A putative *de novo* evolved gene required for spermatid chromatin condensation in *Drosophila melanogaster*

Emily L. Rivard[1,4,*], Andrew G. Ludwig[1,*], Prajal H. Patel[1,*], Anna Grandchamp[2], Sarah E. Arnold[1,5], Alina Berger[2], Emilie M. Scott[1], Brendan J. Kelly[1,6], Grace C. Mascha[1], Erich Bornberg-Bauer[2,3], Geoffrey D. Findlay[1,**]

1. Department of Biology, College of the Holy Cross, Worcester, MA, USA
2. Institute for Evolution and Biodiversity, University of Münster, Münster, Germany
3. Department of Protein Evolution, Max Planck Institute for Developmental Biology, Tübingen, Germany
4. Current address: Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA
5. Current address: Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, MA, USA
6. Current address: Department of Entomology, The Ohio State University, Columbus, OH, USA

*equal contribution
**Correspondence to: gfindlay@holycross.edu

Abstract

Comparative genomics has enabled the identification of genes that potentially evolved *de novo* from non-coding sequences. Many such genes are expressed in male reproductive tissues, but their functions remain poorly understood. To address this, we conducted a functional genetic screen of over 40 putative *de novo* genes with testis-enriched expression in *Drosophila melanogaster* and identified one gene, *atlas,* required for male fertility. Detailed genetic and cytological analyses showed that *atlas* is required for proper chromatin condensation during the final stages of spermatogenesis. Atlas protein is expressed in spermatid nuclei and facilitates the transition from histone- to protamine-based chromatin packaging. Complementary evolutionary analyses revealed the complex evolutionary history of *atlas*. The protein-coding portion of the gene likely arose at the base of the *Drosophila* genus on the X chromosome but was unlikely to be essential, as it was then lost in several independent lineages. Within the last ~15 million years, however, the gene moved to an autosome, where it fused with a conserved non-coding RNA and evolved a non-redundant role in male fertility. Altogether, this study provides insight into the integration of novel genes into biological processes, the links between genomic innovation and functional evolution, and the genetic control of a fundamental developmental process, gametogenesis.

Author Summary

Genomes are in flux, as genes are constantly added and lost throughout evolution. New genes were once thought to arise almost exclusively via the modification or duplication of existing genes. Recently, however, interest has grown in alternative modes of new gene origination, such as *de novo* evolution from genetic material that previously did not encode proteins. Many *de novo* genes are expressed in male reproductive tissues, but their significance for fertility is not well understood. We screened likely *de novo* genes expressed in the *Drosophila* testis for reproductive roles and found one gene, *atlas*, essential for male fertility. We leveraged genetic and cell biological experiments to investigate roles for Atlas protein in reproduction and found that it is required during sperm development for proper packaging of DNA in the sperm nucleus. Evolutionary analyses of this gene revealed a complicated history, including loss in some lineages, movement between chromosomes, and fusion with a non-protein-coding gene. Studying both the functions and evolutionary histories of new proteins illustrates how they might evolve critical roles in biological processes despite their relative novelty. Furthermore, the study of *atlas* identifies an essential genetic player in the fly testis, an important model system for understanding how gametes are produced.

60 **Introduction**

61 The evolution of new genes is integral to the extensive genotypic and phenotypic
62 diversity observed across species. The best-characterized mechanism of novel gene
63 emergence is gene duplication [1, 2]; however, rapid expansion in high-quality genomic
64 resources has provided mounting evidence of lineage-specific sequences and the existence of
65 alternative modes of new gene origination. One such mechanism is *de novo* evolution, the birth
66 of new genes from previously non-genic or intronic regions, which is now a widely
67 acknowledged source of protein-coding and RNA genes [3-5]. Although *de novo* origination
68 was once considered an unlikely event, catalogs of *de novo* genes have now been published for
69 an expansive range of species [6-13]. Multiple models explain how protein-coding *de novo*
70 genes may acquire both an open reading frame (ORF) and regulatory sequences permitting
71 transcription [14-17]. Interrogation of the biochemical and biophysical properties of the proteins
72 encoded by *de novo* genes has offered initial insight into the mechanisms of emergence and
73 functional potential of these genes [17-20].

74 The capacity of protein-coding *de novo* genes to evolve important functions is a topic of
75 interest from evolutionary, physiological and molecular perspectives [21]. In the last couple of
76 decades, the products of *de novo* genes have been shown to play diverse roles in a variety of
77 organisms. For example, *de novo* genes function in fundamental molecular processes in yeast,
78 such as *BSC4*, a gene implicated in DNA repair, and *MDF1,* which mediates crosstalk between
79 reproduction and growth [22, 23]. *De novo* genes also evolve roles in organismal responses to
80 disease and changing environmental factors. A putatively *de novo* evolved gene in rice
81 regulates the plant's pathogen resistance response to strains causing bacterial blight [24].
82 Antifreeze glycoprotein genes, essential for survival in frigid ocean temperatures, evolved *de*
83 *novo* in the ancestor of Arctic codfishes to coincide with cooling oceans in the Northern
84 Hemisphere [25, 26]. *De novo* genes are additionally implicated in the development and
85 physiology of mammals. In house mice, a *de novo* evolved gene expressed in the oviduct

86 functions in female fertility by regulating pregnancy cycles [27]. A *de novo* gene found in
87 humans and chimpanzees regulates the oncogenesis and growth of neuroblastoma, revealing
88 the relevance of novel genes to human disease [28]. These studies have started to
89 demonstrate the significance of *de novo* genes, thereby challenging previous assumptions that
90 only ancient, highly conserved genes can be essential.

91 Across multicellular animals, male reproductive tissues serve as hubs for new gene
92 emergence via numerous mechanisms, including *de novo* evolution [12, 29-34]. Proposed
93 causes of this "out of the testis" phenomenon include the high level of promiscuous transcription
94 in testis cells [35, 36], the relative simplicity of promoter regions driving expression in the testis
95 [37], and preferential retention of novel genes with male-biased expression [38]. Sexual
96 selection also drives rapid evolution of reproductive proteins [39] and could drive the emergence
97 of new genes as a mechanism of improving male reproductive ability [33]. The testis-biased
98 expression of novel genes, combined with growing evidence for new genes acting across a
99 variety of tissue contexts, suggests that many novel genes may function in male reproduction.
100 For example, a pair of young duplicate genes in *Drosophila, apollo* and *artemis,* are essential for
101 male and female fertility, respectively [40]. Continued efforts to identify and characterize testis-
102 expressed novel genes are consequently critical for understanding the genetic basis of male
103 reproductive phenotypes.

104 *Drosophila* serves as an ideal system for interrogating the prevalence, sequence
105 attributes, expression patterns, and functions of testis-expressed *de novo* genes. The
106 availability of well-annotated genomes for numerous *Drosophila* species, the tractability of flies
107 to molecular genetics techniques, and our thorough understanding of *Drosophila* reproductive
108 processes facilitate comprehensive analyses of novel fly reproductive proteins [41, 42]. As
109 observed in other biological systems, *Drosophila de novo* genes retained by selection
110 demonstrate enriched expression in the testis [10, 19, 33, 34]. The expression patterns of
111 emerging *de novo* genes in the *Drosophila* testis were recently analyzed at single cell resolution

112 [43], thereby providing insight into the dynamics of novel gene expression throughout
113 spermatogenesis. In addition to bioinformatic screens that have started to identify *de novo*
114 genes and large-scale expression analyses of testis-expressed genes, RNAi [44] and
115 CRISPR/Cas9-based [45] functional screens have identified putative, testis-expressed *de novo*
116 genes required for fertility. However, a need remains for in-depth experimental and evolutionary
117 characterization of the genes identified in such screens. Detailed examination of the function of
118 *de novo* proteins will enable us to understand how these proteins might integrate into existing
119 gene networks and become essential.

120 We previously conducted a pilot functional screen of *de novo* genes with testis-enriched
121 expression in *D. melanogaster* and identified two novel genes, *goddard* and *saturn*, that are
122 required for full fertility [46]. *Goddard* knockdown males failed to produce any sperm. *Saturn*
123 knockdown males produced fewer sperm, which were inefficient at migrating to female sperm
124 storage organs. Subsequent characterization of Goddard using null deletion alleles and a
125 biochemically tagged rescue construct showed that the protein localizes to elongating sperm
126 axonemes and that, in its absence, individualization complexes associate less efficiently with
127 spermatid nuclei and do not successfully progress along sperm tails [20]. These data
128 suggested that putative *de novo* genes can evolve essential roles in a rapidly evolving
129 reproductive process, spermatogenesis.

