Promoters of germline transposon silencing genes evolve rapidly accompanied by diverging gene expression

DAVID W. J. MCQUARRIE[1,2], AZAD ALIZADA[3], BENJAMIN CZECH NICHOLSON[3] AND MATTHIAS SOLLER[1,2*]

[1]School of Biosciences, College of Life and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[2]Birmingham Centre for Genome Biology, University of Birmingham, Edgbaston, Birmingham, B15 2TT, United Kingdom

[3]Cancer Research UK Cambridge Institute, University of Cambridge, Li Ka Shing Centre, Cambridge CB2 0RE, United Kingdom

[*]Corresponding author: m.soller@bham.ac.uk

Running title: Evolution of promoters of germline transposon silencing factors

Key Words: Piwi-interacting RNA (piRNA), germline transposon silencing, RNA transgenerational inheritance, rapid evolution, promoter evolution, speciation, nuclear pore complex, neuronal wiring, transposon silencing

Abstract

Background: The piRNA pathway in animal gonads functions as an 'RNA-based immune system', serving to silence transposable elements and prevent inheritance of novel invaders. In *Drosophila*, this pathway relies on three gonad-specific Argonaute proteins (Argonaute-3, Aubergine and Piwi) that associate with 23-28 nucleotide piRNAs, directing the silencing of transposon-derived transcripts. Transposons constitute a primary driver of genome evolution, yet the evolution of piRNA pathway factors has not received in-depth exploration. Specifically, channel nuclear pore proteins, which impact piRNA processing, exhibit regions of rapid evolution in their promoters. Consequently, the question arises whether such a mode of evolution is a general feature of transposon silencing pathways.

Results: By employing phylogenomic analysis of coding and promoter regions within genes that function in transposon silencing in *Drosophila*, we demonstrate that the promoters of germ cell-specific piRNA factors are undergoing rapid evolution. Our findings indicate that rapid promoter evolution is a common trait among piRNA factors engaged in germline silencing across insect species, potentially contributing to gene expression divergence in closely related taxa. Furthermore, we observe that the promoters of genes exclusively expressed in germ cells generally exhibit rapid evolution, which correlates with divergent gene expression.

Conclusions: Our results suggest that promoter evolution could have wide impact on transposon silencing and evolution of species through differential expression of genes driven by invading transposons.

46 **Background**

47 Eukaryotic organisms are continuously challenged from genomic parasites called
48 transposable elements (TEs) (1-4). Unchecked transposon activity often results in reduced
49 reproductive fitness (5-8). To negate the detrimental effects posed by TE mobilisation, vital
50 regulatory pathways evolved that efficiently suppress transposon activity. At the centre of these
51 are 23-28 nucleotide small RNAs, called PIWI-interacting RNAs (piRNAs). piRNAs are
52 generated from transposons and from genomic clusters and bind to Argonaute proteins of the
53 PIWI-clade, including P-element induced wimpy testis (Piwi). Aubergine (Aub) and
54 Argonaute-3 (Ago3). The piRNA pathway is often referred to as an 'RNA-based immune
55 system' thanks to its ability to adopt to new TE insertions and the memory of past transposon
56 activity through piRNA clusters through RNA transgenerational inheritance (1, 2, 9, 10).

57 The piRNA pathway is arguably best understood in the ovary of *Drosophila*
58 *melanogaster*. It was shown that ovaries feature two branches of the silencing machinery (5),
59 one that is active in germ cells and a condensed version that is specific to somatic follicle cells
60 that surround the germline (2, 11). Somatic follicle cells only feature unistrand piRNA clusters,
61 such as *flamenco*, and exclusively produce piRNAs via a Zucchini (Zuc)-dependent biogenesis
62 mechanism that takes place on the mitochondria surface involving Armitage (Armi), Papi,
63 Vreteno (Vret), Gasz, Daedalus (Daed), Minotaur (Mino), Shutdown (Shu), and Sister of Yb
64 (SoYb) (Fig. 1A) (3, 12). Licensing of *flam* for piRNA production takes place upstream, at so-
65 called Yb-bodies (1, 12). Somatic cells only express Piwi, which following the association with
66 a mature piRNA shuttles to the nucleus. There, Piwi-piRNA complexes scan for nascent TE
67 transcripts and instruct co-transcriptional gene silencing (cTGS) that requires components of
68 the general chromatin silencing machinery and the piRNA pathway-specific factors Asterix
69 (Arx), Piwi, Panoramix (Panx), Nuclear export factor 2 (Nxf2) and Maelstrom (Mael) (Fig.
70 1A) (1, 2, 12, 13).

71 In germ cells, piRNAs predominantly originate from dual-strand clusters and are
72 produced by a different biogenesis mechanism called the ping-pong amplification cycle (Fig.
73 1B) (1-3, 5, 11, 12, 14). Dual-strand clusters harbour remnants of transposons and are expressed
74 via a non-canonical transcription process that is centred on a complex containing the HP1a
75 homolog Rhino (Rhi), Deadlock (Del) and Cutoff (Cuff), and further requires Moonshiner
76 (Moon) (Fig. 1B). This Rhi-Del-Cuff complex also recruits a non-canonical export machinery
77 consisting of Bootlegger (Boot) and Nuclear export factor 3 (Nxf3), which transport piRNA
78 precursors for processing to a germ-cell specific perinuclear region termed "nuage" (Fig. 1B).
79 Here, the ping-pong cycle takes place and relies on the two PIWI proteins Aubergine (Aub)
80 and Argonaute 3 (Ago3), as well as several RNA binding proteins and Tudor-domain
81 containing factors including Tejas (Tej), Vasa (Vas), Qin, Squash (Squ), Tapas, Brother of Yb
82 (BoYb), Krimper (Krimp), and Spindle-E (Spn-E) (Fig. 1B) (1-3, 12, 14, 15).
83 Channel nuclear pore proteins (Nups) have been implicated in piRNA regulated
84 transposon silencing in *C. elegans* (16). Interestingly, in *Drosophila*, Nups were shown to
85 affect transposon silencing in both the germline (15, 17), and in somatic cells of the ovary, here
86 specifically by contributing to the conversion of precursor RNAs from the *flam* locus into
87 piRNAs (18). Recently, we found that inner and outer ring Nups in the nuclear pore complex
88 (NPC) evolve rapidly through indel accumulation in their promoters (19, 20). Such promoter
89 indel variability in Nup54 has dominant, pleiotropic effects on sexual differentiation including
90 neuronal wiring important for female post-mating behaviours directed by male-derived sex
91 peptide, that as a result of sexual conflict could drive speciation by such mechanism (16, 19,
92 21). Intriguingly, nuclear import/export pathways have been linked to speciation through
93 hybrid incompatibility, but are in addition important for many cellular processes beyond
94 piRNA processing (22, 23). Hence, promoter evolution particularly of piRNA processing genes
95 could have wide impact on transposon silencing and evolution of species through differential

96 expression of genes, but whether regulators of germline transposon silencing generally evolve
97 rapidly through their promoters has not been determined.

98 Here, we analyse the promoters of genes involved in the ovarian piRNA pathway. We
99 compare piRNA factors required in somatic cells (Yb) or in both the somatic and germline
100 compartments (cTGS and piRNA biogenesis, Fig. 1A) with those that are germ cell-specific
101 (dual-strand clusters and the ping-pong cycle, Fig. 1B). We find that the promoters of germline-
102 specific piRNA factors are hot spots for rapid evolution compared to those of somatic piRNA
103 genes and the *Drosophila* genome average throughout 27 insect species. Analysis of genes with
104 significant accumulation of promoter changes reveals a mix of indel and base change
105 accumulation. Further, our analysis reveals that rapid promoter evolution is a general feature
106 of genes specifically expressed in germ cells, while soma-expressed genes evolve at a similar
107 rate to the average of all *Drosophila* genes. Through cross-species differential ovary RNA-seq
108 analysis we reveal that germ cell-specific genes also acquired more divergent expression levels
109 compared to genes expressed in somatic cells. Our findings highlight that promoters of
110 germline genes involved in transposon silencing evolve rapidly and are accompanied by
111 diverging gene expression, suggesting a possible mode of rapid speciation through
112 accumulation of changes in promoters.

113

114 **Results**

115 **Promoter regions, but not protein coding regions, of germline-specific piRNA factors are**
116 **hot spots for rapid evolution**

117 Since Nups have been shown to function in transposon silencing in the germline (12
118 from 14 tested are above the control) and display fast evolving promoters (15, 19, 20), we
119 analysed the rate of evolution in promoters and coding regions of piRNA pathway genes. We
120 separated all piRNA genes based on their genetic requirement into two categories: piRNA

121 factors essential in somatic cells, including the soma-specific piRNA factor Yb and piRNA
122 factors expressed in both somatic and germline cells (cTGS and piRNA biogenesis, Fig. 1A),
123 and germ cell-specific piRNA proteins (dual-strand clusters and ping-pong cycle, Fig 1B). The
124 piRNA factors essential in somatic cells include *armi*, *papi*, *vret*, *Gasz*, *daed*, *mino*, *zuc*, *shu*,
125 *SoYb*, *Yb*, *arx*, *piwi*, *panx*, *nxf2*, and *mael* (Fig. 1A). The group of piRNA factors that are
126 exclusively required in the germline are *aub*, *tej*, *Ago3*, *vas*, *qin*, *squ*, *tapas*, *BoYb*, *krimp*, *spn-*
127 *E, Boot*, *moon*, *nxf3*, *cuff*, *rhi*, and *del* (Fig. 1B) (3, 13, 24).

128 We defined the promoter containing region as a 350-nucleotide (-30 to -380) window
129 upstream of the predicted TATA box region 30 nucleotides upstream of the transcription start
130 site (TSS) (25). In a first approach, we performed a genome-wide promoter evolution analysis
131 of germline-specific as well as somatic piRNA factors and compared it to all genes in the
132 *Drosophila* genome (Fig 1C-F). We used publicly available PhyloP27way data composed of
133 27 insect species as a large-scale conservation score. We observed a significant accumulation
134 of changes in the germ cell-specific piRNA factor promoter regions compared to the genome
135 (Fig. 1C-F). Of note, the somatic piRNA factors evolved slower compared to the genome.

136 We next compared indel and base changes in the promoter regions between five closely
137 related *Drosophila* species (*D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D.*
138 *erecta*). Promoter scores (*d*) were calculated and compared to a control group comprised of the
139 m^6A methylation machinery genes (*Mettl3*, *Mettl14*, *fl(2)d*, *virilizer*, *flacc*, *nito* and *Hakai*,
140 *Ythdc1* and *Ythdf*), chosen due to their high evolutionary conservation and requirement for strict
141 stoichiometry for functionality, making this an ideal control group to monitor promoter
142 evolution (20, 26, 27). As previously reported, the m^6A machinery showed a slight increase in
143 sequence changes in promoter regions compared to coding regions (Fig. 1G-L) (20). In this
144 analysis, the group of somatic piRNA factors showed neutral evolution compared to the control
145 (Fig. 1G-L), while germ cell-specific piRNA factors as a group evolved at a much higher rate

146 (Fig. 1G and H), accruing a significantly increased number of promoter-located indels (Fig. 1I
147 and J) and base changes (Fig. 1K and L). At the individual gene level, 7 out of 14 of germ cell-
148 specific piRNA factors (*Ago3* was omitted due to its low conservation between closely related
149 *Drosophila* species (28-30) showed significant accumulation of promoter changes
150 (Supplementary Fig. 1B). Among the somatic piRNA factors, 3 out of 15 genes had significant
151 accumulation of promoter changes (Supplementary Fig. 1B).

