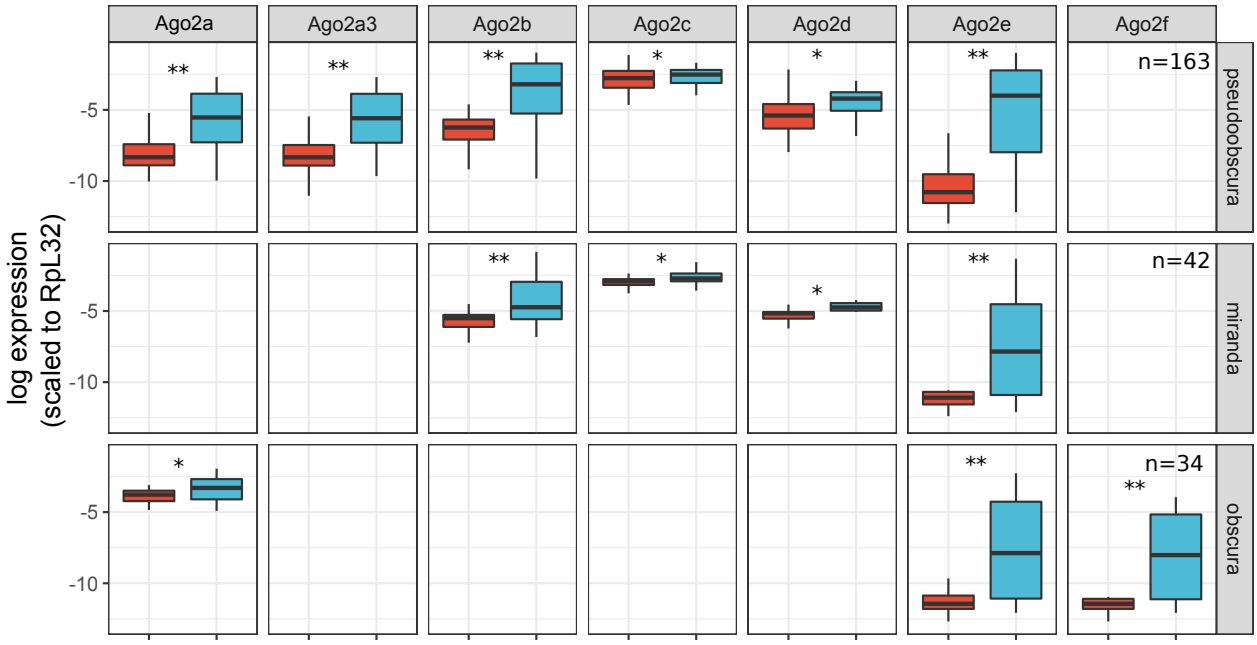


Figure 1 Expression Profile of Argonaute 2

Plots show the difference in expression between female (red) and male (blue) flies, based on public RNAseq data, normalized to rpL32 and plotted on a natural log scale. The significance of differences between the sexes was assessed using a linear model fitted with MCMCglmm and is denoted by asterisks: * 0.001 < pMCMC < 0.05; ** pMCMC <= 0.001). Sample size (n) represents the number of RNAseq datasets used (combined across tissues). *Ago2d* is the ancestral copy in Dpse and Dmir, *Ago2a* is the ancestral copy in *D. obscura* and *Ago2a* is recently duplicated in Dpse become *Ago2a1* (*Ago2a*) and *Ago2a3*.

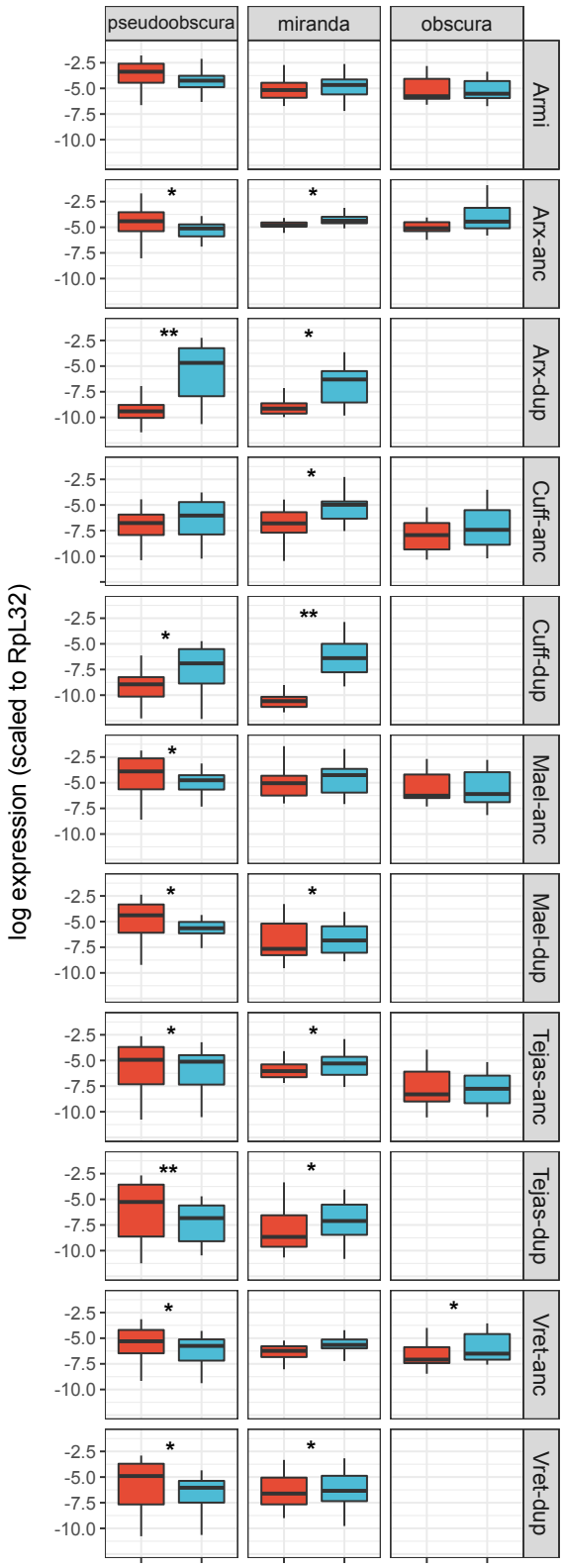


Figure 2 The expression profile of RNAi-accessory protein genes

Plots show the expression pattern between sex (male: blue, female:red) for genes other than *Ago2*; 'anc' ancestral copy, 'dup' duplicate copy as inferred by synteny. The y-axis is the natural log of normalized expression. The significance between sexes is denoted by * (0.001 < pMCMC < 0.05) and ** (pMCMC < 0.001). Sample sizes are the same as **Figure 1**.