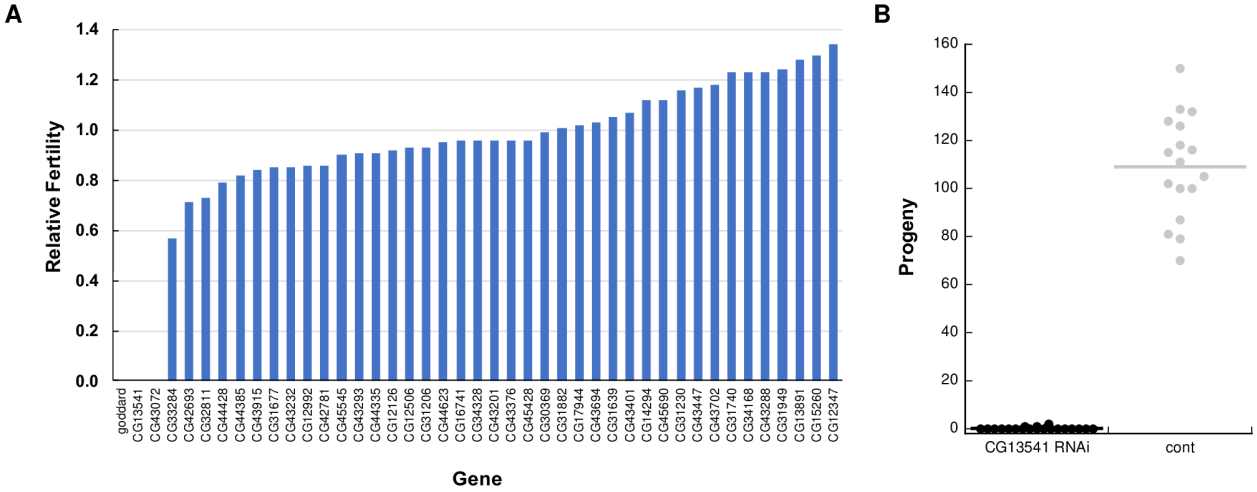
130 Here, we expanded this functional screen by evaluating whether any of 42 putative *de*
131 *novo* genes that show testis-enriched expression in *D. melanogaster* are required for male
132 fertility. We identified one gene, which we named *atlas,* whose knockdown or knockout results
133 in nearly complete male sterility. We found that *atlas* encodes a transition protein that facilitates
134 spermatid chromatin condensation. The *atlas* gene in *D. melanogaster* arose when a likely *de*
135 *novo* evolved protein-coding sequence moved off of the X chromosome and was inserted
136 upstream of a well-conserved non-coding RNA. While the *atlas* protein-coding sequence has
137 undergone multiple, independent gene loss events since its apparent origin at the base of the

138 *Drosophila* genus, the gene has evolved a critical function in *D. melanogaster*. These results
139 underscore the importance of detailed functional and evolutionary characterization in
140 understanding the origins of new protein-coding genes and the selective forces that affect their
141 subsequent evolution.
142

143 **Results**

144 *An RNAi screen identifies a putative* de novo *gene essential for* Drosophila *male fertility*

145 A previous pilot screen of 11 putative *de novo* evolved, testis-expressed genes identified
146 two genes that are critical for male fertility in *Drosophila melanogaster* [46]. This result, and
147 other recent work [e.g., 40, 47, 48], suggested that lineage-specific, newly evolved genes can
148 rapidly become important for fertility, perhaps by gaining interactions with existing protein
149 networks. To determine more comprehensively the frequency with which potential *de novo*
150 evolved genes become essential for fertility, we identified *de novo* or putative *de novo* evolved
151 genes with testis-biased expression. A previous computational analysis identified genes that
152 are detectable only within the *Drosophila* genus, lack identifiable protein domains, and show no
153 homology to other known proteins through BLASTP and TBLASTN searches [19]. We filtered
154 these genes to identify those expressed exclusively or predominantly in the testis, a common
155 site of *de novo* gene expression in animal species [27, 34, 38, 49]. This resulted in a set of 96
156 target genes.



157

158 **Figure 1. An RNAi screen of putative *de novo* genes identifies *CG13541* as a major contributor to**
159 ***Drosophila melanogaster* male fertility.** A) All RNAi lines that showed at least partial knockdown of the
160 target gene were screened in group fertility assays (see Materials and Methods). Relative fertility was
161 calculated by dividing the average number of progeny produced per female mated to knockdown males
162 by the average number of progeny produced per female mated to control males in a contemporaneous
163 experiment. Relative fertility measurements lack error bars because each gene was tested in only 1-2
164 replicates. Knockdown of *goddard* was used as a positive control. B) A single-mating, single-pair fertility
165 assay confirms the observed defect when males are knocked down for *CG13541*, as knockdown males

166 showed significantly reduced fertility (control fertility (mean ± SEM): 109.0 ± 5.3; knockdown fertility: 0.2 ±
167 0.1; two-sample *t*-test assuming unequal variances, $p = 5.6 \times 10^{-13}$).
168
169 We used testis-specific RNA interference to screen these genes for roles in male fertility.
170 We obtained RNAi lines from the Vienna *Drosophila* Resource Center (VDRC) and the
171 Transgenic RNAi Project (TRiP) and constructed additional lines using the TRiP-style
172 pValium20 vector [50], which induces efficient knockdown in the male germline. We tested an
173 RNAi line for each of 57 genes by using the *Bam*-GAL4 driver, which is expressed in the male
174 germline and which we enhanced with a copy of UAS-*Dicer2*. RT-PCR confirmed at least
175 partial knockdown in lines representing 42 genes (see example in Fig. S1). We then screened
176 knockdown males for fertility by allowing groups of 7 knockdown males to mate with 5 wild-type
177 females for 2 days. Progeny counts were standardized to the number of progeny produced by
178 concurrently mated groups of 7 control males and 5 wild-type females. The results are shown in
179 Fig. 1A. This initial screen identified *CG13541*, whose knockdown severely reduced male
180 fertility. We confirmed the result for *CG13541* by performing single-pair mating fertility assays
181 (Fig. 1B). Consistent with our previous convention of naming testis-expressed genes after
182 American rocketry [46], we will from here on refer to *CG13541* as *atlas*. While RNAi transgenes
183 designed to knockdown *CG43072* and *CG33284* caused full and consistently partial sterility,
184 respectively, we do not further consider these genes because subsequent gene knockout using
185 CRISPR genome editing indicated that neither gene plays a role in male fertility. In these
186 cases, the RNAi phenotypes might have been due to off-target knockdown.
187
188 *CRISPR-mediated gene mutation validates* atlas *RNAi results*
189 We validated the observed fertility defect by using CRISPR/Cas9-based genome editing
190 to construct putative loss-of-function alleles for *atlas* (Fig. S2). The principal allele we used for
191 validation and the functional studies described below was a null allele that completely deletes
192 the *atlas* genomic locus. This allele was generated by targeting each end of the locus with a

gRNA. We made three additional frameshift alleles by inducing double-stranded breaks at a gRNA target site just downstream of the *atlas* start codon, which induced non-homologous end joining. Males homozygous for the *atlas* deletion allele have the same fertility defect as knockdown males (Fig. 2A). Males homozygous for any of three frameshift alleles showed significantly reduced, but non-zero, fertility (Fig. 2B). It is possible that residual *atlas* function may be present in these animals, perhaps due to translation initiation at a downstream start codon, which could generate a shorter protein with partial function. Each frameshift allele retains the possibility of encoding an N-terminally truncated, but otherwise in-frame, version of Atlas protein that would lack the first 60 amino acids (out of 172; see Fig. S2). Alternatively, it is possible that the residual fertility of the frameshift alleles is caused by the gene's intact 3' UTR, a topic we discuss in more detail below. Finally, we constructed a genomic rescue construct carrying both the full *atlas* transcribed region and its native regulatory sequences. *atlas* null males that carried a single copy of the rescue construct had fully restored fertility (Fig. S3). Overall, these data demonstrate that *atlas* loss, and not an RNAi or CRISPR off-target, causes nearly complete male sterility.

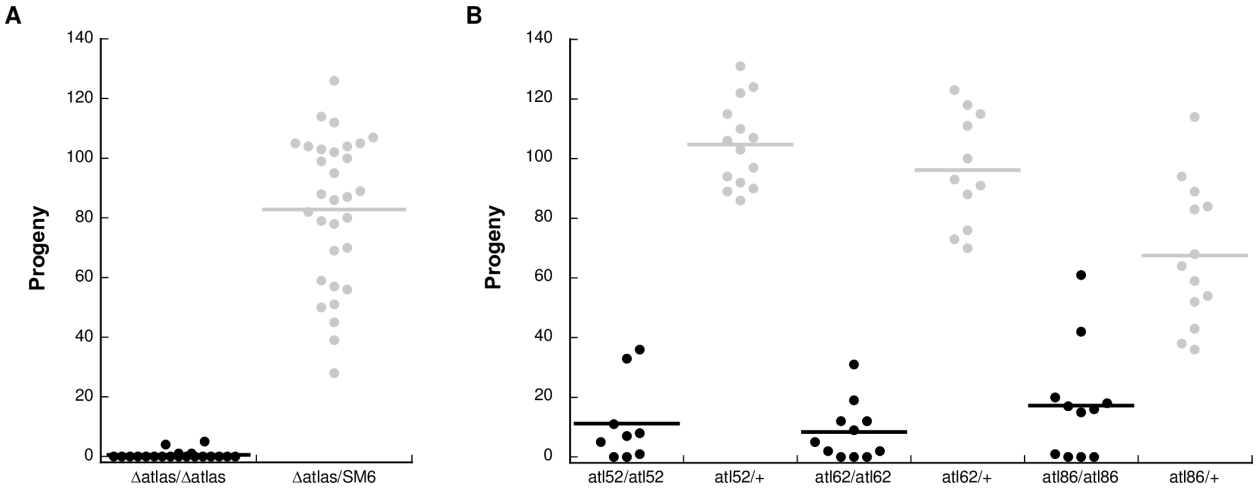


Figure 2. CRISPR-generated deletion and frameshift alleles of *atlas* confirm the gene's requirement for male fertility. A) Single-pair fertility assay for males homozygous for the null (Δ*atlas*) allele or heterozygous controls (Δ*atlas*/SM6). Flies homozygous for the deletion had significantly reduced fertility (control fertility: 82.9 ± 4.5; null fertility: 0.3 ± 0.6; two-sample t-test assuming unequal variances, $p = 5.4 \times 10^{-18}$). B) Single-pair fertility assays for males homozygous or heterozygous for three frameshift alleles of *atlas* generated by imprecise non-homologous end joining at a CRISPR/Cas9 target site just