152 To assess evolution in the coding region, we used McDonald-Kreitman tests (MKTs)
153 and analysed polymorphisms and divergence within *D. melanogaster* and between *D.*
154 *melanogaster* and *D. simulans* (31). None of the somatic piRNA factor genes were under
155 selection, while one of the germ cell-specific piRNA factors (*krimp*) was under selection
156 (Supplementary Fig. 1A).

157

158 **Promoters of genes required for germline transposon silencing are fast evolving**

159 Given the rapid evolution of germ cell-specific piRNA factors among closely related
160 *Drosophila* species, we wanted to determine (1) whether this is a general feature of insect
161 evolution, and (2) whether this is a general feature of genes required for transposon silencing
162 that have been identified in a genetic screen (GTS100 genes) (15). To get a scale of the
163 persistence of evolutionary changes in promoters of GTS100 genes we used publicly available
164 PhyloP27way data from UCSC genome browser (32, 33). Genes were ordered based on
165 transposon de-repression scores (15) and d^P was calculated for individual genes, the genome
166 average as a control group, and for the average of GTS100 gene promoter regions (Fig. 2A-E).
167 Genes without full PhyloP coverage (where ≥1 nucleotide in gene or upstream regions had no
168 homologous sequence in the 27 related species) were removed from the analysis.

169 The promoter score average of GTS100 genes showed a significant increase in
170 promoter nucleotide changes compared to the control (Fig. 2C). Although many of these genes

171 have additional roles, gene ontology (GO) analysis confirmed a primary role in piRNA binding
172 (Fig. 2D) (34-36). To understand whether GTS100 genes were individually fast evolving,
173 individual d^P scores were calculated (Fig. 2E). Compared to the genome average, 49 genes
174 (FEPG, Fast Evolving Promoter Germline genes) were significantly increased in promoter
175 change events, while 12 were significantly decreased (Fig. 2E). Of note *Brother of Yb* (*BoYb*)
176 which is considered the germline replacement of *Yb* (3), also displayed a fast evolving promoter
177 in contrast to its somatic counterpart *Yb* (Fig. 2E). To specifically look at indels at a high
178 confidence level, we analysed FEPG genes for rapid promoter evolution between the
179 previously used five *Drosophila* species (Fig. 2F-H). Here, indels and base changes were
180 measured, showing a significant increase in both indels and base changes compared to the
181 control (Fig. 2G and H).
182

183 **A subset of piRNA factors that are part of distinct protein complexes display hot spots**
184 **for rapid evolution in promoters**

185 In a saturating genetic screen for regulators of transposon silencing in germ cells (15),
186 numerous genes that are part of distinct protein complexes, including the nuclear cap-binding
187 complex (CBC), THO complex, the exon junction complex (EJC) inner core, outer shell and
188 EJC transient factors, the non-specific lethal (NSL) complex, SUMOylation modifiers and the
189 complex of proteins associated with Set1 (COMPASS) were identified (Supplementary Fig. 2).
190 Given the fast evolution of promoters in germline-specific piRNA genes, we analysed their
191 promoters for the rate of evolution. Compared to soma piRNA factors, promoter regions for
192 the nuclear CBC, SUMO modifiers, COMPASS complex, NSL complex and EJC outer shell
193 showed a significant accumulation of sequence changes (Fig. 2I and J). Analysis of individual
194 genes from both significant and non-significant gene groups revealed that 43% of analysed
195 genes have rapidly accumulated sequence changes in their promoters (Supplementary Fig. 2).

196 Specifically looking at genes in fast evolving complexes (Fig. 2I-J) revealed that 56% of genes
197 in these groups rapidly evolve through accumulation of mutations in their promoters
198 (Supplementary Fig. 2).

199 When analysing the base changes and indels of rapid evolving promoters (Fig. S3A-
200 D), indel accumulation was observed for the SUMO modifiers, COMPASS complex, NSL
201 complex and EJC outer shell (4 out of 5), while the nuclear CBC showed a significant decrease
202 (Fig. S3B). The nuclear CBC, SUMO modifiers and COMPASS complex also showed a
203 significant accumulation of base changes in their promoter regions (3 out of 5) (Fig. S3D).
204 Individually, 4 out of 27 genes were significantly enriched in indel events (*Hcf*, *dgt1*, *nsl1*, and
205 *Bin1*), while base change accumulation events indicated 8 out of 27 genes (*Ars2*, *cbp80*, *Ulp1*,
206 *Set1*, *Hcf*, *Cfp1*, *MBD-R2*) being significantly increased (Fig. S3E).

207 Next, we also analysed the rate of evolution in the coding region of these genes using
208 MKT analysis (31). This analysis shows that most of the coding regions were not under
209 selection (37 out of the 40 analysed genes), with the exception of cbc80 (a nuclear CBC
210 protein), Aly (a factor of the EJC outer shell) and Upf2 (a EJC transient factor), which were
211 under selection (Supplementary Fig. 4). In general, this indicates a neutral mode of evolution
212 in the CDS of these gene complexes implicated in transposon silencing.

213

214 **piRNA factors with rapidly evolving promoters show divergent gene expression in ovaries**
215 **of *D. yakuba***

216 Since promoters of FEPG genes are fast evolving, we analysed publicly available RNA-
217 seq data from ovaries between *D. melanogaster* and *D. yakuba* to assess gene expression
218 divergence in the FEPG gene group compared to the somatic piRNA factors (Fig. 3A) (37-39).
219 Here, many FEPG genes showed significant divergence in gene expression in both directions
220 (*D. melanogaster* or *D. yakuba*), but this was true for only few somatic piRNA factors genes

221 including *arx* (Fig 3A). Notably, we observed large expression divergence for *CG32152, MEP-*
222 *1* and *Rbbp5* (Fig. 3A). Aligning the promoters of these three genes revealed an increased
223 number of indels and base changes compared to the 5'UTR or the coding region (Fig. 3B-D),
224 but the cohort is too small to detect whether increased changes in promoters correlate with
225 altered expression between the two species.
226
227 **Germ cell-specific genes show rapid evolution of promoters associated with divergent**
228 **gene expression between species**
229 To investigate whether the rapid evolution of promoters in piRNA factors expressed in
230 germ cells is a common feature of germline-expressed genes, we used publicly available RNA-
231 seq data from FACS-sorted vasa-GFP ovaries to identify genes specifically expressed in either
232 germ cells or somatic cells of the ovary. First, using differential RNA-seq analysis of FACS-
233 sorted vasa-GFP+ (germline) and vasa-GFP- (somatic) cells from cultured *D.mel* ovary GSC
234 line (vasa-GFP;hs-bam;bamΔ86) (GSE119862) (40), we defined germline-enriched (vasa-
235 GFP+; log2FC>0, p-val<0.01) and soma-enriched (vasa-GFP-; log2FC<0, p-val<0.01) genes
236 (Fig. 4A and B). Next, using RNA-seq data from ovarian somatic cells (OSCs) (GSE160860)
237 (41), we further filtered the soma-enriched genes that are highly expressed in OSCs.
238 Additionally, we further filtered the germline-enriched genes that show expression in early fly
239 embryos (0-2hr, modENCODE). In this way, we generated a stringent list of 107 genes
240 specifically expressed in germ cells and of 59 soma-expressed genes (Supplementary Dataset
241 1).
242 We then analysed sequence change accumulation in promoters using PhyloP data and
243 calculated d^P for all *D. melanogaster* genes and genes expressed specifically in somatic cells
244 or the germline (Fig. 4C and D) or the average (Fig 4E). Germ cell-specific expressed genes
245 significantly accumulated changes (d^P) in their promoters compared to soma-expressed genes

246 (Fig. 4F). Comparison of the average promoter PhyloP scores and d^P conservation scores
247 between the genome average and somatic or germ cell-specifically expressed genes equally
248 revealed a significant increase in germ cell gene promoter nucleotide changes compared to the
249 genome average (Fig. 4G and H).

250 Since rapid promoter evolution of germline piRNA silencing factors correlated with
251 gene expression divergence in our previous analysis, we hypothesized that this was a general
252 feature of germ cell genes. To test this, we compared gene expression divergence between
253 somatic and germ cell-specific genes using publicly available RNA-seq data of *D.*
254 *melanogaster* and *D. yakuba* ovaries (Fig. 4I) (37-39). Notably, genes specifically expressed
255 in germ cells showed significantly higher divergence in expression compared to soma-
256 expressed genes (Fig. 4I). When d^P conservation scores were plotted against the gene
257 expression change between *D. melanogaster* and *D. yakuba*, the germ cell-specifically
258 expressed genes showed a higher divergence in expression especially in the higher d^P score
259 groups (>50-75, >75-100) compared to soma-expressed genes, which showed relatively
260 conserved expression between *D. melanogaster* and *D. yakuba* (Fig. 4J-L).

261

262 **Discussion**

263 Through analysis of polymorphisms within and between closely related *Drosophila*
264 species, we show that at the individual gene level, germline and somatic transposon silencing
265 genes and their regulators do not generally evolve through their coding regions. However, we
266 identify hot spots for rapid evolution in promoter regions (-30 to -380 nucleotides upstream of
267 the TSS) of germ cell-specific (dual-strand clusters and the ping-pong cycle), but not soma-
268 expressed piRNA genes (cTGS and piRNA biogenesis). Our analysis suggests that this mode
269 of evolution in the germline pathway could be a general feature of at least insect evolution.
270 Overall, our results point towards a key role for rapid evolution of gene promoters in the germ

271 cell-specific piRNA pathway. Intriguingly, the soma-specific Yb is replaced in germ cells with
272 BoYb and in contrast to its somatic cell counterpart, BoYb also displays a fast-evolving
273 promoter.

274 Transposon mobility has been attributed to the accumulation of sequence changes in
275 promoter regions because of the presence of open chromatin around the transcription start site
276 of genes (42). Interestingly, in core NPC genes, rapid evolution mostly led to accumulation of
277 indels (19, 20), while piRNA genes display equal accumulation of indels and base changes.
278 Likely, indels have more profound effects on changes in transcription than single nucleotide
279 alterations. This feature might reflect that compromised transposon silencing causes sterility,
280 hence not allowing for substantial changes in expression in piRNA processing genes. Likewise,
281 stoichiometric changes in protein complexes can drastically alter complex functionality, e.g.
282 the male-specific lethal (MSL) complex binds less to its targets when one component of the
283 complex is missing (43, 44). Hence, changes in the protein levels of individual piRNA factors
284 likely have dominant effects on their capacity to silence transposon activity.