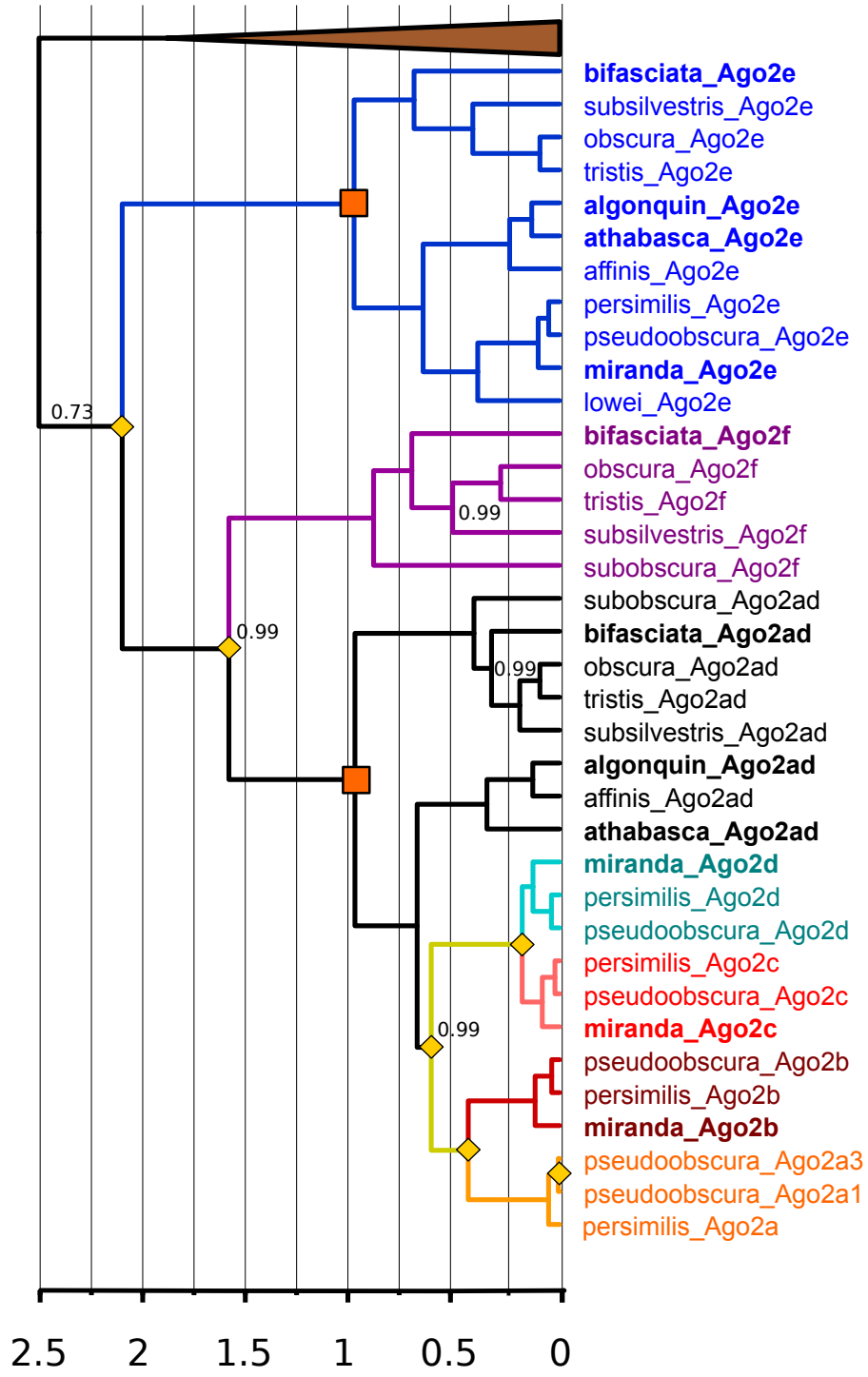
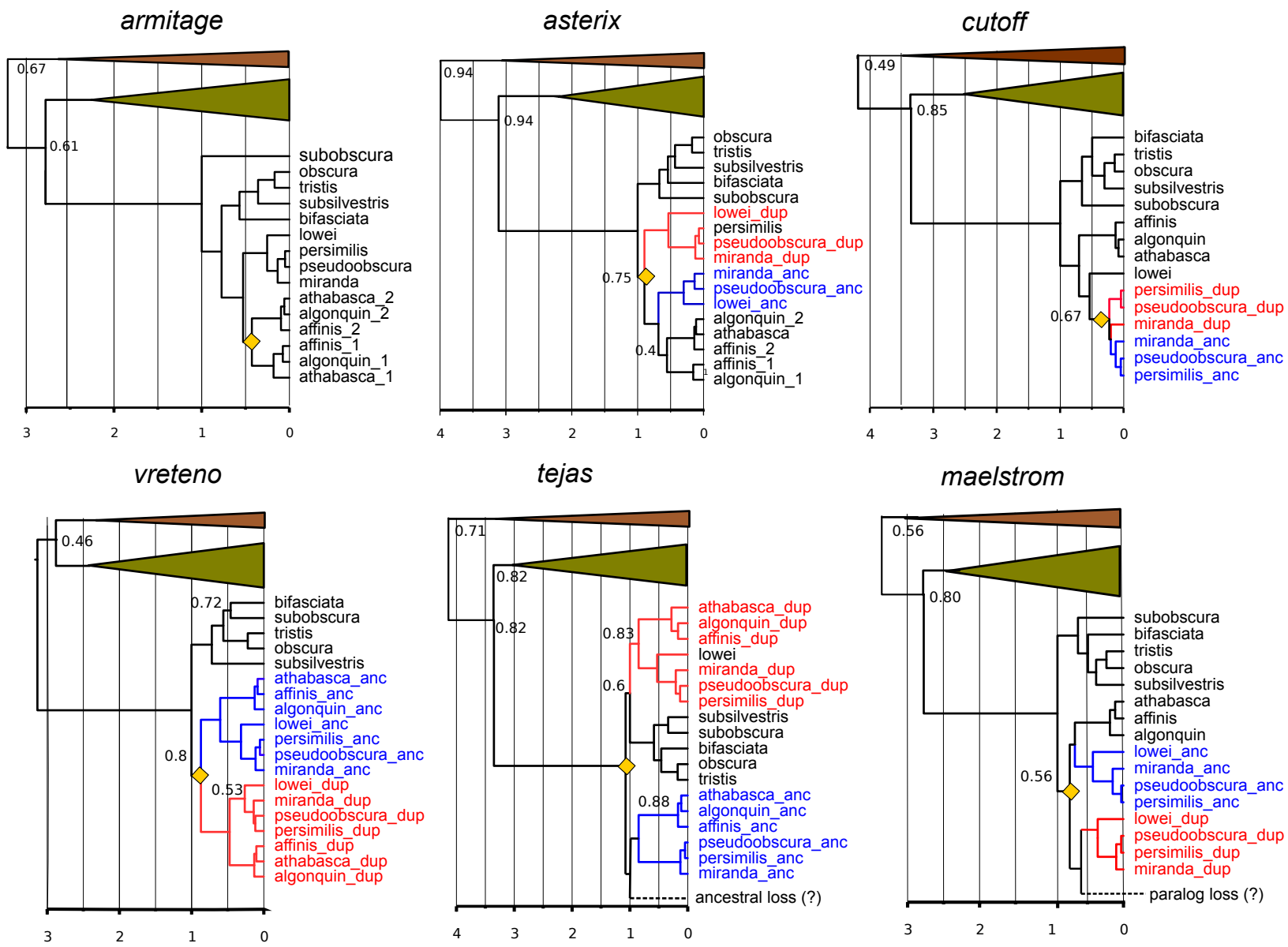