215 downstream of the start codon: *atlas*^{52} (control fertility: 104.7 ± 3.8, mutant fertility: 11.2 ± 4.6; two-sample
216 *t*-test assuming unequal variances: $p = 8 \times 10^{-12}$), *atlas*^{62} (control fertility: 96.2 ± 5.7; mutant fertility: 8.4 ±
217 2.9; two-sample *t*-test assuming unequal variances: $p = 6.1 \times 10^{-10}$) and *atlas*^{86} (control fertility: 67.5 ± 6.6;
218 mutant fertility: 17.3 ± 5.8; two-sample *t*-test assuming unequal variances: $p = 9.5 \times 10^{-6}$).
219
220
221 Atlas *is required for proper spermatid nuclear condensation*

222 We next examined how *atlas* loss of function impacted male fertility at the cellular level.
223 Dissection and phase-contrast imaging of *atlas* deletion null or knockdown male reproductive
224 tracts revealed that while the pre-meiotic and meiotic stages of spermatogenesis appeared
225 normal, sperm accumulated at the basal end of the testes, rather than in the seminal vesicles
226 (SVs), over the first week of adulthood (Fig. 3A and Fig. S4). To further characterize the fertility
227 defects in the absence of *atlas*, we examined the Mst35Bb-GFP ("protamine"-GFP) marker in
228 null or knockdown backgrounds [51]. Mst35Bb encodes one of two protamine-like proteins
229 (highly similar paralogs of each other) that bind DNA in mature sperm. Its GFP fusion construct
230 thus allows visualization of nuclei in late stage spermiogenesis and mature sperm. Consistent
231 with the observed conglomeration of sperm tails at the basal testes, SVs from either *atlas* null or
232 knockdown males contained fewer mature sperm (Fig. 3B, Fig. S4C). The nuclei of sperm from
233 null males also appeared wider and less elongated than those of controls. Together, these data
234 suggest that *atlas* is required after meiosis, as developing spermatids take on their final
235 structures.

236 We next examined two post-meiotic processes: individualization of 64-cell spermatid
237 cysts into mature sperm, and spermatid nuclear condensation. Individualization initiates when
238 an actin-rich individualization complex (IC) associates with the bundle of spermatid nuclei. The
239 IC then proceeds down the sperm tails, expelling cytoplasmic waste and remodeling cell
240 membranes to form 64 individual sperm. We visualized this process in males 0-1 days old,
241 when spermatogenesis occurs at high levels, by staining whole mount testes for actin (Fig. 3C-
242 D). Although ICs associated with nuclear bundles present at the basal end of the testes in both
243 control and *atlas* null males, we observed significantly fewer nuclear bundle-associated ICs in

244 nulls (Fig. 3C-D). While control testes typically had several ICs progressing down sperm tails,
245 we saw a significantly reduced proportion of progressed bundles in nulls (Fig. 3C-D). In some
246 null testes, we also observed individual investment cones dissociated from progressing ICs (Fig.
247 3C).

248 The ability of ICs to assemble at nuclear bundles and progress down sperm tails may be
249 reduced if nuclear condensation is aberrant [reviewed in ref. 52]. During *Drosophila*
250 spermiogenesis, round spermatid nuclei undergo a series of stepwise, morphological changes
251 that are the product of two distinct, but related processes: changes in the chromatin packaging
252 of DNA, and changes in nuclear shape [53-55]. The end result is thin, condensed nuclei. We
253 quantified this process in testes dissected from newly eclosed wild-type and *atlas* null males
254 expressing Mst35Bb-GFP, which marks the final stages of condensation. We shredded the
255 post-meiotic region of the testes in the presence of a fixative and counted the number of nuclear
256 bundles that exhibited each of five stages of condensation [53]: round nuclei, early canoe-stage
257 (unmarked with Mst35Bb-GFP), late canoe-stage (marked with Mst35Bb-GFP), elongated
258 nuclei, and fully condensed nuclei (Fig. 4). Condensation of the nuclear bundles in *atlas* null
259 testes progressed at similar rates to controls through the late canoe stage (Table S1).
260 However, in *atlas* null males, all nuclear bundles that progressed past the canoe stage (which
261 included ~60% [range: 26-100%] of all observed bundles) showed an aberrant “curled”
262 phenotype (Fig. 4; Table S1). These data suggest that Atlas protein is required during the later
263 stages of nuclear condensation and are consistent with the idea that the loss of *atlas* affects
264 nuclear condensation in a way that reduces IC assembly and sperm individualization (see Fig.
265 3C).

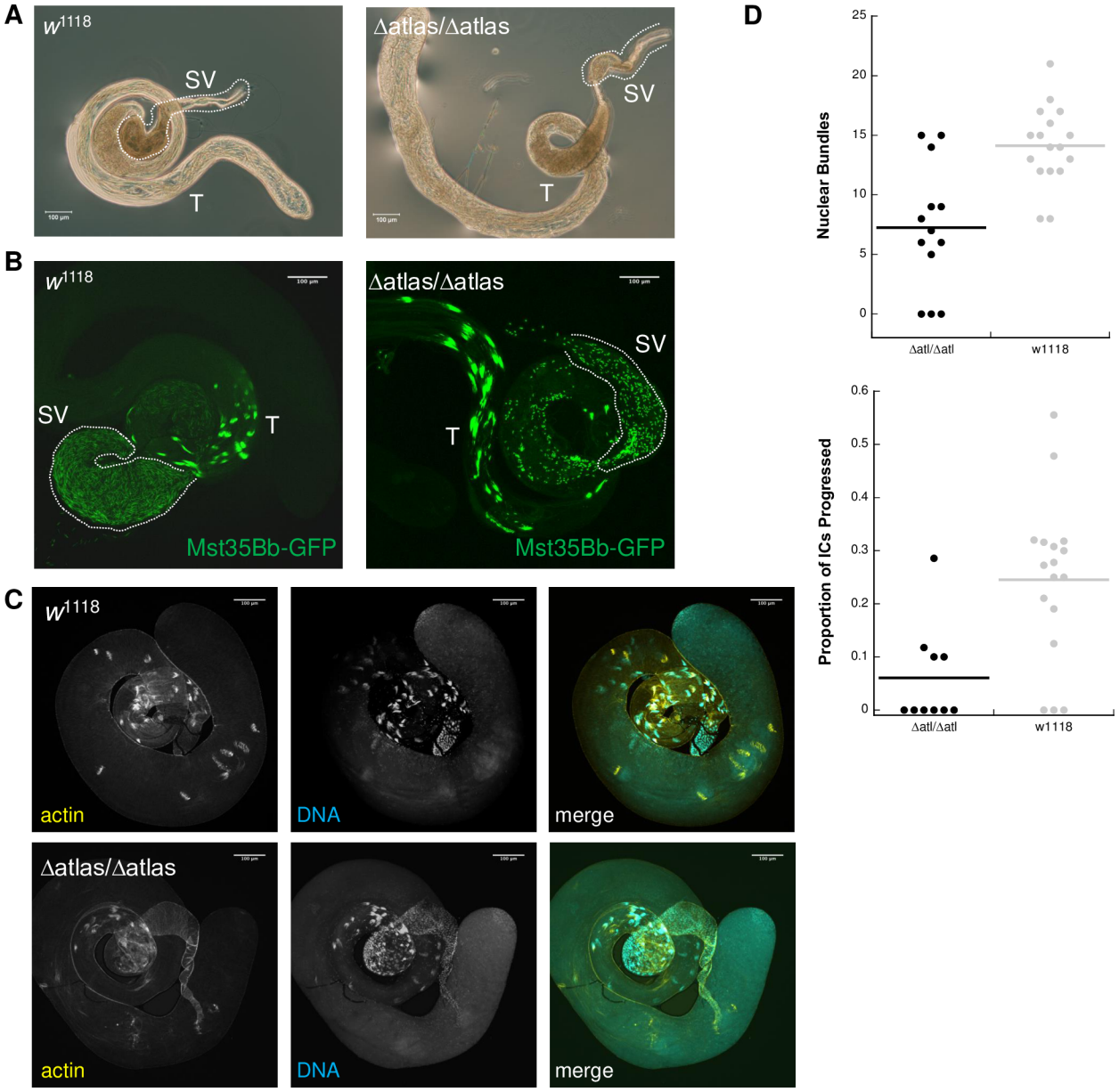


Figure 3. Cytological investigations of the *atlas* mutant fertility defect. A) Phase contrast microscopy of male reproductive tracts dissected from 7-day-old, unmated control (w^{1118}) or *atlas* null males. Control males show the expected accumulation of sperm tails in the seminal vesicle (SV), which appear here as a darker brown shading, while null male have an aberrant accumulation of sperm tails at the basal end of the testis (T). B) Visualization of Mst35Bb-GFP in 4-day old control and *atlas* null testes. While Mst35Bb localizes to spermatid nuclei in the absence of *atlas*, the nuclei appear shorter and much less numerous in the outlined SV. C) Representative images from phalloidin staining of w^{1118} and *atlas* null testes used to assess the association of individualization complexes (ICs) with nuclear bundles and the progression of ICs down the length of sperm tails. D) At top, number of nuclear bundles with ICs associated in control and *atlas* null testes. Significantly more ICs were observed in control testes (control: $N = 17$, median = 14; mutant: $N = 13$, median = 7; Wilcoxon rank-sum test W = 34, $p = 0.0014$). At bottom, proportion of all observed ICs that were intact and that had progressed away from nuclear bundles. Three mutant testes with no observed ICs were excluded from the analysis. A significantly higher proportion of ICs progressed in control testes (control: $N = 17$, median = 0.27; control: $N = 10$, median = 0; Wilcoxon rank-sum test W = 28, $p = 0.0038$).