285 The occurrence of novel transposons has fundamental impact on a species reproductive
286 success through a phenomenon call hybrid dysgenesis (45). Here, if a novel transposable
287 element is transmitted by the male, female fertility is severely compromised and can result in
288 sterility as a result of missing piRNA silencing of the novel invader. Essential to combat these
289 novel active transposable elements is the ping-pong cycle, primed by transcripts from the novel
290 transposon. Accordingly, females can be become resistant to such novel transposons over time
291 through RNA transgenerational inheritance of piRNAs to build up ping-pong cycle
292 amplification in germ cells (9, 10, 46).

293 The pattern of changes observed in fast evolving promoters like germline specific
294 piRNA factors and Nups resembles the outcome of a *P*-element mutagenesis experiment in the
295 *egghead* (*egh*) locus coding for a glycosphingolipid biosynthesis enzyme (47). Here, multiple

296 base changes were observed in the region of the first promoter after *P*-element mutagenesis,
297 but not an actual *P*-element insert, resulting in a sex peptide insensitive allele egh^{cm}. Mutations
298 in the *egh* gene result in pleiotropic phenotypes and the egh^{cm} allele disrupts neuronal wiring
299 required for the female post-mating response and development of the optic lobe (21, 47, 48).
300 Since establishing transposon resistance involves forced selection, genes whose
301 expression changes result in pleiotropic effect, like channel Nup54, could combine adaptations
302 in piRNA processing with changes in neuronal wiring resulting in altered behavioural
303 preferences (19). Of note, in *D. simulans*, a very close relative of *D. melanogaster* projections
304 of *fruitless* P1 sensory neurons that control courtship have changed and alter mate preference
305 (49).
306 The severe impact of hybrid dysgenesis on fertility likely limits fast evolution to species
307 with high numbers of progeny. Perhaps this can explain the differences seen when comparing
308 insect and mammalian promoters that are generally more conserved (25). However, rapid
309 promoter evolution has also been observed in primate promoters, suggesting a common mode
310 of evolution that has yet to be explored (19, 20, 25, 50).
311
312 **Methods**
313 **Sequence/data retrieval and alignment**
314 Gene and promotor sequences were retrieved from UCSC Genome Browser using the
315 UCSC Table Browser sequence retrieval tool (32, 33). A standardised region of 2000 bases
316 upstream of the annotated gene TSS was exported for each gene to ensure inclusion of promoter
317 regions. Sequences were imported and aligned with clustalW in MEGA11 (51). PhyloP27way
318 data were sourced from UCSC genome browser (genome.ucsc.edu) through the Table Browser
319 tool (32, 33). Data points were collected for a region of 1000 nucleotides upstream and 300
320 nucleotides downstream for each of the analysed genes. Genes without full species coverage

321 (where ≥1 species had no conserved genomic sequence) or PhyloP coverage (where ≥1
322 nucleotide in gene or upstream regions had no homologous sequence in the 27 related species)
323 were removed from the analysis.
324
325 **Molecular evolution of open reading frames analysis**
326 Analysis of adaptive protein evolution was performed via MKTs (31) using the PopFly
327 online database (imkt.uab.cat) which uses Drosophila Genome Nexus project sequence data
328 (52-54). Polymorphism and divergence data were collected from the ancestral Congo
329 population, chosen since it is a sub-Saharan population with comparably high ancestral stability
330 compared to other populations. The 'Standard MKT' test was used to calculate results for
331 individual genes. Fisher's Exact Test with significance defined as Bonferroni corrected $p \leq 0.05$
332 was used to calculate significance between nonsynonymous to synonymous polymorphisms
333 within *D. melanogaster* or between *D. melanogaster* and *D. simulans*.
334
335 **Identification of promotor hot spots and comparison of substitution rates**
336 All gene models were manually inspected in FlyBase using the JBrowse browser and
337 dominant transcripts were chosen based on comparative modENCODE expression data (55).
338 Gene and promoter sequence alignments were trimmed to -1000/+300 nucleotides from the
339 TSS of each gene. Alignments were translated into 'events' for each species compared to *D.*
340 *melanogaster*, where 0 signified no measurable sequence difference between all analysed
341 species and 1 signified sequence divergence (56). Events were calculated for base changes and
342 indels concatenated, as well as base changes and indels individually. The sum of events was
343 calculated for concatenated gene groups and individual genes at each nucleotide position. A
344 sliding event (*Se*) score was calculated from this using a sliding window of five bases along
345 the sequence, from which heatmaps were generated. To calculate the percentage of events

346 greater than the average control promoter *Se* score (*d*), a 350-nucleotide region upstream of the
347 estimated TATA box region was analysed where the total number of *Se* scores exceeding the
348 control group average sliding event score (Se^C) was divided by the total number of events in
349 that region (*N*).

350 $$d = \frac{\text{Total number of } Se \text{ events where } Se > Se^C}{N} \text{x}100$$

351 To calculate the promoter region *d* scores from PhyloP data, the total number of PhyloP
352 (*p*) scores in the 350 nucleotide regions less than the control group average promoter region
353 (p^C) took the place of the total number of *Se* events where *Se* is greater than Se^C (see equation
354 below).

355 $$d^P = \frac{\text{Total number of } p \text{ scores} < p^C}{N} \text{x}100$$

356 Significance was calculated using non-parametric chi-squared tests compared to the
357 control group *d* score. Significance values where $p \leq 0.05$ with Bonferroni correction were
358 considered statistically significant.

359

360 **Identification of somatic and germ cell-specifically expressed genes in *D. melanogaster***

361 We defined somatic and germline genes using several steps. First, using processed RNA-seq
362 data from vasa-GFP+/- cells sorted from cultured ovary GSC line (vasa-GFP;hs-bam;bamΔ86)
363 (GSE119862) (40), we defined germline-enriched (vasa-GFP+) using logFC>0, p-value<0.01
364 parameters and soma-enriched (vasa-GFP-) genes using logFC<0, p-value<0.01 parameters
365 (Figures 4A and 4B). Next, using processed RNA-seq data from fly ovaries (modENCODE),
366 we filtered genes that are expressed in adult fly ovaries using RPKM>10 parameter for both
367 soma and germline genes. To stringently define the germline genes, we additionally used
368 processed RNA-seq data from early fly embryos (0-2hr, modENCODE; mRNA-seq fly embryo
369 0-2hr; RPKM; mE_mRNA_em0-2hr; FBlc0000086) to further filter the germline-enriched

370 genes using >9 RPKM parameter. To stringently define the somatic-specific genes, we
371 additionally used processed RNA-seq data from ovarian somatic cells (OSCs) (n=4,
372 GSE160860) (41) and filtered the soma-enriched genes using >20 RPKM to select genes that
373 are expressed in OSCs as well as parameter <2 RPKM for early fly embryo (0-2hr,
374 modENCODE) to filter out genes that are expressed in early embryos. This analysis generated
375 a list of 107 germ cell-specifically expressed genes and of 59 somatic cell-specifically
376 expressed genes (Fig. 4A and B, Supplementary Dataset 1).
377
378 **Comparative gene expression analysis between *D. melanogaster* and *D. yakuba***
379 Using publicly available raw RNA-seq data from *D. melanogaster* (n=3) and *D. yakuba* (n=4)
380 ovaries, we trimmed adapter sequences from the raw reads using Cutadapt tool (v1.18, default
381 parameters) and aligned the trimmed reads to the genome assemblies (dm6 and droYak2) using
382 the RNA-seq aligner STAR (v2.7.3a). Gene counts were quantified using the featureCounts
383 tool (Subread package v1.5.3). Orthologues genes were identified using blastDm2FB in UCSC
384 Table Browser. Gene count normalization and differential gene expression analysis between
385 the species were performed using DESeq2 (v1.30.1) with an additional species-specific gene
386 lengths normalization step.
387
388 **Supplementary information**
389 Supplementary information includes Fig. S1-4, Table S1 and Supplementary Dataset 1.
390
391 **Acknowledgements**
392 We thank the Soller lab for helpful discussion.
393
394 **Funding**

This work was supported by the Medical Research Council Integrated Midlands Partnership for Biomedical Training Program to D.W.J.M. and the Biotechnology and Biological Science Research Council to M.S. Work by B.C.N and A.A. was funded by Cancer Research UK (A21143) and the Wellcome Trust (110161/Z/15/Z).

Authors' contributions

D.W.J.M and M.S. conceived and directed the project. D.W.J.M. and A.A. analysed data. D.W.J.M. wrote the initial draft and all authors contributed to the final manuscript. All authors read and approved the final manuscript.

Availability of data and materials

All data generated or analysed during this study are included in the supplementary information files.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

419 1. Iwakawa H-o, Tomari Y. Life of RISC: Formation, action, and degradation of RNA-
420 induced silencing complex. Molecular Cell. 2022;82(1):30-43.
421 2. Czech B, Munafò M, Ciabrelli F, Eastwood EL, Fabry MH, Kneuss E, et al. piRNA-
422 Guided Genome Defense: From Biogenesis to Silencing. Annual Review of Genetics.
423 2018;52(1):131-57.
424 3. Yamashiro H, Siomi MC. PIWI-Interacting RNA in Drosophila: Biogenesis,
425 Transposon Regulation, and Beyond. Chemical Reviews. 2018;118(8):4404-21.
426 4. Slotkin RK, Martienssen R. Transposable elements and the epigenetic regulation of
427 the genome. Nature Reviews Genetics. 2007;8(4):272-85.
428 5. Malone CD, Brennecke J, Dus M, Stark A, McCombie WR, Sachidanandam R, et al.
429 Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary.
430 Cell. 2009;137(3):522-35.
431 6. Khurana JS, Theurkauf W. piRNAs, transposon silencing, and Drosophila germline
432 development. Journal of Cell Biology. 2010;191(5):905-13.
433 7. Senti K-A, Brennecke J. The piRNA pathway: a fly's perspective on the guardian of
434 the genome. Trends in Genetics. 2010;26(12):499-509.
435 8. Sakakibara K, Siomi MC. The PIWI-Interacting RNA Molecular Pathway: Insights
436 From Cultured Silkworm Germline Cells. BioEssays. 2018;40(1):1700068.
437 9. Moelling K. Epigenetics and transgenerational inheritance. J Physiol. 2023.
438 10. Grentzinger T, Armenise C, Brun C, Mugat B, Serrano V, Pelisson A, et al. piRNA-
439 mediated transgenerational inheritance of an acquired trait. Genome Res. 2012;22(10):1877-
440 88.
441 11. Wang X, Ramat A, Simonelig M, Liu M-F. Emerging roles and functional
442 mechanisms of PIWI-interacting RNAs. Nature Reviews Molecular Cell Biology.
443 2023;24(2):123-41.