Figure 3 Bayesian Relaxed Clock Gene Tree for Argonaute2

The duplication events are marked by yellow diamonds, species other than the obscura group are collapsed (brown triangle), and paralog clades are colored. Bayesian posterior supports are only shown for the nodes with support less than 1. Genes not previously included in the analysis of [4] are marked in bold. Time is expressed relative to the split between the obscura and subobscura subgroups (orange boxes), which was constrained to be 1 using a strongly informative prior.

Figure 4 Bayesian Relaxed Clock Trees for 6 RNAi accessory protein genes
Ancestral genes are marked by bold blue, duplicates in bold red. Yellow diamonds indicate duplication events. Species other than *obscura* group are collapsed (green triangle; melanogaster group and brown triangle: other *Drosophila* species). Posterior Bayesian Supports are only shown in the nodes with support less than 1. Duplicated genes which could not be assigned as ancestral or duplicate is marked by _1 or _2. Scale axis is in the time relative to the *obscura* speciation, which was set to 1.


armitage
0.67
0.61
subobscura
obscura
tristis
subsilvestris
bifasciata
lowei
persimilis
pseudoobscura
miranda
athabasca_2
algonquin_2
affinis_2
affinis_1
algonquin_1
athabasca_1
3 2 1 0
asterix
0.94
0.94
0.75
0.4
obscura
tristis
subsilvestris
bifasciata
subobscura
lowei_dup
persimilis
pseudoobscura_dup
miranda_dup
miranda_anc
pseudoobscura_anc
lowei_anc
algonquin_2
athabasca
affinis_2
affinis_1
algonquin_1
4 3 2 1 0
cutoff
0.49
0.85
0.67
bifasciata
tristis
obscura
subsilvestris
subobscura
affinis
algonquin
athabasca
lowei
persimilis_dup
pseudoobscura_dup
miranda_dup
miranda_anc
pseudoobscura_anc
persimilis_anc
4 3 2 1 0
vreteno
0.46
0.72
0.8
0.53
bifasciata
subobscura
tristis
obscura
subsilvestris
athabasca_anc
affinis_anc
algonquin_anc
lowei_anc
persimilis_anc
pseudoobscura_anc
miranda_anc
lowei_dup
miranda_dup
pseudoobscura_dup
persimilis_dup
affinis_dup
athabasca_dup
algonquin_dup
3 2 1 0
tejas
0.71
0.82
0.82
0.83
0.6
0.88
athabasca_dup
algonquin_dup
affinis_dup
lowei
miranda_dup
pseudoobscura_dup
persimilis_dup
subsilvestris
subobscura
bifasciata
obscura
tristis
athabasca_anc
algonquin_anc
affinis_anc
pseudoobscura_anc
persimilis_anc
miranda_anc
ancestral loss (?)
4 3 2 1 0
maelstrom
0.56
0.80
0.56
subobscura
bifasciata
tristis
obscura
subsilvestris
athabasca
affinis
algonquin
lowei_anc
miranda_anc
pseudoobscura_anc
persimilis_anc
lowei_dup
pseudoobscura_dup
persimilis_dup
miranda_dup
paralog loss (?)
3 2 1 0