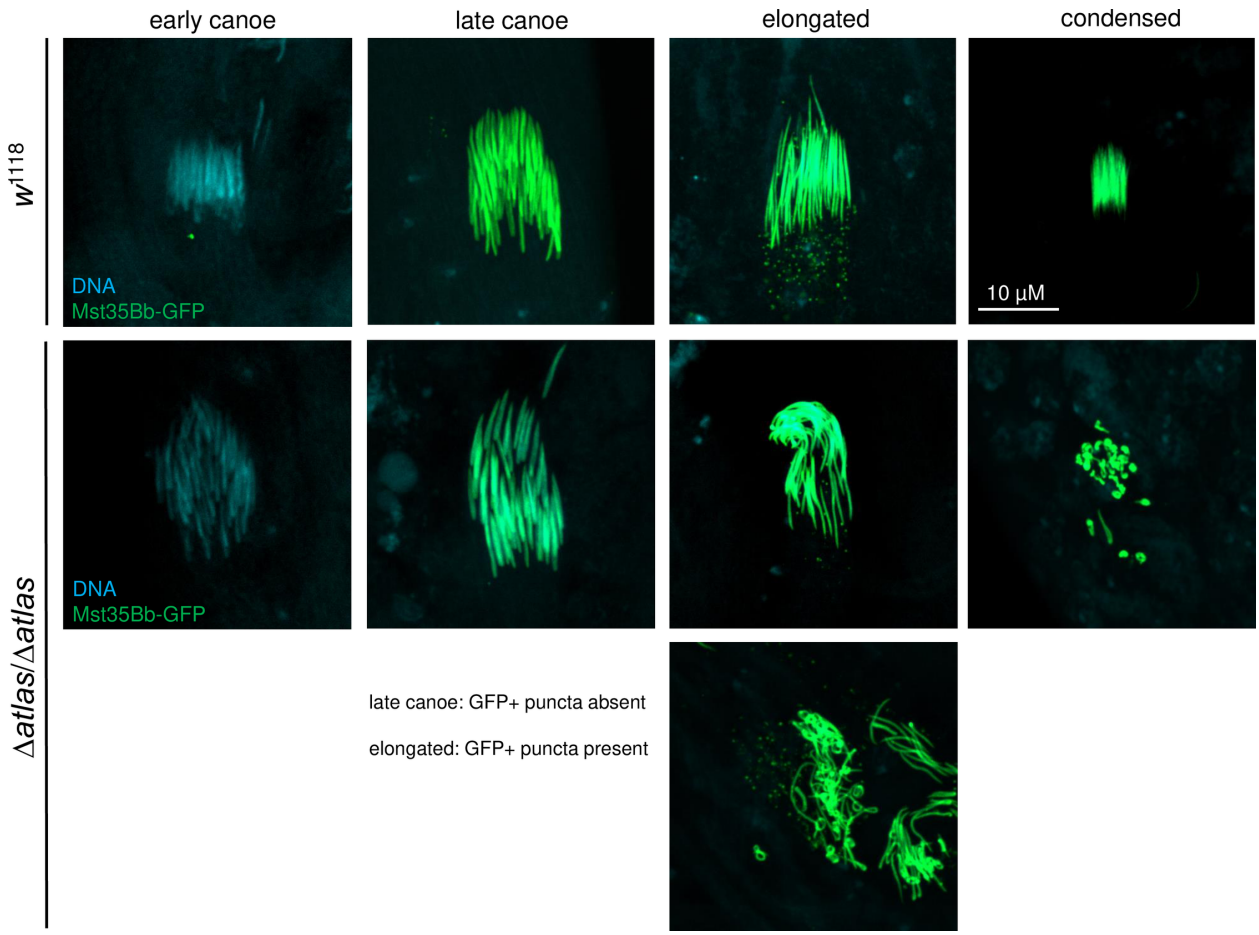


Figure 4. *Atlas* null males show aberrant nuclear shaping at and beyond the elongated stage of spermatid nuclear condensation. Early and late canoe stages were distinguished by the absence or presence of Mst35Bb-GFP, respectively. Late canoe and elongated stages were distinguished by the absence or presence of GFP-positive puncta, respectively. Condensed nuclei were distinguished from elongated nuclei by size. As shown in Table S1, nuclear bundles from *atlas* null testes consistently took on a curved shape after the canoe stage, though the degree of curvature was variable, as exemplified by the two examples of elongated nuclei from *atlas* null testes above.

That condensing spermatid nuclei are misshapen in the absence of *atlas* suggests the possibility that Atlas protein is critical for nuclear condensation. This idea is further supported by its predicted biochemical properties. Previously characterized spermatid chromatin binding proteins are small and highly basic [53, 56, 57], as the excess of positively charged amino acid side chains facilitates ionic interactions with negatively charged DNA. Many such proteins (i.e., Tpl94D, Mst35Ba, Mst35Bb, Prtl99C and Mst77F) also contain a conserved protein domain, the high-mobility-group box (HMG-box) domain [55, 58-64], suggesting that this type of chromatin

298 binding protein could have originated through gene duplication and divergence. Consistent with
299 its putative *de novo* origin, Atlas lacks a detectable HMG-box domain. However, Atlas is
300 otherwise similar to these other sperm chromatin binding proteins: the ~20 kDa protein has a
301 highly basic predicted isoelectric point of 10.7, and its primary sequence contains the sequence
302 KRDK, which matches the canonical consensus sequence for nuclear import, K(K/R)X(K/R)
303 [65]. To test the hypothesis that Atlas is nuclear localized, and could thus bind DNA, we
304 generated an *atlas-GFP* transgene under UAS control and expressed it ubiquitously using
305 *tubulin*-GAL4 and in the early male germline using *Bam*-GAL4. In both larval salivary glands
306 and early male germline cells, Atlas-GFP appeared to be nuclear localized (Fig. S5).
307 While these results were consistent with Atlas protein localizing to the nucleus, they did
308 not allow us to visualize Atlas in the cells in which it is normally expressed. To do so, we used
309 CRISPR/Cas9-induced homology directed repair (https://flycrispr.org/scarless-gene-editing/)
310 [66-68] to create an *atlas*-GFP fusion at the endogenous *atlas* locus (see Fig. S6 and Materials
311 and Methods). We first confirmed the functionality of the knock-in allele by showing that males
312 with the *atlas* locus genotype *atlas*-GFP/Δ*atlas* had equivalent fertility to males of genotype
313 *atlas*+/Δ*atlas* (Fig. 5A). We then visualized Atlas-GFP fusion protein in whole-mount testes in
314 conjunction with phalloidin-stained actin (Fig. 5C and Fig. S7A). Atlas-GFP was absent from
315 seminal vesicles, consistent with its absence from the proteome of mature *D. melanogaster*
316 sperm [69, 70]. Instead, Atlas-GFP colocalized with condensing nuclear bundles near the basal
317 end of the testes (Fig. 5C). Actin-based ICs were also observed in the basal testes, but
318 generally did not co-localize with Atlas-GFP, suggesting that Atlas-GFP is present in
319 condensing nuclei before IC association (Fig. 5C). This result, taken together with the aberrant
320 nuclear condensation in the absence of *atlas* (Fig. 4), is consistent with the idea that Atlas is a
321 transition protein. Transition proteins are chromatin components that act transiently during
322 spermatid nuclear condensation. A series of transition proteins first replace histones as the

323 primary DNA binding proteins in the nucleus and then give way to protamines, the proteins that
324 package chromatin in mature sperm [53, 55, 58].
325 To further elucidate the role of *atlas* in nuclear condensation, we next examined Atlas-
326 GFP localization in the presence of either an early spermatid nuclear marker, histone H2Av-
327 RFP [55, 71], or Mst35Bb-dsRed [51], a marker of nuclei from the late canoe stage through final
328 condensation. Atlas-GFP showed no co-localization with H2Av-RFP, suggesting that Atlas
329 functions after histone removal (Fig. 5B). In contrast, some GFP-positive bundles co-localized
330 with Mst35Bb-dsRed, but others did not (Fig. 5D and Fig. S7B). These data suggest that Atlas
331 may be one of the final transition proteins used in nuclear condensation before the chromatin
332 becomes fully condensed with protamines.
333