444 12. Sato K, Siomi MC. The piRNA pathway in Drosophila ovarian germ and somatic
445 cells. Proceedings of the Japan Academy, Series B. 2020;96(1):32-42.
446 13. Czech B, Hannon GJ. One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-
447 Guided Silencing. Trends in Biochemical Sciences. 2016;41(4):324-37.
448 14. Haase AD. An introduction to PIWI-interacting RNAs (piRNAs) in the context of
449 metazoan small RNA silencing pathways. RNA Biology. 2022;19(1):1094-102.
450 15. Czech B, Preall J, B. , McGinn J, Hannon G, J. . A Transcriptome-wide RNAi Screen
451 in the Drosophila Ovary Reveals Factors of the Germline piRNA Pathway. Molecular Cell.
452 2013;50(5):749-61.
453 16. Brown JS, Zhang D, Chen W, Lee H-C. Sensitized piRNA reporter identifies multiple
454 RNA processing factors involved in piRNA-mediated gene silencing. 2023.
455 17. Handler D, Meixner K, Pizka M, Lauss K, Schmied C, Franz, et al. The Genetic
456 Makeup of the Drosophila piRNA Pathway. Molecular Cell. 2013;50(5):762-77.
457 18. Munafò M, Lawless VR, Passera A, Macmillan S, Bornelöv S, Haussmann IU, et al.
458 Channel nuclear pore complex subunits are required for transposon silencing in Drosophila.
459 eLife. 2021;10.
460 19. Nallasivan MP, Haussmann IU, Civetta A, Soller M. Channel nuclear pore protein 54
461 directs sexual differentiation and neuronal wiring of female reproductive behaviors in
462 Drosophila. BMC Biology. 2021;19(1).
463 20. McQuarrie DWJ, Read AM, Stephens FHS, Civetta A, Soller M. Indel driven rapid
464 evolution of core nuclear pore protein gene promoters. Scientific Reports. 2023;13(1).
465 21. Haussmann IU, Hemani Y, Wijesekera T, Dauwalder B, Soller M. Multiple pathways
466 mediate the sex-peptide-regulated switch in female Drosophila reproductive behaviours. Proc
467 Biol Sci. 2013;280(1771):20131938.

468 22. Tang S, Presgraves DC. Evolution of the Drosophila Nuclear Pore Complex Results
469 in Multiple Hybrid Incompatibilities. Science. 2009;323(5915):779-82.
470 23. Tang S, Presgraves DC. Lineage-Specific Evolution of the Complex Nup160 Hybrid
471 Incompatibility Between Drosophila melanogaster and Its Sister Species. Genetics.
472 2015;200(4):1245-54.
473 24. Iwasaki YW, Siomi MC, Siomi H. PIWI-Interacting RNA: Its Biogenesis and
474 Functions. Annu Rev Biochem. 2015;84:405-33.
475 25. Taylor MS, Kai C, Kawai J, Carninci P, Hayashizaki Y, Semple CAM. Heterotachy in
476 Mammalian Promoter Evolution. PLoS Genetics. 2006;2(4):e30.
477 26. Balacco DL, Soller M. The m(6)A Writer: Rise of a Machine for Growing Tasks.
478 Biochemistry. 2019;58(5):363-78.
479 27. Bawankar P, Lence T, Paolantoni C, Haussmann IU, Kazlauskiene M, Jacob D, et al.
480 Hakai is required for stabilization of core components of the m6A mRNA methylation
481 machinery. Nature Communications. 2021;12(1):3778.
482 28. Van Lopik J, Trapotsi M-A, Hannon GJ, Bornelov S, Czech Nicholson B. Unistrand
483 piRNA clusters are an evolutionary conserved mechanism to suppress endogenous
484 retroviruses across the Drosophila genus. bioRxiv. 2023.
485 29. Fridrich A, Moran Y. Some flies do not play ping-pong. PLOS Biology.
486 2023;21(6):e3002152.
487 30. Chary S, Hayashi R. Mechanistic divergence of piRNA biogenesis in Drosophila.
488 2022.
489 31. McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in
490 Drosophila. Nature. 1991;351(6328):652-4.
491 32. Karolchik D. The UCSC Table Browser data retrieval tool. Nucleic Acids Research.
492 2004;32(90001):493D-6.

493 33. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The
494 Human Genome Browser at UCSC. Genome Research. 2002;12(6):996-1006.
495 34. Consortium GO. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids
496 Res. 2021;49(D1):D325-d34.
497 35. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene
498 ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet.
499 2000;25(1):25-9.
500 36. Mi H, Muruganujan A, Ebert D, Huang X, Thomas PD. PANTHER version 14: more
501 genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools.
502 Nucleic Acids Research. 2018;47(D1):D419-D26.
503 37. Sanfilippo P, Wen J, Lai EC. Landscape and evolution of tissue-specific alternative
504 polyadenylation across Drosophila species. Genome Biology. 2017;18(1).
505 38. Smolko AE, Shapiro-Kulnane L, Salz HK. The H3K9 methyltransferase SETDB1
506 maintains female identity in Drosophila germ cells. Nature Communications. 2018;9(1).
507 39. Vankuren NW, Vibranovski MD. A Novel Dataset for Identifying Sex-Biased Genes
508 in Drosophila. Journal of Genomics. 2014;2:64-7.
509 40. Story B, Ma X, Ishihara K, Li H, Hall K, Peak A, et al. Defining the expression of
510 piRNA and transposable elements in Drosophila ovarian germline stem cells and somatic
511 support cells. Life Science Alliance. 2019;2(5):e201800211.
512 41. Eastwood EL, Jara KA, Bornelöv S, Munafò M, Frantzis V, Kneuss E, et al.
513 Dimerisation of the PICTS complex via LC8/Cut-up drives co-transcriptional transposon
514 silencing in Drosophila. eLife. 2021;10:e65557.
515 42. Treiber CD, Waddell S. Transposon expression in the Drosophila brain is driven by
516 neighboring genes and diversifies the neural transcriptome. Genome Research.
517 2020;30(11):1559-69.

518 43. Laverty C, Lucci J, Akhtar A. The MSL complex: X chromosome and beyond.
519 Current Opinion in Genetics & Development. 2010;20(2):171-8.
520 44. Straub T, Becker PB. Dosage compensation: the beginning and end of generalization.
521 Nature Reviews Genetics. 2007;8(1):47-57.
522 45. Brennecke J, Malone CD, Aravin AA, Sachidanandam R, Stark A, Hannon GJ. An
523 Epigenetic Role for Maternally Inherited piRNAs in Transposon Silencing. Science.
524 2008;322(5906):1387-92.
525 46. Luo Y, He P, Kanrar N, Fejes Toth K, Aravin AA. Maternally inherited siRNAs
526 initiate piRNA cluster formation. Mol Cell. 2023.
527 47. Soller M, Haussmann IU, Hollmann M, Choffat Y, White K, Kubli E, et al. Sex-
528 peptide-regulated female sexual behavior requires a subset of ascending ventral nerve cord
529 neurons. Curr Biol. 2006;16(18):1771-82.
530 48. Fan Y, Soller M, Flister S, Hollmann M, Müller M, Bello B, et al. The egghead gene
531 is required for compartmentalization in Drosophila optic lobe development. Dev Biol.
532 2005;287(1):61-73.
533 49. Seeholzer LF, Seppo M, Stern DL, Ruta V. Evolution of a central neural circuit
534 underlies Drosophila mate preferences. Nature. 2018;559(7715):564-9.
535 50. Main BJ, Smith AD, Jang H, Nuzhdin SV. Transcription Start Site Evolution in
536 Drosophila. Molecular Biology and Evolution. 2013;30(8):1966-74.
537 51. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics
538 Analysis Version 11. Molecular Biology and Evolution. 2021;38(7):3022-7.
539 52. Lack JB, Cardeno CM, Crepeau MW, Taylor W, Corbett-Detig RB, Stevens KA, et
540 al. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila
541 melanogaster Genomes, Including 197 from a Single Ancestral Range Population. Genetics.
542 2015;199(4):1229-41.

543 53. Lack JB, Lange JD, Tang AD, Corbett-Detig RB, Pool JE. A Thousand Fly Genomes:
544 An Expanded Drosophila Genome Nexus. Mol Biol Evol. 2016;33(12):3308-13.
545 54. Mackay TFC, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The
546 Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482(7384):173-8.
547 55. Gramates LS, Agapite J, Attrill H, Calvi BR, Crosby MA, dos Santos G, et al.
548 FlyBase: a guided tour of highlighted features. Genetics. 2022;220(4).
549 56. Tang H, Lewontin RC. Locating Regions of Differential Variability in DNA and
550 Protein Sequences. Genetics. 1999;153(1):485-95.
551

552 **Table and Figure Legends**

553 **Fig. 1: Promoter regions of germ cell-expressed piRNA processing factors are hot spots**
554 **for rapid evolution**

555 A and B) Schematic depiction of somatic piRNA factors involved in co-transcriptional gene
556 silencing and piRNA biogenesis (A) and germline-specific piRNA pathway genes involved in
557 dual-strand cluster regulation and the ping-pong cycle (B) in the *Drosophila* ovary.
558 C and D) Evolutionary conservation per nucleotide around the TSS depicted by PhyloP27way
559 conservation score averages for the *D. melanogaster* dm6 genome (black) and the somatic
560 (grey, C) or the germ cell-specific piRNA factors (grey, D).
561 E and F) PhyloP27way conservation score averages (E) and PhyloP27way conservation
562 promoter d^P scores (F) for the 350 nucleotide promoter regions of the somatic piRNA factors
563 and the germ cell-specific piRNA factors compared to all genes in the *Drosophila* genome.
564 Statistically significant differences from unpaired student t-tests (E) and non-parametric chi-
565 squared tests (F) are indicated by asterisks (*** $p \leq 0.001$, **** $p \leq 0.0001$ following Bonferroni
566 correction).

567 G-L) Heatmaps indicating sequence (E, purple), indel (G, blue) or base change (I, green)
568 accumulation, and their quantification (F-J) depicted as *d* scores for each of the analysed gene
569 groups compared to the control among closely related *D. melanogaster*, *D. simulans*, *D.*
570 *sechellia*, *D. yakuba* and *D. erecta*. Regions of 1000 nucleotides upstream and 300 nucleotides
571 downstream of the TSS were analysed. Analysis was performed for the control group (m^6A
572 writer complex and readers), the somatic piRNA factors, and the germ cell-specific piRNA
573 factors. The blue line indicates the promoter region used for quantification of the substitution
574 rate based on the gene and intergenic region schematic. Statistically significant differences
575 from non-parametric chi-squared tests are indicated by asterisks (**** $p \leq 0.0001$ following
576 Bonferroni correction).
577
578 **Fig. 2: Promoters of genes required for germline transposon silencing evolve fast**
579 A) PhyloP27way nucleotide scores for the top 100 genes affecting transposon silencing in germ
580 cells based on transposon derepression *z* scores and ordered by promoter conservation d^P
581 scores. A region of 1000 nucleotides upstream and 300 nucleotides downstream are shown.
582 Red represents lower conservation while blue represents higher.
583 B) Average PhyloP27way nucleotide scores for the top 100 genes affecting transposon
584 silencing in germ cells compared to the genome average.
585 C) Comparison of PhyloP27way conservation promoter scores (d^P) for each of the 100 genes
586 affecting transposon silencing in germ cells compared to the genome average. Statistically
587 significant differences from non-parametric chi-squared tests are indicated by asterisks (***
588 $p \leq 0.001$ following Bonferroni correction).
589 D) Representative GO analysis for the top 100 gene candidates affecting transposon silencing.
590 E) Individual gene comparison of PhyloP27way conservation promoter scores (d^P) for the top
591 100 genes affecting transposon silencing in germ cells compared to the genome average.