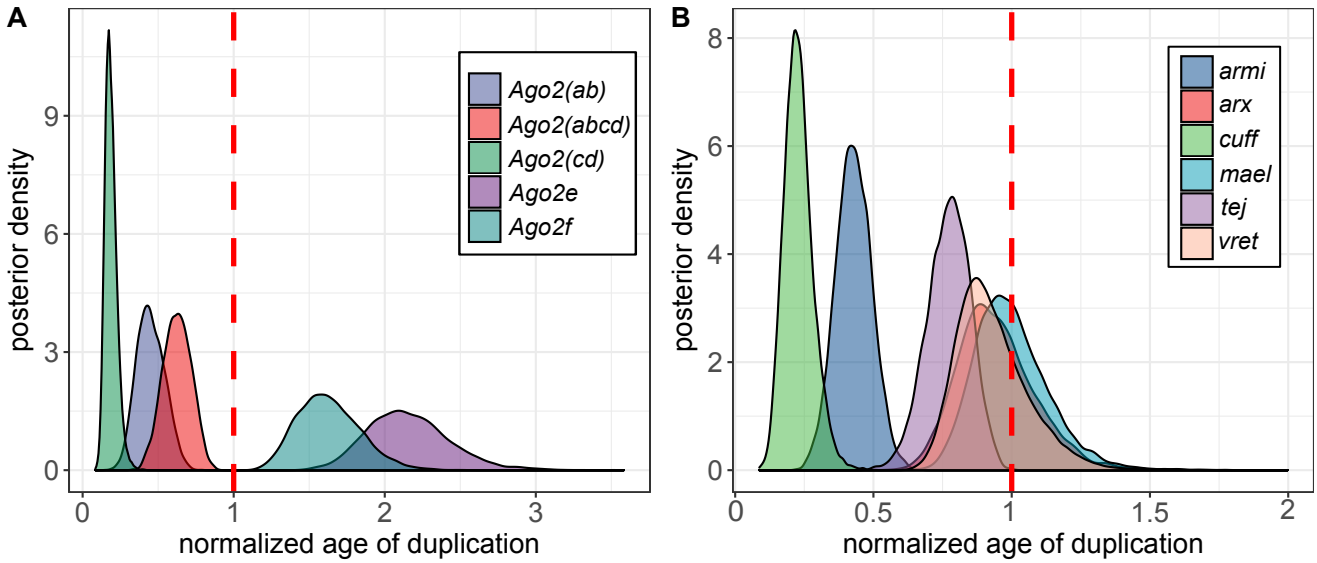


Figure 5 Density plot for posterior distributions of the duplication age

The MCMC posterior of the age of duplication node after 25% burn-in. (**A**) *Argounate 2* and (**B**) RNAi accessory proteins. The broken red line denotes speciation event in *obscura* group which was normalised to be 1.

Tables

Table 1 The details of RNAi accessory genes duplicated in the obscura group as reported by Palmer *et al.* [22]
*The tissue gene expression in *Dmel* is based on the FlyAtlas2 [76]. We report tissues with enrichment > 0.4.

Gene	Involvement in piRNA pathway	Function of the protein product	Tissue expression in *Dmel*	Reference
cutoff (*cuff*)	piRNA transcription	Forms a complex with Rhino-Deadlock-Cutoff (Rhi-Del-Cuff) to protect uncapped non-canonical (dual-strand cluster) piRNA transcript from degradation, splicing and transcription termination	testis	[77]
vreteno (*vret*)	piRNA biogenesis	A Tudor-domain protein which essential for an early primary piRNA processing	larval brain, adult female salivary gland, ovary, testis	[78]
armitage (*armi*)	piRNA biogenesis	A RNA helicase which unwinds the piRNA intermediates before loading into Piwi	ubiquitously expressed	[79] [80]
tejas *(tej)*	Secondary piRNA production (Ping-pong cycle)	A Tudor-domain protein which physically interact with *Vas*, *Spn-E* and *Aub* for a proper ping-pong cycle in the *nuage*	testis, accessory glands, adult female salivary gland, ovary	[81]
asterix (*arx*)	TGS (Transcriptional Gene Silencing)	A zinc-finger protein which directly interacts with Piwi to scan and identify the transposon transcriptions as target for histone modifications	ubiquitously expressed	[82] [83]
maelstrom (*mael*)	TGS	Act downstream of Piwi to establish histone modification and prevent the spreading of the silencing marker to the surrounding genes	brain, testis, adult female salivary gland, ovary	[84]

Table 2 Joint estimates of adaptive evolution across genes

Maximum-likelihood extension of MK test model fitted with different constraints on α [29]. LnL is the log likelihood of the model, AIC is the Akaike Information Criterion with corresponding relative probability as Akaike Weight (w_i). The most supported model is in bold.

Model	Model description	LnL	AICc	Akaike weight(w_i)	Maximum likelihood α estimate	
					ancestral	duplicate
M0	α-anc=0, α-dup=0	-300.61	633.2207	4.00×10^{-12}	0	0
M1	α-anc > 0, α-dup>0 α-anc = α-dup	-284.105	602.2092	2.6×10^{-05}	0.539	0.539
M2	α-anc>0, α-dup=0	-301.857	637.7133	5.10×10^{-13}	0.047	0
M3	α-anc=0, α-dup>0	-275.249	584.4975	0.186	0	0.618
M4	**α-anc > 0, α-dup>0 α-anc ≠ α-dup**	**-272.774**	**581.5471**	**0.813**	**0.2**	**0.676**