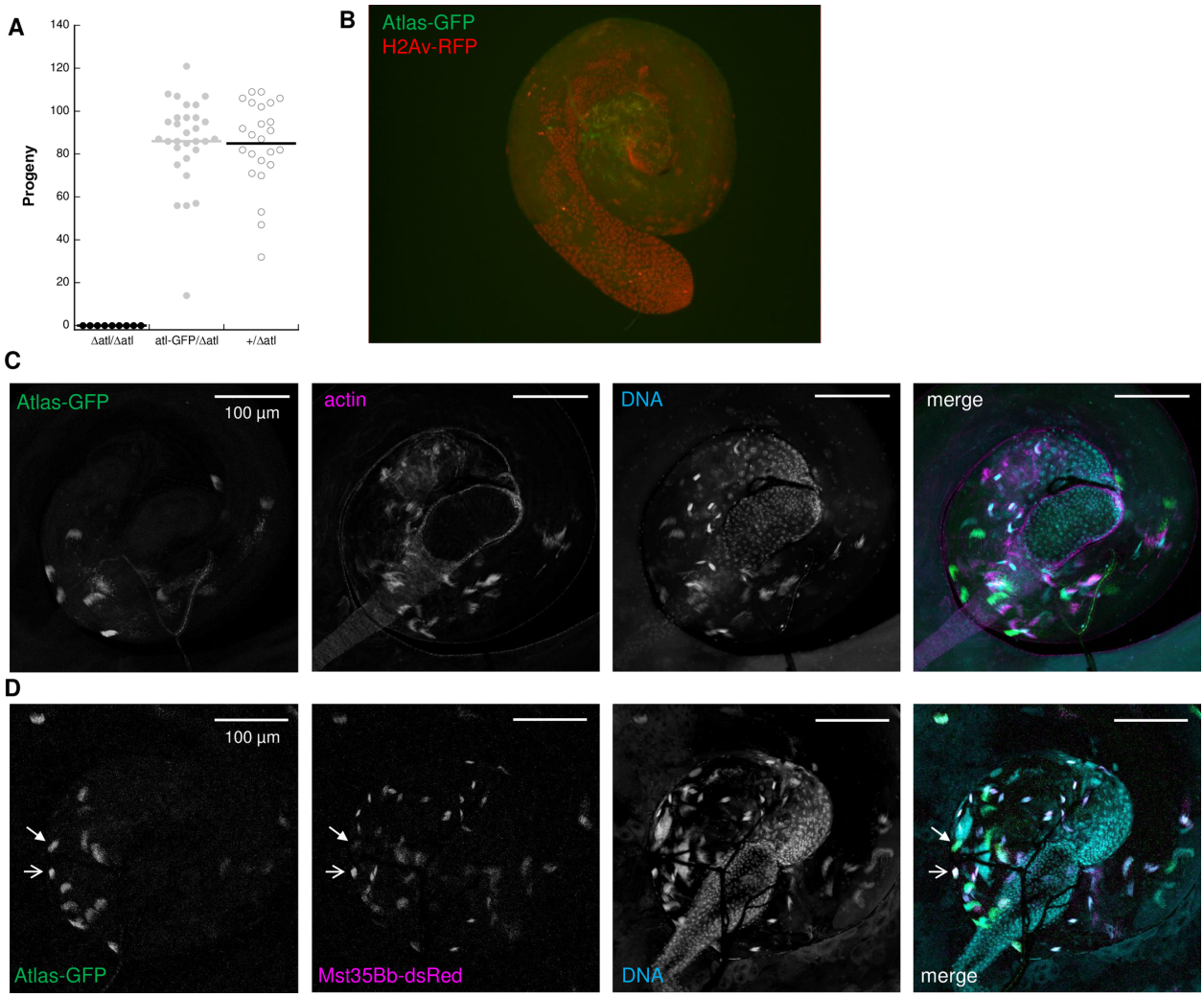


Figure 5. An *atlas*-GFP allele generated at the endogenous *atlas* locus is fully functional for male fertility and encodes a protein that localizes to condensing spermatid nuclei. A) A single copy of the *atlas*-GFP allele is sufficient for full fertility when paired with the Δ*atlas* allele, as compared to males heterozygous for the wild-type *atlas* allele (Δ*atlas*/*atlas*-GFP fertility: 86.0 ± 3.2; Δ*atlas*/+ fertility: 84.9 ± 4.1; two-sample t-test assuming unequal variances, $p = 0.85$). B) Atlas-GFP does not co-localize with histone H2Av-RFP, a marker of the initial stages of spermatid nuclear condensation. C) Visualization of Atlas-GFP in the basal portion of whole-mount testes from *atlas-GFP* homozygotes shows that the fusion protein co-localizes with a subset of condensing spermatid nuclear bundles. While actin associates with fully condensed nuclei at the basal testis, Atlas-GFP does not overlap and is also absent from the seminal vesicle. D) Atlas-GFP partially colocalizes with Mst35Bb-dsRed, a marker of the final stage of nuclear condensation, in the basal portion of whole-mount testes. Open arrow: example of co-localization. Filled arrowhead: example of Atlas-GFP that does not co-localize with Mst35Bb-dsRed. Collectively, these data suggest that atlas may serve as a transition protein involved in the final stages of nuclear condensation. The whole testes from which the basal portions are shown in panels C and D are shown in Fig. S7.

To determine the stage(s) of nuclear condensation at which *atlas* functions, we analyzed the shape of fixed nuclear bundles from shredded testes isolated from *atlas*-GFP males on the

352 day of eclosion. Based on the stage of the defect in *atlas* null males (Fig. 3-4) and the pattern
353 of Atlas-GFP-positive bundles in whole-mount testes (Fig. 5), we hypothesized that Atlas-GFP
354 would localize to the later stages of nuclear condensation. Consistent with this hypothesis, we
355 did not detect Atlas-GFP in round or early canoe stage bundles (Fig. 6A-B). Atlas-GFP co-
356 localized with DNA in late canoe stage bundles (Fig. 6C). Interestingly, when nuclei elongated
357 further, GFP was detected not in the nucleus, but as puncta basal to the nuclei (Fig. 6D; see
358 also Fig. S7A). Since Atlas-GFP is not observed in mature sperm in the SV (Fig. 5C), these
359 data suggest that Atlas may function as a transition protein that facilitates the condensation of
360 spermatid nuclei from histone-based DNA packaging to protamine-like-based DNA packaging
361 [53] and is then removed from nuclei once protamines bind DNA. Indeed, the appearance of
362 Atlas in nuclei during the late canoe stage of condensation is similar to the pattern observed for
363 a previously characterized transition protein, Tpl94D [53]. We hypothesize that the failure of
364 *atlas* null sperm to form needle-like nuclei can be explained by the absence of Atlas from the
365 late canoe nucleus. It is also possible that the apparent removal of Atlas-GFP from nuclei (Fig.
366 6D) represents a mechanism for removing transition proteins from the nucleus after they exert
367 their functions. We observed above that some Mst35Bb-GFP also appears to be removed from
368 the nucleus in puncta during the elongation stage of nuclear condensation (see elongated stage
369 of control nuclear bundles, Fig. 4), even though other Mst35Bb-GFP molecules ultimately
370 package DNA in mature, individualized sperm. This could occur if Mst35Bb-GFP is present in
371 excess of what is needed to package DNA.
372

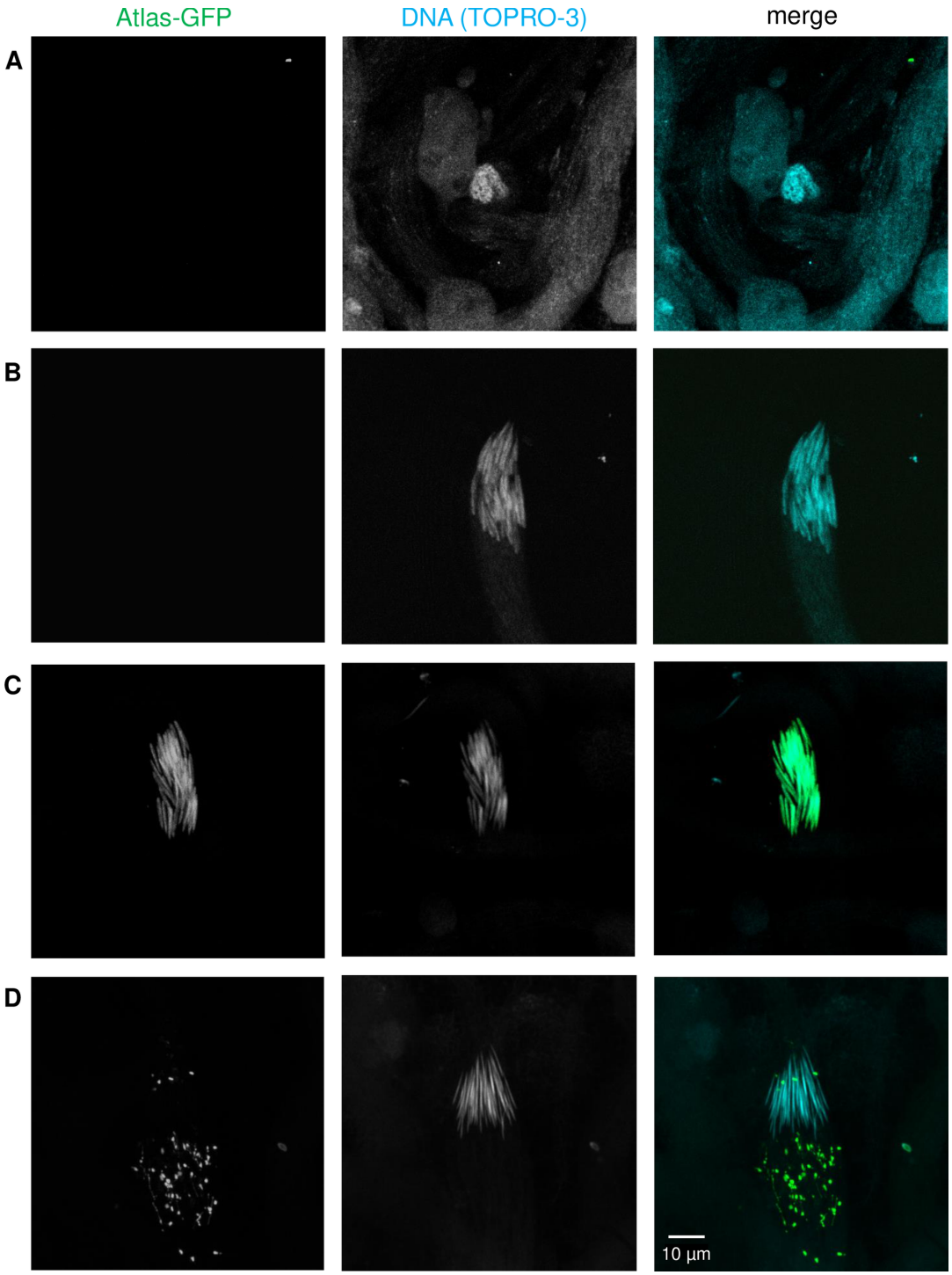
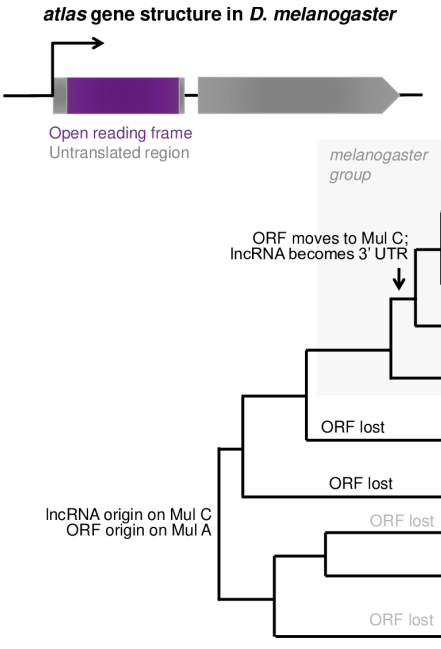