592 Statistically significant differences from non-parametric chi-squared tests are indicated by
593 asterisks (* p≤0.05, ** p≤0.01, *** p≤0.001, **** p≤0.0001 following Bonferroni correction).
594 F-H) Sequence analysis for change accumulation among closely related *D. melanogaster*, *D.*
595 *simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. Comparisons were performed for sequence
596 change (F), indel (G), and base change (H) *d* scores for each of the positive 49 genes affecting
597 transposon silencing in FEPG gene group compared to the somatic piRNA factors. Statistically
598 significant differences from non-parametric chi-squared tests are indicated by asterisks (****
599 p≤0.0001 following Bonferroni correction).
600 I) Comparison of *d* scores for protein complex genes involved in germ cell transposon silencing
601 compared to the somatic piRNA factors. Statistically significant differences from non-
602 parametric chi-squared tests are indicated by asterisks (**** p≤0.0001 following Bonferroni
603 correction).
604 J) Heatmaps indicating sequence change accumulation for protein complex genes involved in
605 germ cell transposon silencing compared to the piRNA factors shown in purple among closely
606 related *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. Regions of 1000
607 nucleotides upstream and downstream of the TSS were analysed. The blue line indicates the
608 promoter region used for quantification of the substitution rate.
609
610 **Fig. 3: Divergent expression of FEPG genes in fly ovaries is accompanied by rapid**
611 **promoter evolution**
612 A) Volcano plot showing differential gene expression (DESeq2) between *D. melanogaster* and
613 *D. yakuba* ovaries (RNA-seq data; Supplementary Dataset 1) for FEPG (blue) and the somatic
614 piRNA factors (red). Horizontal dashed lines indicate significance thresholds (adjusted
615 p≤0.05).

616 B-D) Alignment of the promoter regions of *CG32152, MEP-1* and *Rbbp5* for *D. melanogaster*
617 and *D. yakuba* with indels and base changes indicated as black boxes. The black lines indicate
618 gene 5'UTRs.
619
620 **Fig. 4: Germ cell-specifically expressed, but not somatic cell-specifically expressed gene**
621 **promoters evolve rapidly and are accompanied by divergence in gene expression**
622 A) Ranking of genes using vasa-GFP signal from FACS-sorted vasa-GFP+/- cells of cultured
623 ovary GSC line (vasa-GFP;hs-bam;bamΔ86) (GSE119862) that defines germline-enriched
624 (vasa-GFP+) and soma-enriched (vasa-GFP-) genes.
625 B) Volcano plot showing differential gene expression between the FACS-sorted vasa-GFP+
626 (germline) and vasa-GFP- (somatic) cells from cultured ovary GSC line (vasa-GFP;hs-
627 bam;bamΔ86) (GSE119862).
628 C and D) PhyloP27way nucleotide scores for somatic cell-specific (C) and germ cell-specific
629 (D) expressed genes ordered by promoter conservation d^P scores. A region of 1000 nucleotides
630 upstream and 300 nucleotides downstream are shown. Red represents lower conservation while
631 blue represents higher.
632 E) Average PhyloP27way nucleotide scores for all genes in the *Drosophila* genome, somatic
633 cell and germ cell-specific genes. Purple represents lower conservation while blue represents
634 higher.
635 F) The percentage of somatic cell and germ cell-specific genes with significant promoter
636 nucleotide changes (d^P) versus all genes in the *Drosophila* genome average, divided by
637 significantly increased (low conservation) or decreased (high conservation). Statistically
638 significant differences from non-parametric chi-squared tests are indicated by asterisks (****
639 $p \leq 0.0001$ following Bonferroni correction).

640 G and H) PhyloP27way conservation score averages (G) and PhyloP27way conservation
641 promoter scores (d^P) (H) for the 350 nucleotide promoter regions compared for all genes in the
642 *Drosophila* genome, somatic cell-specifically expressed and germ cell-specifically expressed
643 genes. Statistically significant differences from unpaired student t-tests (G) and from non-
644 parametric chi-squared tests (H) are indicated by asterisks (** $p \leq 0.01$, **** $p \leq 0.0001$
645 following Bonferroni correction).
646 I) Volcano plot of differential gene expression (DESeq2) between *D. melanogaster* and *D.*
647 *yakuba* ovaries (RNA-seq data; Supplementary Dataset 1) for all (grey), germ cell-specific
648 (blue) and somatic cell-specific (red) gene groups. Horizontal dashed lines indicate
649 significance (adjusted $p \leq 0.05$).
650 J) Scatterplot of individual gene d^P scores plotted against fold-change expression differences
651 between *D. melanogaster* and *D. yakuba* ovaries (RNA-seq data; Supplementary Dataset 1) for
652 germ cell-specific (blue) and somatic cell-specific (red) gene groups. The grey dotted lines
653 indicate log2Fold-change threshold values of ≥ 0.5 and ≤ 0.5.
654 K and L) Comparison of the percentage of genes with significant changes in *D. melanogaster*
655 log2FoldChange (≥ 0.5, K) or significant changes in *D. yakuba* log2FoldChange (≤ 0.5, L)
656 expression changes in PhyloP conservation promoter score (d^P) ranges 25-50, >50-75, and >75-
657 100. Expression changes for germ cell-specific (blue) and somatic cell-specific (red) gene
658 groups were analysed between *D. melanogaster* and *D. yakuba*. Statistically significant
659 differences from non-parametric chi-squared tests are indicated by asterisks (** $p \leq 0.01$, ***
660 $p \leq 0.001$, **** $p \leq 0.0001$ following Bonferroni correction).
661

Figure 1

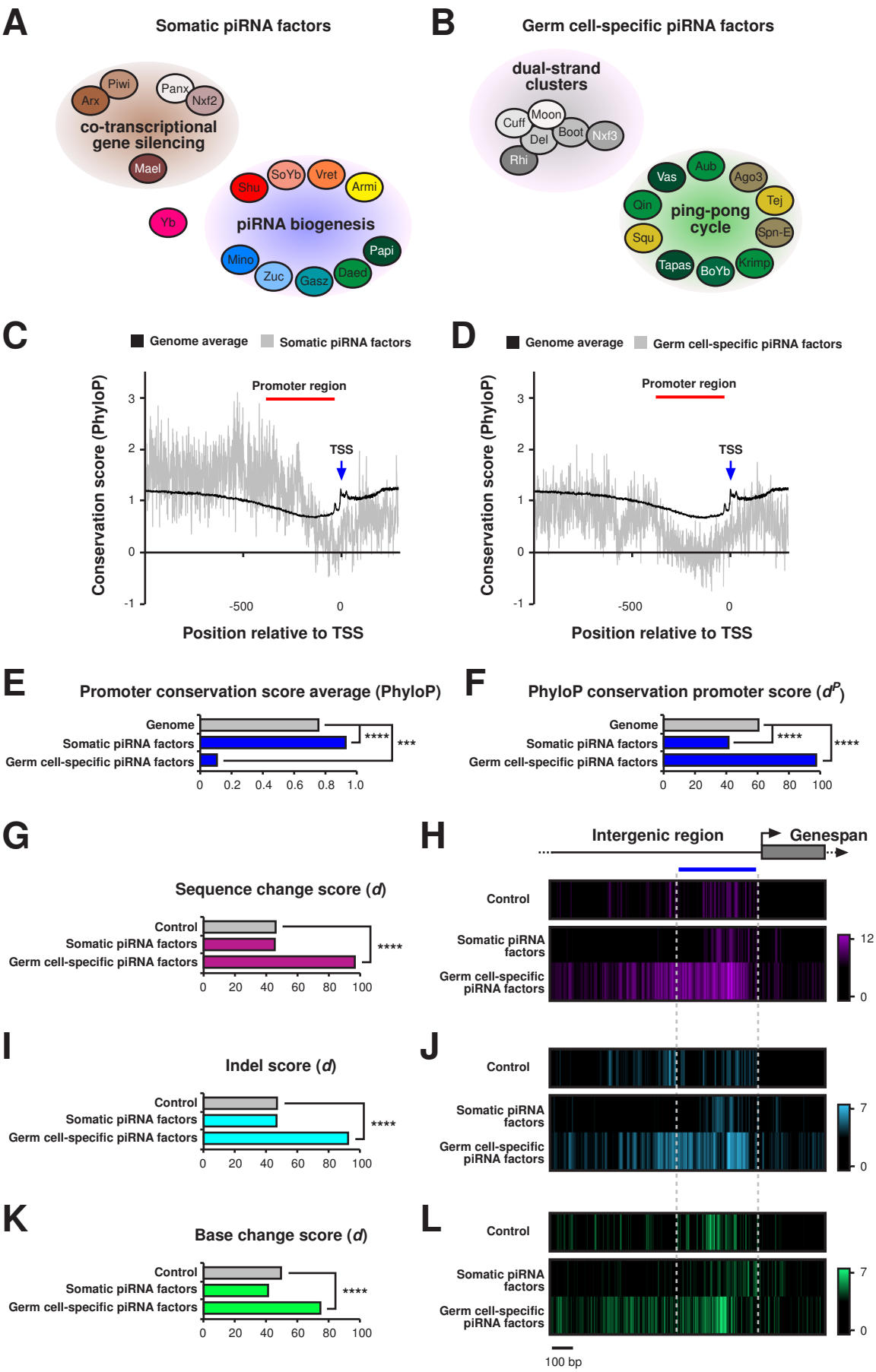

A
Somatic piRNA factors
Piwi
Panx
Arx
Nxf2
co-transcriptional gene silencing
Mael
Shu
SoYb
Vret
Armi
Yb
piRNA biogenesis
Papi
Mino
Zuc
Gasz
Daed
B
Germ cell-specific piRNA factors
dual-strand clusters
Moon
Cuff
Boot
Del
Nxf3
Rhi
Vas
Aub
Ago3
Tej
Qin
ping-pong cycle
Spn-E
Squ
Tapas
BoYb
Krimp
C
Genome average
Somatic piRNA factors
Conservation score (PhyloP)
Promoter region
TSS
Position relative to TSS
D
Genome average
Germ cell-specific piRNA factors
Conservation score (PhyloP)
Promoter region
TSS
Position relative to TSS
E
Promoter conservation score average (PhyloP)
Genome
Somatic piRNA factors
Germ cell-specific piRNA factors
F
PhyloP conservation promoter score (dP)
Genome
Somatic piRNA factors
Germ cell-specific piRNA factors
G
Sequence change score (d)
Control
Somatic piRNA factors
Germ cell-specific piRNA factors
H
Intergenic region
Genespan
Control
Somatic piRNA factors
Germ cell-specific piRNA factors
I
Indel score (d)
Control
Somatic piRNA factors
Germ cell-specific piRNA factors
J
Control
Somatic piRNA factors
Germ cell-specific piRNA factors
K
Base change score (d)
Control
Somatic piRNA factors
Germ cell-specific piRNA factors
L
Control
Somatic piRNA factors
Germ cell-specific piRNA factors
100 bp