Additional files

Additional file 1 Table S1

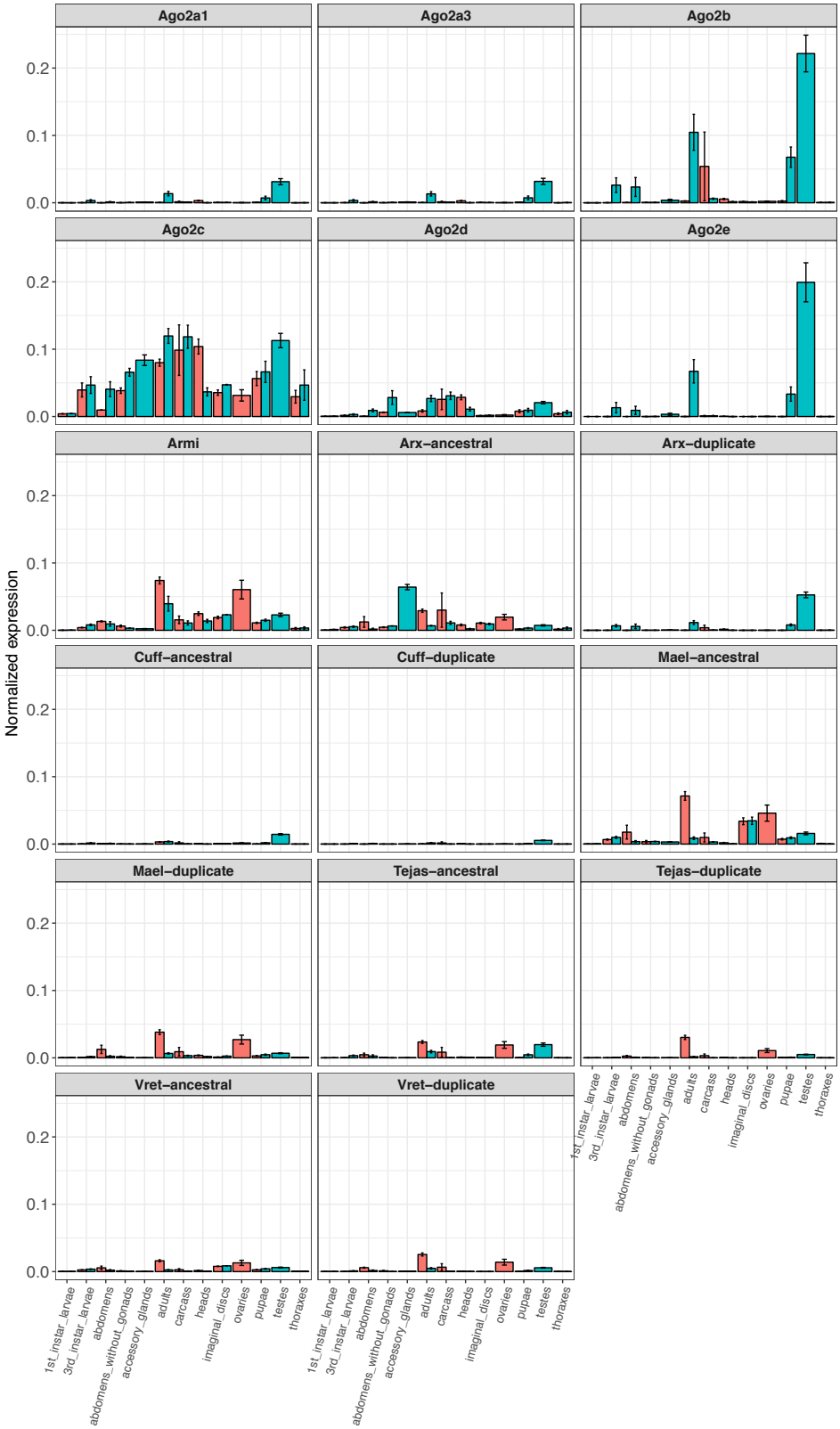
File format: xlsx
Title: **The detailed RNAi genes and its duplicate in** *D. pseudoobscura*
Description: The genomic position is based on the D. pseudoobscura assembly 3.0

Gene	Flybase ID	Locus Tag	Chromosomal location	Muller Element	Start Position	Gene length	Duplication mechanism
Armi	FBgn0246685	GA25304	4_group1	D	981803	3507	
Arx-ancestral	FBgn0077765	GA17756	XR_group8	D	6958861	495	
Arx-duplicate	FBgn0247462	GA26086	4_group3	B	10387204	501	Retrotransposition
Cuff-ancestral	FBgn0246456	GA25073	3	C	17023607	1113	
Cuff-duplicate	FBgn0244163	GA22760	XL_group1a	A	3011154	1119	Direct DNA duplication
Tejas-ancestral	FBgn0081173	GA21185	3	C	17440474	1782	
Tejas-duplicate	FBgn0248235	GA26863	2	E	1079477	1413	Retrotransposition
Vret-ancestral	FBgn0078422	GA18420	2	E	6266048	2085	
Vret-duplicate	FBgn0244928	GA23527	XL_group1e	A	10126307	2110	
Mael-ancestral	FBgn0248264	GA26892	2	E	54104	1314	
Mael-duplicate	FBgn0249827	GA28467	XR_group8	D	396866	1113	Direct DNA duplication
Ago2a	FBgn0249477	GA28114	Unknown_group_265		16703	2008	Direct DNA duplication
Ago2b	FBgn0248821	GA27454	2	E	23856171	2940	Direct DNA duplication
Ago2c	FBgn0248778	GA27411	2	E	21862037	2915	Direct DNA duplication
Ago2d(ancestral)	FBgn0245029	GA23629	XR_group6	D	200896	2839	
Ago2e	FBgn0247385	GA23629	4_group3	B	6716295	2285	Direct DNA duplication