Figure 6. Atlas-GFP is present in late canoe-stage spermatid nuclei and then appears to leave the nucleus in puncta. Staging of condensing spermatid nuclei fixed in paraformaldehyde from *atlas*-GFP males stained with TO-PRO-3 DNA stain. Atlas-GFP is not detectable in (A) round stage or (B) early canoe stage nuclei. Atlas-GFP is nuclear localized in the late canoe stage (C). When nuclei become fully elongated (D), puncta of Atlas-GFP appear to be removed from the nucleus.

381 *Evolutionary origins of* atlas

382 To better understand the evolutionary origin of *atlas* and its evolution since emergence,
383 we used a combination of BLAST- and synteny-based approaches to identify *atlas* orthologs
384 throughout the genus [46, 72]. One notable feature of this two-exon gene is that the protein-
385 coding region (519 nucleotides) is contained entirely within the first exon (622 nt); the longer,
386 second exon (910 nt) appears to be entirely non-coding (Fig. 7). Surprisingly, the second exon
387 is more widely conserved. BLASTN detected significant matches to this region (range of hit
388 length: 185-864 nt) in the same genomic location on Muller element C [see ref. 73 for
389 explanation of Muller elements], as assessed by synteny, in all *Drosophila* species examined,
390 including distantly related species such as *D. virilis* and *D. grimshawi* (Fig. 7 and S8). The
391 protein-coding first exon shows a more limited phylogenetic distribution. In most members of
392 the *melanogaster* group of *Drosophila* (gray box in Fig. 7), this exon is found in a conserved
393 position, adjacent to the non-coding region on the equivalent of *D. melanogaster* chromosome
394 2R (Table S2). In *D. ananassae*, however, the protein-coding region is found on the X
395 chromosome (Muller element A). A putative ortholog for the protein-coding sequence is
396 detectable by BLASTP in *D. virilis* in a partially syntenic region on the same Muller A element
397 (Table S2, Fig. S9). These data suggest that the *atlas* protein-coding sequence initially arose
398 on Muller element A and then moved to Muller element C, giving rise to the gene structure
399 observed in extant *D. melanogaster* and its sister species.
400



	Predicted ORF			3' UTR Sequence		
	Location		Male expressed?	Location		Male expressed?
	Chr. or Scaffold	Muller Element		Chr. or Scaffold	Muller Element	
D. simulans	2R	C	Yes	2R	C	Yes
D. sechellia	2R	C	Yes	2R	C	Yes
D. melanogaster	2R	C	Yes*	2R	C	Yes*
D. yakuba	2R	C/B	Yes*	2R	C/B	No
D. erecta	2R	C/B	Yes	2R	C/B	No
D. ananassae	X	A	Yes*	2R	C	No
D. pseudoobscura	—	—	—	3	C	Yes*
D. persimilis	—	—	—	s2	C	Yes
D. willistoni	—	—	—	s2	C	No
D. mojavensis	—	—	—	s6496	C	Yes
D. virilis	X	A	Yes*	s12875	C	No/Low
D. grimshawi	—	—	—	s15112	C	No

Figure 7. Molecular evolution and gene expression of *atlas* across the *Drosophila* genus. The gene structure of *atlas* in *D. melanogaster* is shown at top left. The predicted protein-coding sequence is contained entirely within exon 1, while exon 2 encodes the presumed 3' UTR. The gene is located on chromosome 2R, equivalent to Muller element C. The phylogeny shows BLAST- and synteny-based detection of sequences orthologous to the protein-coding sequence and the 3' UTR sequence across *Drosophila* species. Sex-specific adult RNA-seq data were used to assess male expression across species, with RT-PCR verification performed in species marked with asterisks. RNA-seq data for the syntenic region of the 3' UTR in *D. virilis* were ambiguous; see Figs. S8 and S10.

To confirm the lack of *atlas* protein-coding sequences identifiable by BLASTP or TBLASTN in most non-*melanogaster* group species, we identified the regions syntenic to those containing *atlas* in *D. ananassae* and *D. virilis* in 11 additional *Drosophila* species and used more sensitive methods to search for potential orthologs [72]. Specifically, we: a) relaxed the BLAST cut-offs for detection, since default parameters can cause false-negative results when searching for potential *de novo* genes in divergent species [74]; b) used adult male RNA-seq data to detect transcribed areas within each syntenic region that did not match annotated genes; and, c) predicted the isoelectric point of the potential proteins encoded, under the hypothesis that Atlas orthologs would have conserved, DNA-binding functions. The results are summarized in Fig. S9 and Table S2. These searches detected no evidence for *atlas* orthologs in the following *Drosophila* species: *obscura*, *miranda, willistoni, hydei, arizonae, mojavensis,*

422 *navojoa,* and *grimshawi*. Some of these species have unannotated, male-expressed transcripts
423 in the regions syntenic to the Muller A location of *atlas* in *D. ananassae* and *D. virilis*, but when
424 each was compared with BLASTP to the *D. melanogaster* proteome, all matched proteins other
425 than Atlas, suggesting they may be lineage-specific paralogs of other genes (Fig. S9). In sister
426 species *D. pseudoobscura* and *D. persimilis*, we detected a male-expressed transcript predicted
427 to encode a protein with a pI > 10 in the region syntenic to the location of *atlas* in *D. virilis*, but
428 the predicted protein sequences showed no significant BLASTP similarity to *atlas* orthologs
429 (Table S2). While this predicted protein may represent a divergent *atlas* ortholog, the abSENSE
430 method predicts low probabilities of BLASTP detection failure when searching for Atlas protein
431 in these species (0.02 and 0.04, respectively), so we favor the hypothesis of a lineage-specific,
432 newly evolved gene in the region. Conversely, in *D. busckii*, we detected, in the region syntenic
433 to the *D. virilis atlas* locus, a male-expressed transcript predicted to encode a protein with
434 significant BLASTP identity to *D. melanogaster* Atlas (e = 4e-8), but with a predicted pI of 5.1
435 and a ~50 percent shorter open-reading frame (Table S2). The ortholog status of this predicted
436 protein is also unclear, but because of its dramatically altered size and pI, it is unlikely to have a
437 functional role equivalent to that of *D. melanogaster* Atlas.

438 To investigate whether the protein-coding region may have reproductive functions in
439 other species, we used sex-specific RNA-seq data from numerous *Drosophila* species curated
440 by the Genomics Education Partnership [72; thegep.org] and verified several of these results by
441 RT-PCR (Fig. 7, Fig. S10). In all species in which *atlas* was detected, the protein-coding region
442 is expressed specifically in males regardless of its genomic location (Fig. 7). Interestingly, the
443 non-coding region shows male-specific expression in species lacking an unambiguous,
444 orthologous coding region, such as *D. pseudoobscura* and *D. mojavensis*. Conversely, while *D.*
445 *yakuba* and *D. erecta* express the protein-coding region robustly, we found no RNA-seq
446 evidence to support expression of the non-coding second exon, in spite of its sequence
447 conservation (Fig. S8). Based on its high level of sequence conservation, consistent genomic