Figure 2

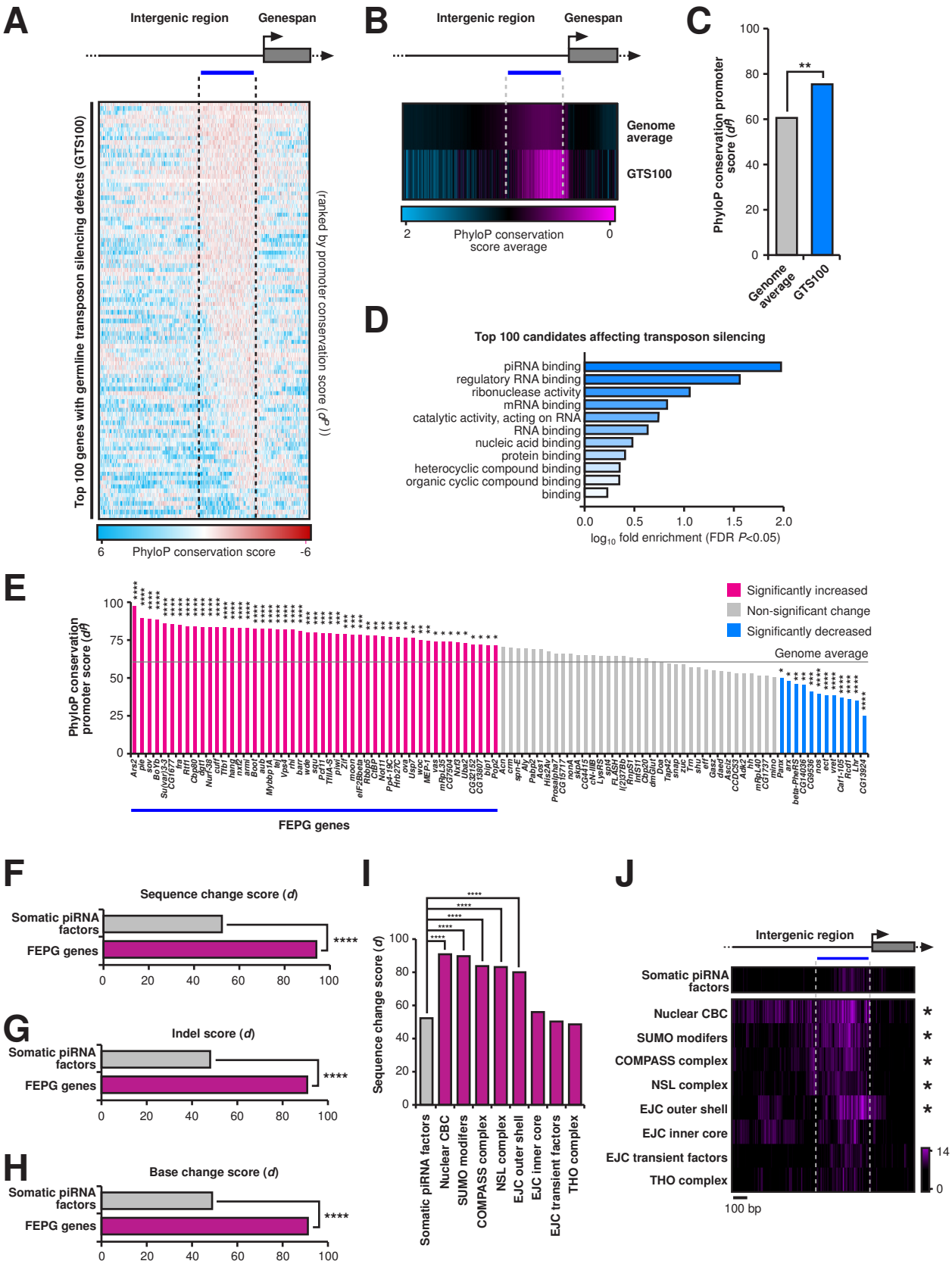
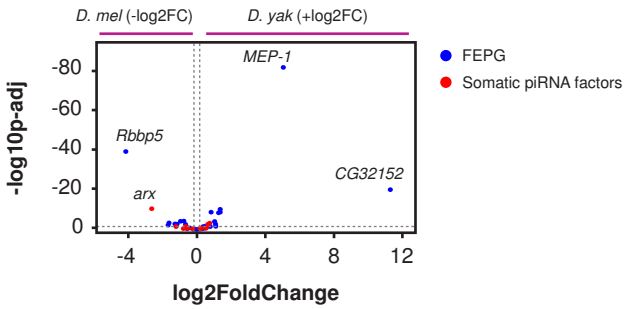
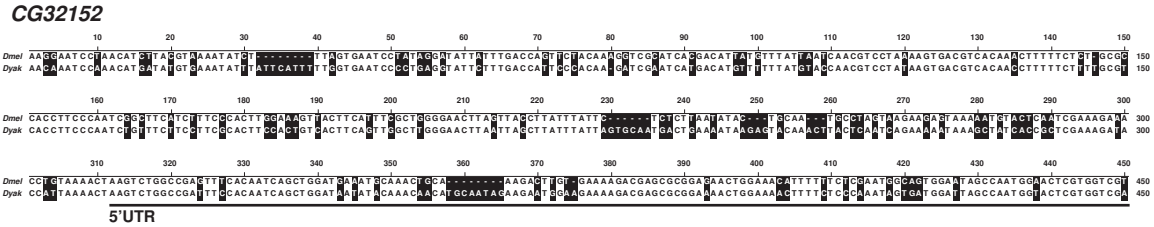

A
Intergenic region
Genespan
Top 100 genes with germline transposon silencing defects (GTS100)
(ranked by promoter conservation score (dP))
6
PhyloP conservation score
-6
B
Intergenic region
Genespan
Genome average
GTS100
2
PhyloP conservation score average
0
C
PhyloP conservation promoter score (dP)
100
80
60
40
20
0
**
Genome average
GTS100
D
Top 100 candidates affecting transposon silencing
piRNA binding
regulatory RNA binding
ribonuclease activity
mRNA binding
catalytic activity, acting on RNA
RNA binding
nucleic acid binding
protein binding
heterocyclic compound binding
organic cyclic compound binding
binding
0.0
0.5
1.0
1.5
2.0
log10 fold enrichment (FDR P<0.05)
E
PhyloP conservation promoter score (dP)
100
75
50
25
0
Significantly increased
Non-significant change
Significantly decreased
Genome average
FEPG genes
F
Sequence change score (d)
Somatic piRNA factors
FEPG genes
****
0
20
40
60
80
100
G
Indel score (d)
Somatic piRNA factors
FEPG genes
****
0
20
40
60
80
100
H
Base change score (d)
Somatic piRNA factors
FEPG genes
****
0
20
40
60
80
100
I
Sequence change score (d)
100
80
60
40
20
0
Somatic piRNA factors
Nuclear CBC
SUMO modifers
COMPASS complex
NSL complex
EJC outer shell
EJC inner core
EJC transient factors
THO complex
J
Intergenic region
Somatic piRNA factors
Nuclear CBC
SUMO modifers
COMPASS complex
NSL complex
EJC outer shell
EJC inner core
EJC transient factors
THO complex
*
*
*
*
*
14
0
100 bp

Figure 3

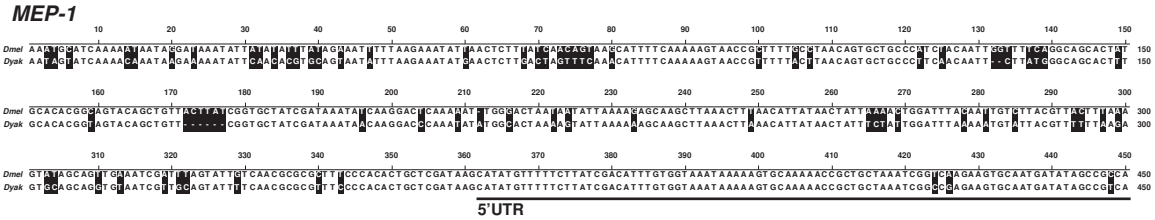
A



B



C



D

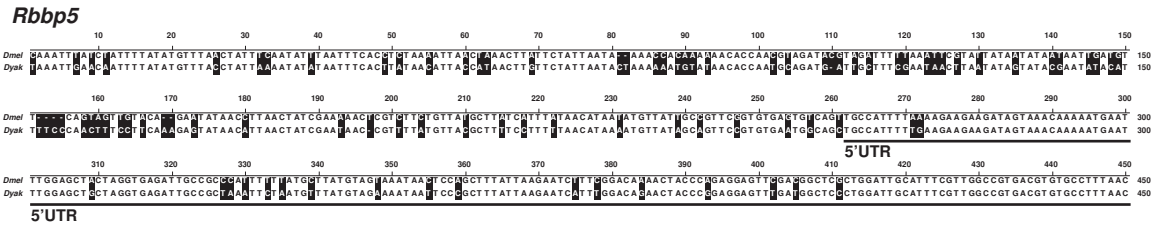
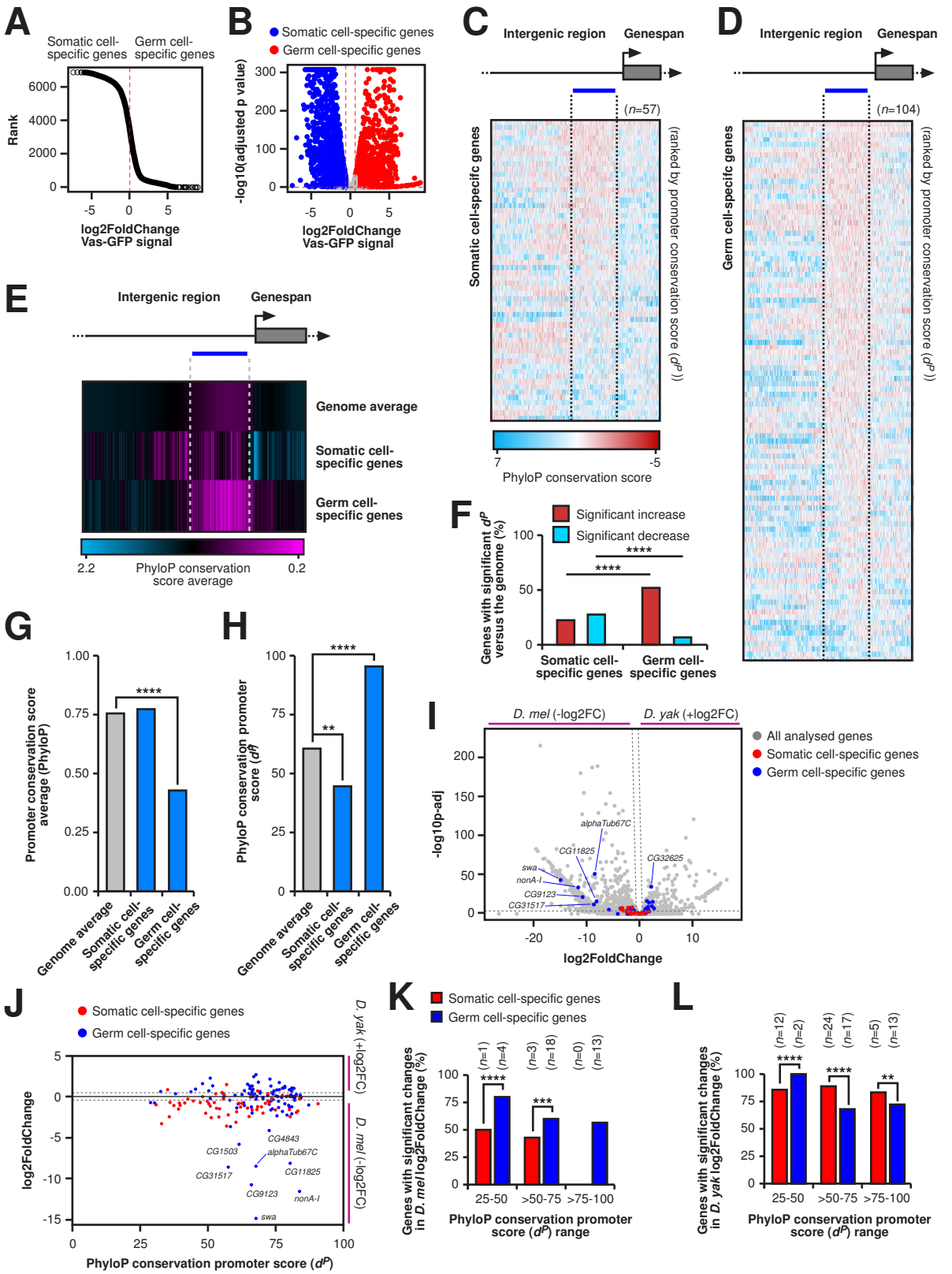