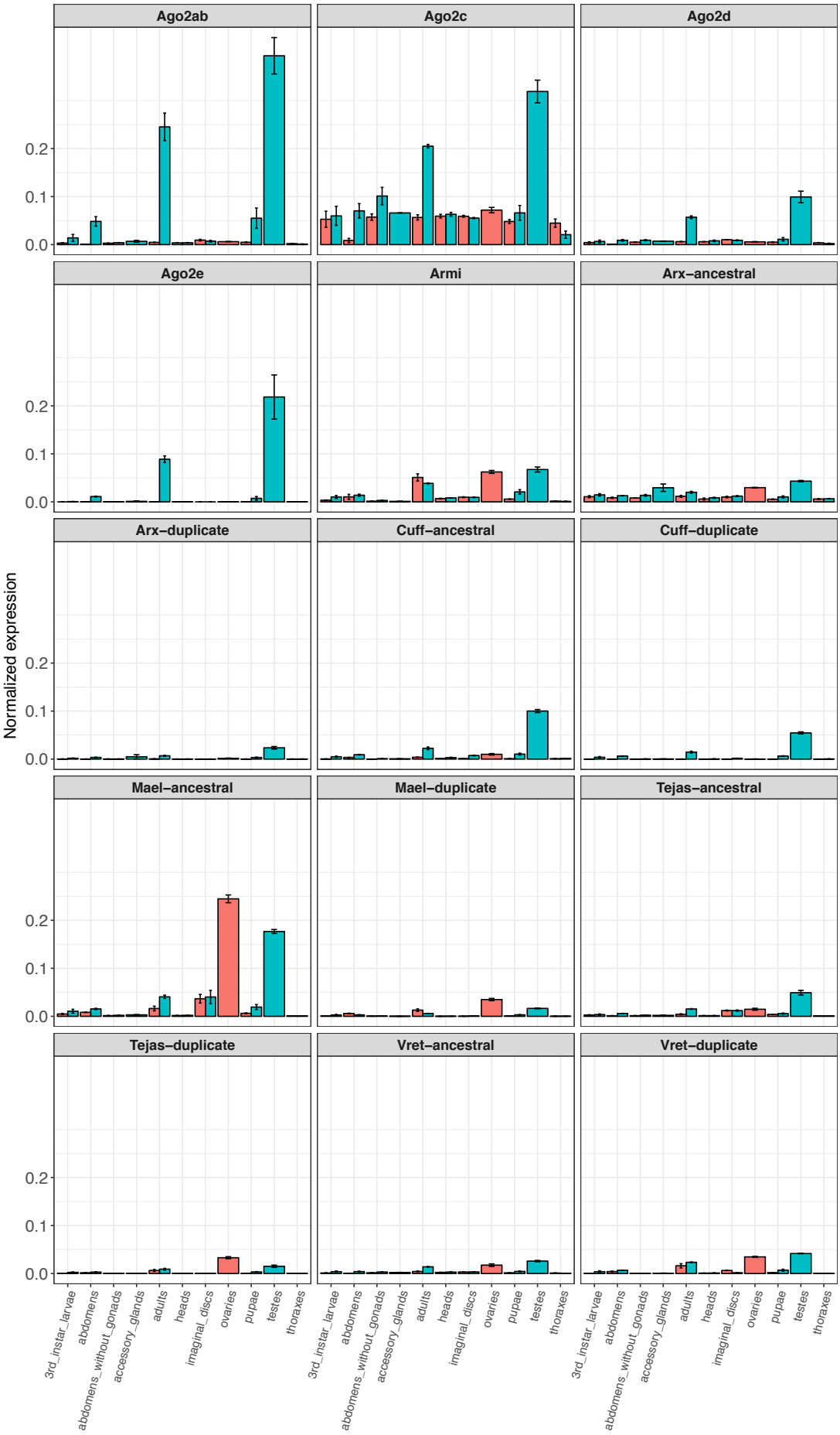
Additional file 2 Figure S1

File format: pdf
Title: **The expression profile of RNAi across tissue**
Description: The normalized expression plotted across tissues. The error bars denote the standard error for the given tissue. Blue bar indicates male and female is indicated by red bar. The plot is shown for *D. pseudoobscura* (n=163), *D. miranda* (n=42) and *D. obscura* (n=34), respectively.