448 location and expression in a variety of species, it is possible that what we now consider to be
449 the 3' untranslated region of *atlas* from *D. melanogaster* was, ancestrally, a non-coding RNA.
450 The FlyBase database reports two transcript isoforms of *atlas* in *D. melanogaster*: the
451 *atlas*-RA isoform is 986 nucleotides, while the *atlas*-RB isoform is 1528 nt. These isoforms
452 differ in how much of the second, non-coding exon is included in the transcript. We used RT-
453 PCR of whole male cDNA to assess the presence of these isoforms and their relative
454 abundances. Primers designed to amplify a region present in both isoforms produced products
455 that appeared more abundant than primers designed to amplify only the long isoform, even
456 though both primer pairs appeared to amplify genomic DNA with equal efficiency. Based on
457 RT-PCR band intensities and controlling for product size and genomic PCR band intensities, we
458 estimated that the short isoform is about 3-fold more abundant. This difference in abundance is
459 mirrored in available RNA-seq data, which show approximately 3- to 4-fold higher levels of
460 expression in the upstream part of exon 2 (Fig. S8), a pattern that also appears in *D. simulans*
461 and *D. sechellia*. Evaluating the potential significance of this finding awaits functional
462 characterization of the non-coding region.
463 As we have observed for other putative *de novo* genes with essential male reproductive
464 functions [46], the pattern of *atlas* protein-coding sequence presence/absence across the
465 phylogeny is difficult to explain parsimoniously. If we assume that gene birth events are less
466 frequent than gene deaths, since the latter can occur through many possible mutational events
467 and can happen separately along multiple phylogenetic lineages, our data support the
468 hypothesis of a single origin of the protein-coding sequence at the base of the genus, followed
469 by independent loss events on the lineages leading to *D. grimshawi*, *D. mojavensis* and *D.*
470 *willistoni*, and potentially also *D. pseudoobscura*/*persimilis* (Table S2). We summarize these
471 findings for 12 representative species of *Drosophila* in Fig. 7. The general patterns of loss do
472 not change when all species of Table S2 are considered, though an additional loss in the
473 *melanogaster* group is likely due to the absence of a detectable ortholog in *D. kikkiwai* and *D.*

474 *serrata*. As noted above, the pattern of gene loss can also appear due to orthology detection
475 failure [74], for which we tried to account with our additional search methods described above.
476 We also note, however, that the probability of BLASTP-based ortholog detection failure is
477 relatively low for some *Drosophila* species that lack *atlas*, including *D. pseudoobscura*
478 (probability of non-detection due to divergence = 0.02), *D. persimilis* (p = 0.04) and *D. willistoni*
479 (p = 0.06). The probability is higher for other species, *D. mojavensis* (p = 0.33) and *D.*
480 *grimshawi* (p = 0.66), underscoring the importance of our additional search strategies. Overall,
481 our data support the hypothesis of multiple, independent loss events within *Drosophila*.
482 AbSENSE produces a 1.00 probability of BLASTP-based Atlas ortholog detection failure
483 outside of *Drosophila*, reflecting the protein's short length and relatively rapid divergence (see
484 below). Indeed, the protein-coding and non-coding transcriptomes from each species showed
485 no matches to Atlas protein or cDNA sequences by BLAST. We thus used another synteny-
486 based approach, summarized in Fig. S11, to look for the protein-coding gene in other Dipterans
487 with well-resolved genomes: *Musca domestica*, *Glossina morsitans*, *Lucilia cuprina*, *Aedes*
488 *aegypti*, *Anopheles darlingi*, *Anopheles gambiae*, *Culex quinquefasciatus* and *Mayetiola*
489 *destructor*. In none of these species was a putative homolog found in any potential syntenic
490 region.
491 Recognizing the limitation of even this approach, we also used HMMER [75] to search
492 iteratively either all genomes in ENSEMBL, or all metazoan genomes in ENSEMBL, for
493 annotated proteins with identity to Atlas from *D. melanogaster* or *D. virilis*. These searches
494 initially identified significant hits to the Atlas orthologs we identified above from other *Drosophila*
495 species. When these collections of orthologs were used as queries, no further proteins outside
496 of *Drosophila* were a significant match. As a control, we performed the same search strategy
497 with *D. melanogaster* Mst35Bb, a protein whose length, amino acid composition, and function
498 are similar to Atlas. These searches readily identified orthologs throughout Diptera, consistent
499 with predictions of its conservation from the OrthoDB database [76]. We also note that a

500 previous HMMER-based analysis to identify HMG-box-containing spermatid chromatin
501 condensation proteins in *Drosophila* and other related insects did not detect *atlas* as a member
502 of the gene family [64], further diminishing the possibility that *atlas* is a divergent paralog of
503 other transition and protamine-like proteins. Thus, we conclude that *atlas* is a putative *de novo*
504 evolved gene that is limited phylogenetically to the *Drosophila* genus.

505 Finally, we used standard tests of molecular evolution to examine the selective
506 pressures that have shaped Atlas protein within the *melanogaster* group. We aligned the *atlas*
507 protein-coding sequences from 12 species and used PAML to ask whether a model (M8)
508 allowing for positive selection, as well as neutral evolution and purifying selection, explained the
509 data better than models (M7 and M8a) that allowed only neutral evolution and purifying
510 selection [77, 78]. These data showed that while the *atlas* protein-coding sequence's rate of
511 evolution was accelerated relative to most *Drosophila* proteins (whole-gene estimated d_N/d_S, ω =
512 0.41 by PAML model M0), there was no significant evidence for positive selection acting to
513 recurrently diversify a subset of sites within the protein (Table 2).

514

515 **Table 2. PAML sites tests for positive selection acting on *atlas* in the *melanogaster***
516 **group.**
517

PAML Model	ω estimate	ln L	Number of parameters	Likelihood ratio test with M8
M0 (uniform ω across all sites)	0.41	-4032.24	23	n/a
M7 (10 site classes, each with $0 \leq \omega \leq 1$)	n/a	-3961.28	24	χ^2 = 1.74, df = 2, p = 0.42
M8a (10 site classes as in M7, plus one class with ω = 1)	n/a	-3961.02	25	χ^2 = 1.22, df = 1, p = 0.27
M8 (10 site classes as in M7, plus one class with $\omega \geq 1$)	1.39 (9.1% of sites)	-3960.41	26	n/a

518

519

520 **Discussion**

521 Across taxa, many *de novo* evolved genes are expressed in the male reproductive
522 system [12, 31, 33, 34]. Identifying those genes that have evolved essential roles in
523 reproduction will provide insight into how newly evolved genes integrate with existing cellular
524 networks [47] and how evolutionary novelties permit adaptation in the face of sexual selection.
525 Here, we screened 42 putatively *de novo* evolved genes for major effects on male *D.*
526 *melanogaster* reproduction. Our primary screen identified three genes whose knockdown
527 caused an apparent reduction in male fertility. However, subsequent CRISPR-mediated gene
528 deletion revealed that only one of these genes, *atlas*, was truly essential. This result
529 underscores the importance of validating genes identified in RNAi screens through traditional
530 loss-of-function genetics and other approaches.

531 Loss *of atlas* function reduces fertility by affecting mature sperm production. During
532 spermiogenesis, *atlas* mutants show aberrant nuclear condensation and an inability to
533 individualize spermatid bundles successfully. GFP-tagged Atlas protein localizes to condensing
534 spermatid nuclei in the basal testis and partially co-localizes with Mst35Bb, a protamine around
535 which DNA is wrapped in mature, individualized sperm. Evolutionary analysis showed that the
536 *atlas* protein-coding sequence likely arose at the base of the *Drosophila* genus but was unlikely
537 to have played an essential role immediately upon birth, as the gene was subsequently lost
538 along several independent lineages. Within the *melanogaster* group of *Drosophila*, however,
539 the gene moved from the X chromosome to an autosome, where it formed a single
540 transcriptional unit with a conserved, non-coding sequence. Since this point, the gene has
541 encoded a protein with a conserved length, isoelectric point and male-specific expression
542 pattern, suggesting potential functional conversation over the last ~15 million years.

543

544

545

546 *Comparison of* atlas *null and frameshift allele phenotypes*

547 The null deletion and frameshift alleles of *atlas* all caused significantly reduced fertility
548 (Fig. 2), but the deletion allele resulted in essentially complete sterility, while residual fertility
549 remained in males homozygous for each frameshift allele. We noted above and depicted in Fig.
550 S2 how the frameshift alleles have the potential to encode an N-terminally truncated form of
551 Atlas that would contain amino acids 61-172 of the wild-type protein, if these alleles allow
552 translation initiation at the methionine-encoding codon 61. Such a truncated protein would be
553 35 percent shorter than wild-type, lack the predicted nuclear localization sequence, and have a
554 reduced isoelectric point of 7.1. Each of these factors could contribute to reduced functionality.

555 The broad sequence conservation of the 3' UTR across the *Drosophila* genus, including
556 in species that lack the *atlas* coding sequence, suggests the alternative possibility that the 3'
557 UTR also contributes to fertility. In one scenario, the 3' UTR could act as part of the *atlas* locus
558 by regulating RNA stability and/or the spatiotemporal control of Atlas protein translation. An
559 alternative possibility is that the 3' UTR functions on its own as a non-coding RNA. Indeed,
560 since the protein-coding region and the *D. melanogaster* 3' UTR were initially separate genetic
561 entities (and continue to be in *D. ananassae* and *D. virilis*), it is possible that each contributes to
562 fertility in a unique way. Future experiments that could distinguish these possibilities are
563 described below.

564 Based on the current evidence, however, we think that the primary way in which *atlas*
565 impacts fertility is through its protein-coding sequence. This conclusion is supported by: the
566 ~60-90 percent reduction in fertility in even the frameshift mutants; the stability of the Atlas-GFP
567 fusion protein and its presence in the spermatid nuclear condensation stages that immediately
568 precede the timing of the null mutant phenotype; and, the observation that the protein-coding
569 region shows a more highly conserved expression pattern than the UTR region within the
570 *melanogaster* group (Fig. 7 and Fig. S8).