Figure 4


A
Somatic cell-specific genes
Germ cell-specific genes
Rank
6000
4000
2000
0
-5
0
5
log2FoldChange
Vas-GFP signal
B
Somatic cell-specific genes
Germ cell-specific genes
-log10(adjusted p value)
300
200
100
0
-5
0
5
log2FoldChange
Vas-GFP signal
C
Intergenic region
Genespan
(n=57)
Somatic cell-specific genes
(ranked by promoter conservation score (dP))
7
-5
PhyloP conservation score
D
Intergenic region
Genespan
(n=104)
Germ cell-specific genes
(ranked by promoter conservation score (dP))
E
Intergenic region
Genespan
Genome average
Somatic cell-specific genes
Germ cell-specific genes
2.2
PhyloP conservation score average
0.2
F
Significant increase
Significant decrease
Genes with significant dP versus the genome (%)
100
50
0
****
****
Somatic cell-specific genes
Germ cell-specific genes
G
Promoter conservation score average (PhyloP)
1.00
0.75
0.50
0.25
0.00
****
Genome average
Somatic cell-specific genes
Germ cell-specific genes
H
PhyloP conservation promoter score (dP)
100
75
50
25
0
****
**
Genome average
Somatic cell-specific genes
Germ cell-specific genes
I
D. mel (-log2FC)
D. yak (+log2FC)
All analysed genes
Somatic cell-specific genes
Germ cell-specific genes
-log10p-adj
200
150
100
50
0
alphaTub67C
CG11825
CG32625
swa
nonA-I
CG9123
CG31517
-20
-10
0
10
log2FoldChange
J
Somatic cell-specific genes
Germ cell-specific genes
D. yak (+log2FC)
D. mel (-log2FC)
log2FoldChange
5
0
-5
-10
-15
CG4843
CG1503
alphaTub67C
CG11825
CG31517
CG9123
nonA-I
swa
0
25
50
75
100
PhyloP conservation promoter score (dP)
K
Somatic cell-specific genes
Germ cell-specific genes
Genes with significant changes in D. mel log2FoldChange (%)
(n=1)
(n=4)
(n=3)
(n=18)
(n=0)
(n=13)
100
75
50
25
0
****
***
25-50
>50-75
>75-100
PhyloP conservation promoter score (dP) range
L
(n=12)
(n=2)
(n=24)
(n=17)
(n=5)
(n=13)
Genes with significant changes in D. yak log2FoldChange (%)
100
75
50
25
0
****
****
**
25-50
>50-75
>75-100
PhyloP conservation promoter score (dP) range

Supplementary Materials

Promoters of germline transposon silencing genes evolve rapidly accompanied by diverging gene expression

DAVID W. J. MCQUARRIE[1,2], AZAD ALIZADA[3], BENJAMIN CZECH NICHOLSON[3] AND MATTHIAS SOLLER[1,2*]

Supplementary Fig. 1: Coding region evolution rates for somatic and germline piRNA processing genes.

A) Evolution of coding regions analysed by 'Standard MKT' tests for polymorphisms and divergence between *D. melanogaster* and *D. simulans* for core genes in the somatic and germ cell transposon silencing pathways. R, replacement; S, synonymous. Analyses were omitted for genes without a Popfly record. Statistically significant differences from Fisher's Exact Tests are indicated by asterisks (* $p \leq 0.05$ following FDR correction).

B) Individual gene sequence change, indel, and base change *d* scores and statistically significant accumulation of sequence changes for the somatic piRNA factors and germ cell-specific piRNA factors. Statistically significant differences from non-parametric chi-squared tests are indicated by asterisks (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$ following FDR correction).

Supplementary Fig. 2: Genes coding for members of a subgroup of protein complexes implicated in piRNA silencing accumulate sequence changes in their promoters

Individual gene *d* scores and statistically significant accumulation of sequence changes for each gene separated by protein complex compared to the somatic piRNA factors. Statistically significant differences from non-parametric chi-squared tests are indicated by asterisks (* $p \leq 0.05$, ** $p \leq 0.01$, **** $p \leq 0.0001$ following Bonferroni correction).

Supplementary Fig. 3: Promoter evolution of protein complex genes involved in germ cell transposon silencing

A and C) Heatmaps indicating indel (A, blue) or base change (C, green) accumulation in protein complex genes involved in germ cell transposon silencing compared to Yb body genes among closely related *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. Regions of 1000 nucleotides upstream and 300 nucleotides downstream of the TSS were analysed. The blue line indicates the promoter region used for quantification of the substitution rate.

B and D) Comparison of indel (B, blue) or base change (D, green) *d* scores for each of the protein complex gene groups compared to the somatic piRNA factors. Statistically significant differences from non-parametric chi-squared tests are indicated by asterisks (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$ following Bonferroni correction).

E) Individual gene *d* scores and statistically significant accumulation of indels or base changes for each gene separated by protein complex compared to the somatic piRNA factors. Comparison of indel (B, blue) or base change (D, green) *d* scores for each of the protein complex gene groups compared to the control group. Statistically significant differences from non-parametric chi-squared tests are indicated by asterisks (* $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, **** $p \leq 0.0001$ following Bonferroni correction).

Supplementary Fig. 4: Coding sequences evolution of protein complex genes involved in germ cell transposon silencing.

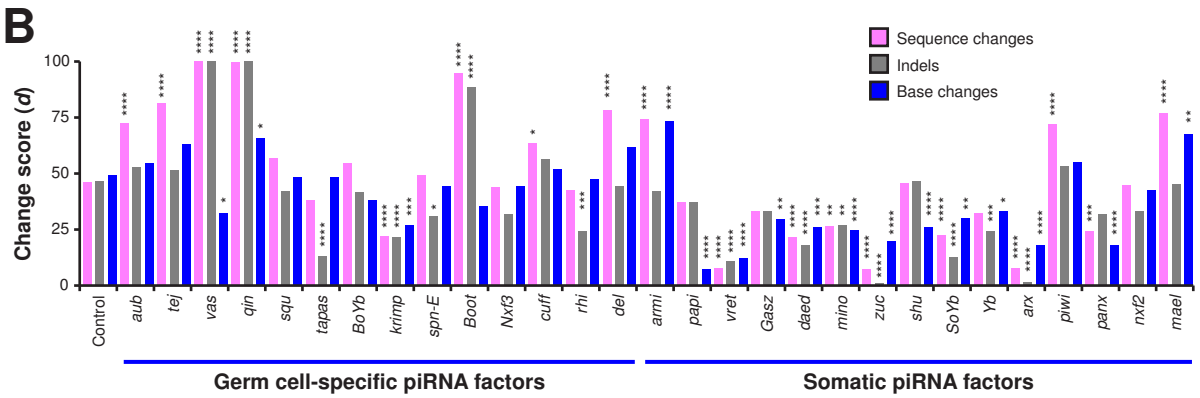
50 A) Impact of protein complex genes involved in germ cell transposon silencing on transposon
51 de-repression. Unprocessed data were taken from (1). Derepression scores (*z*) were
52 concatenated for protein complexes in each transposon type (HeTA, TAHRE, blood, burdock)
53 and an average of all.
54 B) Evolution of coding regions analysed by 'Standard MKT' tests for polymorphisms and
55 divergence between *D. melanogaster* and *D. simulans* for individual genes from the selected
56 germline transposon silencing implicated protein complexes. R, replacement; S, synonymous.
57 Analyses were omitted for genes without a Popfly record. Statistically significant differences
58 from Fisher's Exact Tests are indicated by asterisks (** $p \leq 0.01$ following FDR correction).
59
60 **References**
61 1. Czech B, Preall J, B. , McGinn J, Hannon G, J. . A Transcriptome-wide RNAi Screen
62 in the Drosophila Ovary Reveals Factors of the Germline piRNA Pathway. Molecular Cell.
63 2013;50(5):749-61.
64