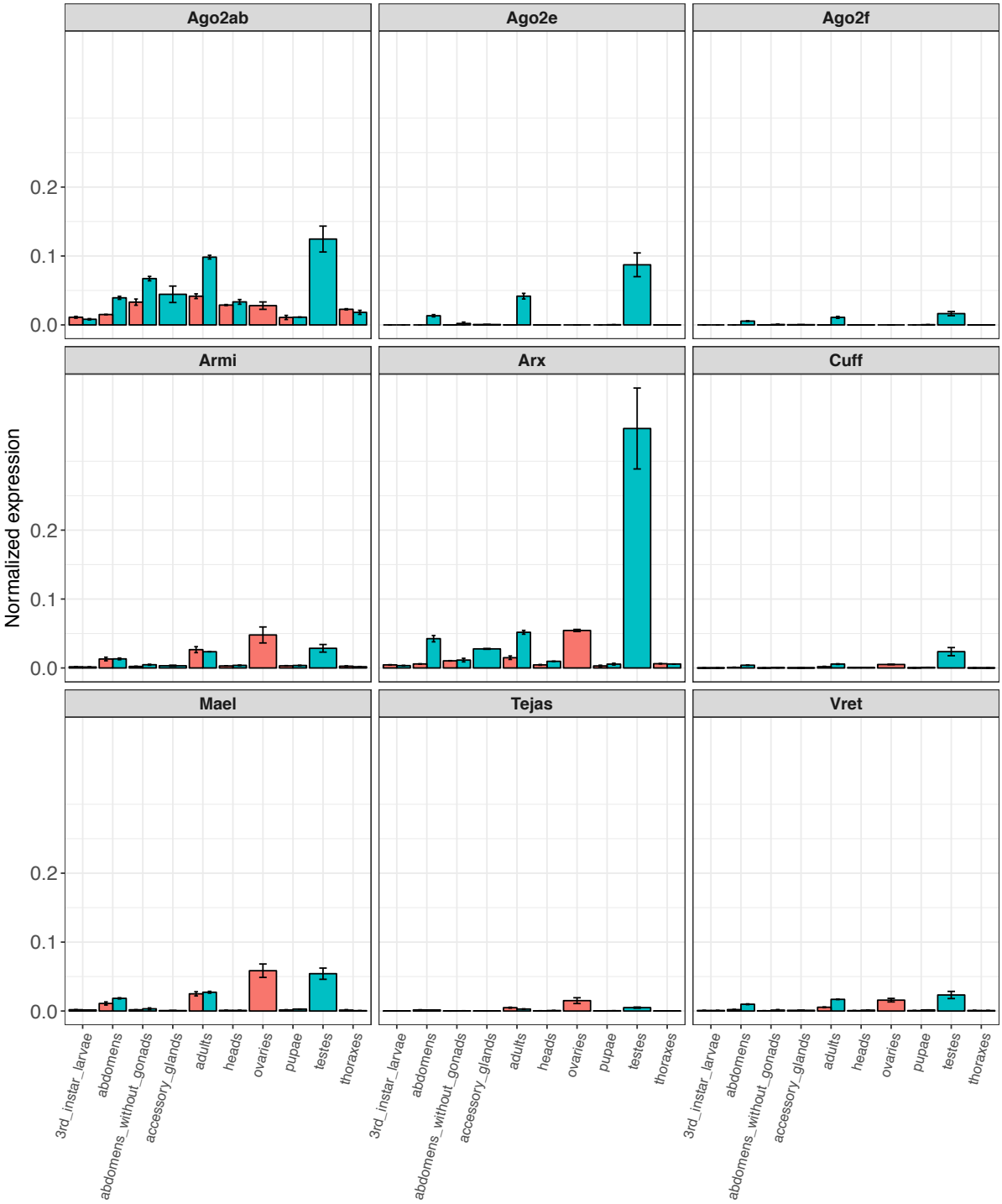
D. pseudoobscura



D. miranda


Ago2ab
Ago2c
Ago2d
Ago2e
Armi
Arx−ancestral
Arx−duplicate
Cuff−ancestral
Cuff−duplicate
Mael−ancestral
Mael−duplicate
Tejas−ancestral
Tejas−duplicate
Vret−ancestral
Vret−duplicate
Normalized expression
0.0
0.1
0.2
3rd_instar_larvae
abdomens
abdomens_without_gonads
accessory_glands
adults
heads
imaginal_discs
ovaries
pupae
testes
thoraxes

D. obscura


Ago2ab
Ago2e
Ago2f
Armi
Arx
Cuff
Mael
Tejas
Vret
Normalized expression
0.0
0.1
0.2
3rd_instar_larvae
abdomens
abdomens_without_gonads
accessory_glands
adults
heads
ovaries
pupae
testes
thoraxes

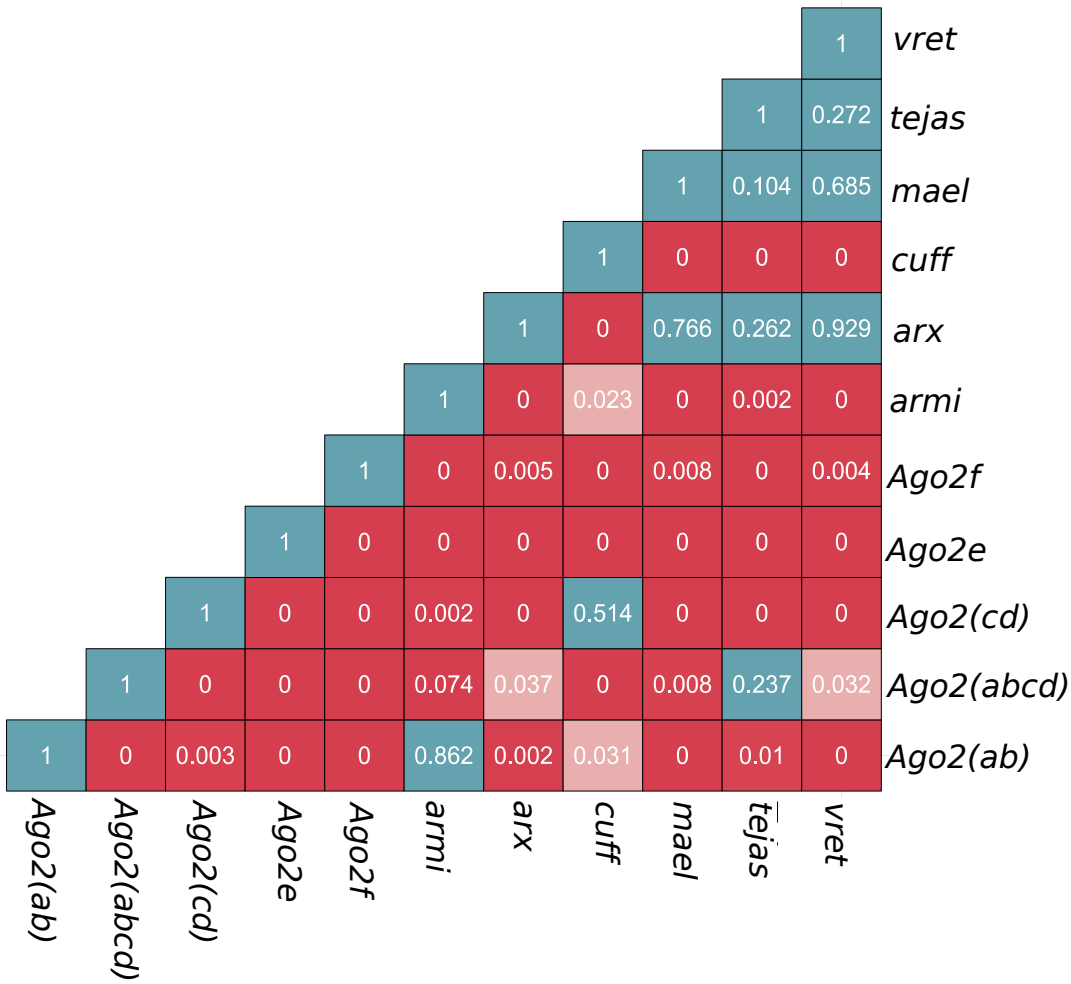
Additional file 3 Table S2

File format: xlsx
Title: **MK test of duplicated RNAi genes in** *D. pseudoobscura-D. miranda*
Description: Ds is synonymous divergence, Dn is non-synonomous divergence, Pn is the non-synonomous polymorphisms, Ps is the synonymous polymorphisms, α represents proportion of substitutions that are adaptive, a is the absolute number of adaptive substitution. Ln is the number of non-synonymous sites, Ls is the number of synonymous sites. Ka is the number of non-synonymous mutations per non-synonymous sites and Ks the number of synonymous mutations per synonymous sites from single randomly chosen strain for each species (Li, 1993 [85] calculated using R package seqinr). Parameter ω_a is identical to Ka/Ks ratio except that the numerator only takes adaptive divergence (α * Dn)/Ln)/(Ds/Ls).

Gene	Ds	Dn	Ps	Pn	alpha(α)	Fisher p-value	A	Ln	Ls	Ka	Ks	Ka/Ks	ω_a
Armi	34	72	22	30	0.356	0.220	26	2639.83	810.17	0.0380	0.0538	0.707	0.252
Arx-ancestral	2	9	7	4	**0.905**	**0.017**	8	383.67	102.33	0.0246	0.0232	**1.059**	**0.958**
Arx-duplicate	3	5	10	10	0.400	0.680	2	386.67	105.33	0.0204	0.0400	0.509	0.204
Cuff-ancestral	5	29	10	10	**0.828**	**0.010**	24	863.67	255.33	0.0352	0.0275	**1.281**	**1.060**
Cuff-duplicate	6	39	14	17	**0.813**	**0.003**	32	856.83	250.17	0.0297	0.0417	**0.714**	**0.580**
Tejas-ancestral	11	24	10	23	-0.054	1.000	0	1082.17	312.83	0.0312	0.0293	1.067	0.000
Tejas-duplicate	16	68	9	13	**0.660**	**0.046**	45	1264.83	379.17	0.0222	0.0352	**0.631**	**0.416**
Mael-ancestral	8	27	9	22	0.276	0.586	7	1001.67	303.33	0.0312	0.0293	1.067	0.294
Mael-duplicate	4	28	11	13	**0.831**	**0.007**	23	865.67	247.33	0.0323	0.0162	**2.000**	**1.662**
Vret-ancestral	17	34	37	70	0.054	1.000	2	1610.33	471.67	0.0211	0.0360	0.586	0.032
Vret-duplicate	25	65	23	26	**0.565**	**0.027**	37	1582.83	466.17	0.0517	0.0617	**0.838**	**0.473**
Ago2b	41	52	3	12	**-2.154**	**0.097**	0	882.00	264.00	0.0590	0.1553	**0.380**	**0.000**
Ago2c	2	34	2	22	0.353	1.000	12	1571.83	483.17	0.0216	0.0041	5.226	1.845
Ago2d	27	44	5	16	-0.964	0.300	0	1228.83	367.17	0.0358	0.0735	0.487	0.000
Ago2e	3	18	54	16	**0.951**	**0.000**	17	1410.17	464.83	0.0128	0.0065	**1.978**	**1.881**

Additional file 5 Figure S2

File format: pdf
Title: **The heat-map p-value of the difference between pairwise posterior distributions**
Description: Large p-value (>0.05) indicates the overlapping distribution and the duplication time might be shared (blue box). Red boxes denote comparison with p-value < 0.05 which indicate non-overlapping posterior distribution and an asynchronous duplication events . Pink colored box indicates the marginally significant (0.01 < p-value <0.05).



Additional file 6 Table S3

File format: multiple-tab xlsx
Title: **Additional MK test analysis**
Description: Table the results of additional MK test analysis where minor allele frequency is removed, repetition with larger dataset and results in *D. melanogaster*

Table 1 MK test result in *D. pseudoobscura* with MAF < 12.5% removed

Gene	Ds	Dn	Ps	Pn	α	Fisher p-value
Armi	38	74	7	11	0.193	0.79
Arx-ancestral	**2**	**9**	**3**	**0**	**1**	**0.0274**
Arx-duplicate	3	5	1	1	0.4	1
Cuff-ancestral	5	29	2	8	0.31	1
Cuff-duplicate	**6**	**41**	**5**	**4**	**0.883**	**0.0099**
Tejas-ancestral	11	24	5	11	-0.008	1
Tejas-duplicate	17	69	1	2	0.507	0.496
Mael-ancestral	8	28	4	4	0.714	0.1846
Mael-duplicate	**4**	**31**	**2**	**1**	**0.935**	**0.0592**
Vret-ancestral	18	35	16	27	0.132	0.831
Vret-duplicate	**25**	**68**	**8**	**7**	**0.665**	**0.069**
Ago2b	41	53	2	6	-1.321	0.4621
Ago2c	2	34	0	5	0	0
Ago2d	27	44	2	8	-1.455	0.318
Ago2e	**4**	**19**	**13**	**2**	**0.968**	**0.000037**

Table 2 MK test result for larger pseudoobscura dataset Fuller *et. al* [31]

Gene	Ds	Dn	Ps	Pn	NI	α	Fisher p-value
Armi	**33**	**71**	**36**	**38**	**0.491**	**0.509**	**0.02259**
Arx-ancestral	**2**	**7**	**14**	**4**	**0.082**	**0.918**	**0.011**
Arx-duplicate	3	5	14	11	0.471	0.529	0.438
Cuff-ancestral	**4**	**29**	**18**	**16**	**0.123**	**0.877**	**0.000569**
Cuff-duplicate	**6**	**39**	**15**	**18**	**0.185**	**0.815**	**0.002**
Tejas-ancestral	12	23	11	28	1.328	-0.328	0.62
Tejas-duplicate	**14**	**67**	**16**	**23**	**0.3**	**0.7**	**0.0069**
Mael-ancestral	8	27	7	22	0.724	0.276	0.568
Mael-duplicate	**4**	**28**	**11**	**15**	**0.195**	**0.805**	**0.015**
Vret-ancestral	17	33	44	85	0.995	0.005	1
Vret-duplicate	**25**	**65**	**25**	**27**	**0.415**	**0.585**	**0.018**

Table 3 MK test from DGRP Freeze 1 Dataset (*D. melanogaster* - *D. simulans*)

Gene	Ds	Dn	Ps	Pn	NI	α	Fisher p-value
Arx	**9**	**3**	**3**	**3**	**0**	**0.333**	**0.04**
Ago2	53	100	1	4	2.12	-1.12	0.66
Tejas	35	87	4	20	2.011	-1.011	0.314
Mael	22	36	4	5	0.764	0.236	0.72
Armi	**54**	**64**	**17**	**6**	**0.298**	**0.702**	**0.02**
Vret	45	30	13	17	1.962	-0.962	0.134
Cuff	33	62	5	5	0.532	0.468	0.49