Supplementary Figure 1

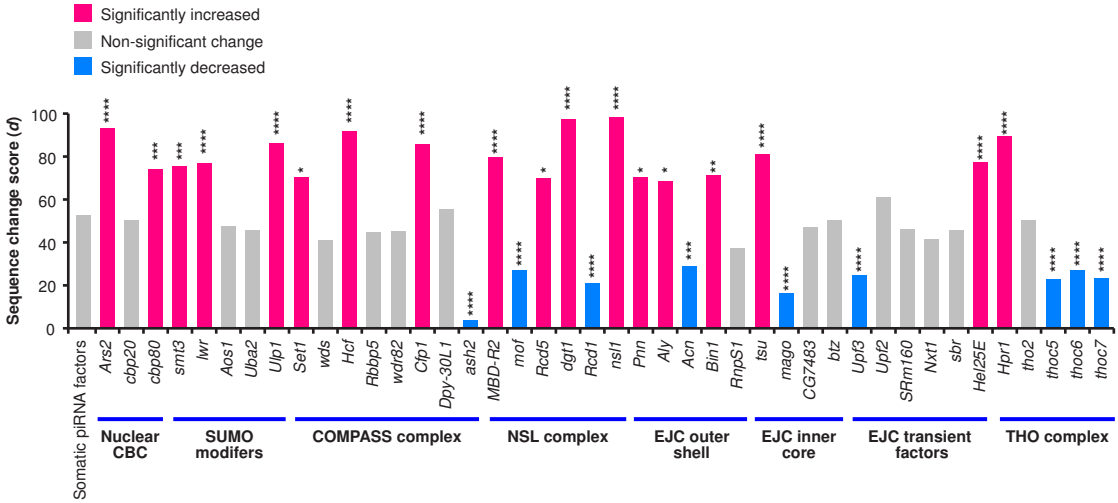
A

	Gene	Polymorphism R	Polymorphism S	Polymorphism R/S	Divergence R	Divergence S	Divergence R/S	Fisher's exact *P*-value (FDR)	
Germ cell-specific piRNA factors	*aub*	10	9	1.111	96	33	2.909	0.35700	
	tej	10	6	1.667	73	32	2.281	1.01238	
	ago3	2	1	2.000	54	41	1.317	1.00000	
	qin	33	28	1.179	175	79	2.215	0.26175	
	squ	2	1	2.000	26	16	1.625	1.03448	
	tapas	14	12	1.167	122	55	2.218	0.59603	
	BoYb	3	1	3.000	102	50	2.040	1.07143	
	krimp	19	16	1.188	126	25	5.040	0.01500	*
	spn-E	6	14	0.429	82	65	1.262	0.34100	
	Boot	3	0	-	51	18	2.833	1.06520	
	moon	0	0	-	0	0	-	1.11111	
	Nxf3	7	2	3.500	69	23	3.000	1.15385	
	cuff	3	3	1.000	52	18	2.889	0.78053	
	rhi	14	4	3.500	91	20	4.550	1.24064	
	del	7	5	1.400	152	44	3.455	0.68032	
Somatic piRNA factors	*armi*	4	12	0.333	45	32	1.406	0.38400	
	papi	2	8	0.250	9	30	0.300	1.20000	
	vret	12	12	1.000	34	30	1.133	1.28700	
	Gasz	3	1	3.000	25	8	3.125	1.25000	
	daed	5	1	5.000	13	16	0.813	0.66532	
	mino	4	7	0.571	15	23	0.652	1.30435	
	zuc	0	1	0.000	28	16	1.750	0.80957	
	shu	3	2	1.500	22	22	1.000	1.36364	
	SoYb	76	27	2.815	86	52	1.654	0.35800	
	Yb	15	8	1.875	130	38	3.421	0.60810	
	arx	4	1	4.000	6	3	2.000	1.42857	
	piwi	9	12	0.750	23	46	0.500	0.89020	
	panx	7	2	3.500	62	26	2.385	1.50000	
	nxf2	3	1	3.000	28	44	0.636	0.74580	
	mael	2	2	1.000	45	14	3.214	0.72277	

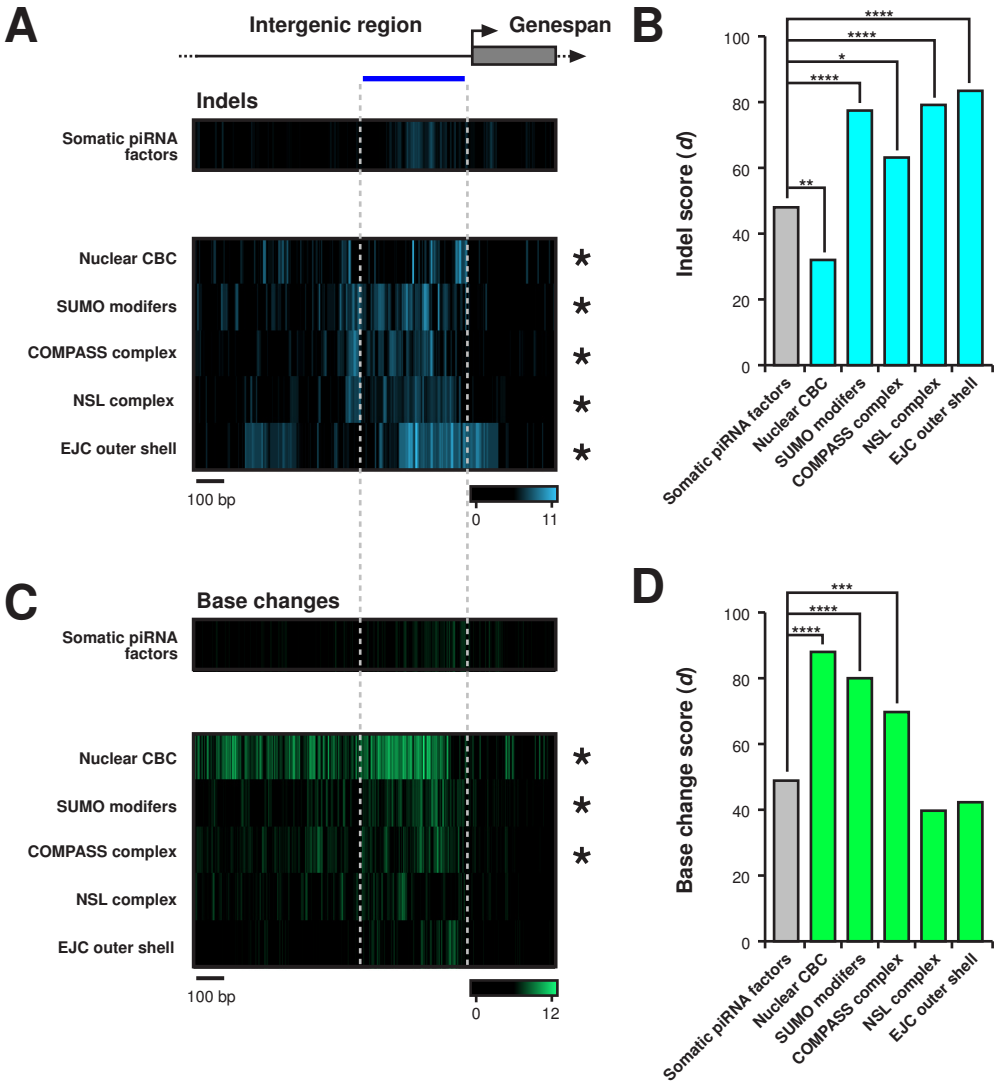
B

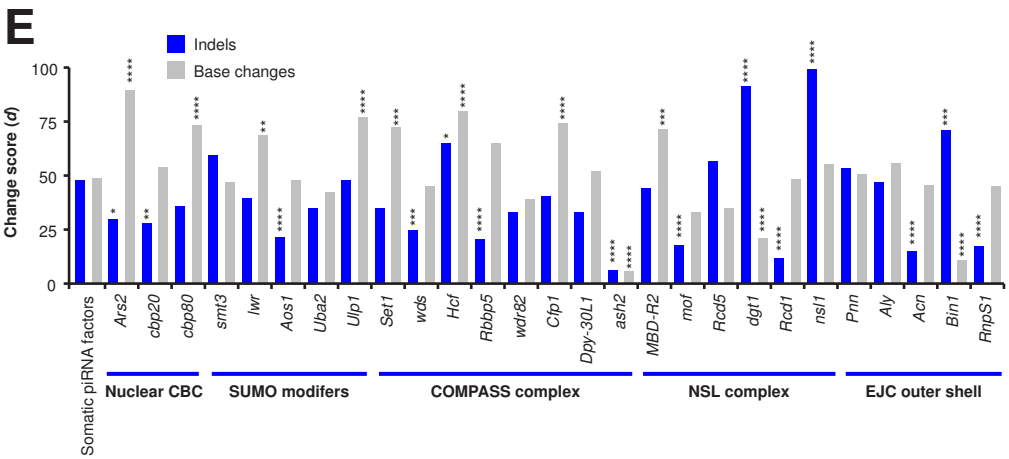


Supplementary Figure 2



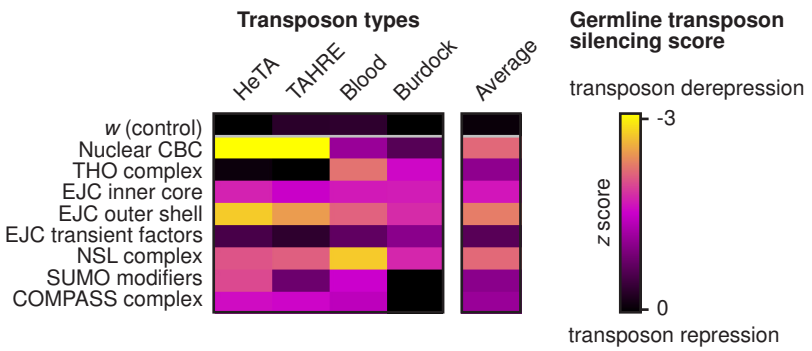
Supplementary Figure 3





Supplementary Figure 4

A



B

		Gene	Polymorphism R	Polymorphism S	Polymorphism R/S	Divergence R	Divergence S	Divergence R/S	Fisher's exact *P*-value (FDR)	
Nuclear CBC		*Ars2*	9	6	1.500	29	47	0.617	0.77150	
		cbp80	0	14	0.000	56	49	1.143	0.00292	**
		cbp20	0	0	-	1	10	0.100	1.00000	
THO complex		*thoc5*	13	9	1.444	18	14	1.286	1.02564	
		thoc7	0	0	-	4	7	0.571	1.05263	
		thoc6	8	6	1.333	7	15	0.467	0.70120	
		tho2	2	7	0.286	34	58	0.586	1.21150	
		Hpr1	1	7	0.143	14	30	0.467	1.09893	
Exon junction complex	Inner core	*tsu*	0	0	-	2	6	0.333	1.08108	
		mago	0	3	0.000	0	2	0.000	1.11111	
		CG7483	0	0	-	14	12	1.167	1.14286	
		btz	7	8	0.875	10	32	0.313	0.75267	
	Outer shell	*Pnn*	2	0	-	21	4	5.250	1.17647	
		Aly	7	20	0.350	39	18	2.167	0.00533	**
		Acn	6	5	1.200	28	26	1.077	1.21212	
		Bin1	0	0	-	1	8	0.125	1.25000	
		RnpS1	0	1	0.000	185	79	2.342	0.92892	
	Transient factors	*Upf3*	4	6	0.667	19	24	0.792	1.29032	
		Upf2	7	32	0.219	36	32	1.125	0.00800	**
		SRm160	5	15	0.333	11	29	0.379	1.33333	
		Nxt1	0	0	-	4	2	2.000	1.37931	
		sbr	2	7	0.286	40	29	1.379	0.72600	
		Hel25E	0	6	0.000	0	8	0.000	1.42857	
NSL complex		*MBD-R2*	9	7	1.286	64	53	1.208	1.48148	
		mof	3	7	0.429	38	33	1.152	0.70364	
		Rcd5	4	10	0.400	9	29	0.310	1.45800	
		dgt1	6	14	0.429	9	12	0.750	1.09558	
		Rcd1	9	14	0.643	61	48	1.271	0.75911	
		nsl1	11	10	1.100	185	78	2.372	0.74000	
SUMO modifiers		*smt3*	0	1	0.000	0	5	0.000	1.53846	
		lwr	0	6	0.000	0	3	0.000	1.60000	
		Aos1	1	1	1.000	6	25	0.240	1.09857	
		Uba2	5	5	1.000	15	26	0.577	1.14400	
		Ulp1	11	6	1.833	328	88	3.727	0.74767	
COMPASS complex		*Set1* (not analysed)	-	-	-	-	-	-	-	
		wds	0	6	0.000	1	13	0.077	1.66667	
		Hcf (not analysed)	-	-	-	-	-	-	-	
		Rbbp5	0	6	0.000	3	9	0.333	1.14378	
		wdr82	0	9	0.000	0	16	0.000	1.73913	
		Cfp1	3	6	0.500	10	22	0.455	1.81818	
		Dpy-30L1	2	1	2.000	9	9	1.000	1.90476	
		ash2	1	13	0.077	9	21	0.429	0.75543	