571

572 *Atlas is an essential transition protein*

573 Several lines of evidence suggest that Atlas is a transition protein that facilitates the
574 change from histone-based to protamine-based chromatin packaging in spermatid nuclei. Atlas
575 localizes throughout spermatid nuclei (Fig. 6) and has biochemical properties consistent with
576 direct DNA interaction. The protein appears specifically at the late canoe stage of nuclear
577 compaction (Fig. 5B-D). Its lack of overlap with testis-specific histones (Fig. 5B), partial overlap
578 with Mst35Bb (Fig. 5D), removal from needle-stage nuclei (Fig. 6) and absence from mature
579 sperm (Figs. 5-6) are all consistent with the expression profile of a transition protein. Several
580 other transition proteins have been characterized in *D. melanogaster*, including Tpl94D, thmg-1,
581 thmg-2, and Mst84B [53, 58, 59]. Collectively, the transition proteins vary in the stage of
582 nuclear condensation at which they first appear and the range of nuclear shapes over which
583 they are found [79], but otherwise match Atlas in their biochemical properties, transient
584 expression, and localization throughout the nucleus. Compared to these other transition
585 proteins, Atlas is present over a fairly narrow range of nuclear condensation stages and reaches
586 its peak expression just prior to the onset of individualization. *Atlas* is also the only transition
587 protein gene characterized to date whose removal disrupts fertility, as *Tpl94D*, *thmg-1*, *thmg-2*
588 and *Mst84B* mutants are all fertile [53, 58, 59]. This may reflect the relatively later timing of
589 Atlas's expression in spermatid nuclei, reduced functional redundancy between DNA-binding
590 proteins at the later stages of condensation, a potential interaction between Atlas and an
591 essential protamine-like protein, and/or a more stringent requirement for DNA binding at these
592 stages.

593 Transition proteins give way in spermatid nuclei to protamine-like proteins, which bind
594 DNA in mature sperm and persist through fertilization. In this way, protamine-like proteins
595 function analogously to vertebrate protamines, though they are believed to be evolutionarily
596 independent [57, 64]. In *D. melanogaster*, protamine-like proteins include Mst35Ba, Mst35Bb,
597 Prtl99C and Mst77F [63, 80, 81]. Interestingly, while all characterized protamine-like proteins

598 are present in mature sperm, only some are essential for fertility. Knockouts of *Mst35Ba*,
599 *Mst35Bb*, or both show occasional nuclear shaping defects, but male fertility is normal [80,
600 82]. In contrast, mutants of *Prtl99C* or *Mst77F* are sterile. Prtl99C and Mst35Ba/b bind
601 condensed DNA independently of each other and contribute additively to the shortening of
602 needle-stage nuclei, but Prtl99C's effect is ~3-fold greater [63, 64]. This difference is apparently
603 great enough to reduce fertility only in *Prtl99C* mutants. In contrast, *Mst77F* and *Mst35Ba/b*
604 show a genetic interaction, as *Mst35Ba/b* null flies become nearly sterile in an *Mst77F*
605 heterozygous background [81]. Furthermore, while Mst35Bb-GFP is expressed in *Mst77F* nulls,
606 these flies show deformed spermatid nuclei that do not reach a recognizable needle-like stage.
607 Because *atlas* nulls show considerable phenotypic similarity to *Mst77F* nulls, but not *Prtl99C*
608 nulls, we hypothesize that Atlas may act in a pathway with Mst77F. Our observation of
609 inefficient IC movement down sperm tails in *atlas* null testes is reminiscent of a similar
610 phenotype in *Mst77F* nulls [81], providing further evidence that these proteins may act in a
611 common pathway. In both cases, ICs can form at misshapen canoe-stage nuclei, but fail at a
612 subsequent step. While the exact relationship between nuclear abnormalities in the late canoe
613 stage and individualization is not entirely understood, it is possible that nuclear shape and the
614 organization of nuclear bundles impact the ability of IC association and IC progression, as is
615 also observed in mutants of another gene, *dPSMG1*, which controls nuclear shape [59].
616

617 *Evolution of* atlas *and other de novo genes*

618 Because *de novo* genes emerge from non-coding sequences, they typically encode
619 proteins that are short and lack complex structure [3, 16]. Indeed, expanding the length of the
620 protein-coding region and evolving higher-level protein structures are hypothesized to be among
621 the final stages of new gene evolution [4]. In light of these constraints, what kinds of cellular
622 functions might be available to newly evolved proteins? Vakirlis et al. [83] overexpressed
623 emerging proto-genes in *S. cerevisiae* and found that those encoding proteins with predicted

624 transmembrane (TM) domains were more likely to be adaptive, as assessed by the effect of
625 proto-gene overexpression on growth rate. Such proteins may arise when thymine-rich
626 intergenic regions undergo mutations that allow protein-coding gene birth and expression, since
627 many codons with multiple U nucleotides encode amino acids commonly found in TM domains
628 [83]. Our imaging data, in addition to the prediction tools employed by Vakirlis et al. [83],
629 suggest that Atlas does not contain a TM domain. However, just as the amino acid
630 compositional requirements of a TM domain are not overly complex, neither are those of DNA
631 binding proteins. In essence, these proteins must simply be small, have a high concentration of
632 positively charged residues, and contain a nuclear localization signal, which itself requires a
633 small patch of positively charged residues [65]. Thus, DNA binding proteins may be a relatively
634 easy class of protein to evolve *de novo*.

635 While many putative *de novo* genes are expressed in the *D. melanogaster* testis [Fig. 1A
636 and ref. 19], *atlas* was the only verified hit from our screen that was essential for male fertility.
637 This result raises two related questions. First, why are so many of these other genes expressed
638 if their knockdown causes no obvious effect? In general, it is common for the knockdown of
639 protein-coding genes expressed in reproductive tissues in *D. melanogaster* to result in no
640 detectable fertility defects [46, 84-86]. One set of explanations center on technical issues: a
641 gene could be expressed outside of the cells targeted by the RNAi driver, knockdown level
642 could be insufficient to cause a phenotype, and/or a gene may be functionally redundant such
643 that its knockdown causes no apparent effect. Another hypothesis to explain this pattern is that
644 while the loss of function of such genes may cause small reductions in fertility that would be
645 subject to strong negative selection in nature, the conditions used to assay such knockdown
646 animals in primary screens are rarely tailored to detect differences of this magnitude. A third
647 possibility, not mutually exclusive with the above, is that while some genes may be expendable
648 in non-competitive, non-exhaustive mating conditions, their absence may result in lower fitness
649 in sperm competitive environment, environments in which males mate several times in quick

650 succession, or environments in which sperm must persist in storage for longer intervals or
651 during less optimal conditions [87-89]. These latter two possibilities are illustrated by some of
652 the other spermatid chromatin binding proteins previously characterized as “non-essential” (e.g.,
653 *Mst35Ba/Bb* and *Tpl94D*) [58, 80]. While these proteins are not *de novo* evolved, they provide
654 examples of genes whose mutations cause aberrant cellular phenotypes, such as the
655 abnormally shaped spermatid nuclei in *Mst35Ba/Bb,* and that may contribute to fertility defects
656 when placed in a sensitized genetic background [81].

657 It is also possible that some of the genes we screened are *de novo* genes that have no
658 functional effects and have thus never been selectively advantageous. For example, 8 of the 42
659 screened genes have annotated orthologs in either *D. melanogaster* alone or in only *D.*
660 *melanogaster* and *D. simulans*. Such recently “born” *de novo* genes could have been
661 propagated by neutral evolutionary processes but not yet experienced inactivating mutations
662 that would abolish their coding sequences and/or expression. This pattern would be consistent
663 with *de novo* genes’ high rates of both birth and death observed previously in *Drosophila* [38]
664 and yeast [83].

665 A second question raised by our finding that *atlas* encodes an essential transition protein
666 is: how might *atlas* have evolved to become essential for fertility in *D. melanogaster*, particularly
667 when other transition proteins appear functionally redundant? Other proteins involved in
668 spermatid chromatin compaction show variable levels of conservation across *Drosophila*. For
669 example, the protamines that contribute to DNA packaging in mature sperm [57] are found
670 across all sequenced *Drosophila* species [61], and orthologs are also reported in FlyBase from
671 other Dipteran and non-Dipteran insects [90]. However, transition protein Tpl94D is reported to
672 be restricted to species ranging from *D. melanogaster* to *D. pseudoobscura* [60], as are the
673 related proteins tHMG1 and tHMG2 with high-mobility group domains [58, 90]. Results like
674 these suggest that while some protamine-like proteins (i.e., Mst35Ba and Mst35Bb) have
675 consistently been among the final chromatin-packaging proteins, the specific proteins facilitating

676 the transition from histones to protamines have likely varied over evolutionary time. Against this
677 backdrop, and based on our analyses of the protein's presence/absence, biochemical
678 properties, and expression patterns in extant species, we hypothesize that while the Atlas
679 protein likely had some DNA-binding ability and male-specific expression upon its origin, it was
680 only one of several proteins involved in spermatid chromatin compaction. Since *atlas* was lost
681 independently in several lineages after its birth (Fig. 7), Atlas was likely non-essential at its
682 outset, but rather evolved an essential function within the *melanogaster* group of species. Such
683 evolution of essentiality could have occurred because of the loss of a protein with a
684 complementary function and/or changes in the process of spermiogenesis that thrust Atlas into
685 a functionally unique role. It is also worth noting that species that have evidently lost *atlas* might
686 have undergone other compensatory changes in their repertoires of spermatid DNA binding
687 proteins. For instance, *D. willistoni* lacks *atlas* but appears to have several additional paralogs
688 of the protamines found only in duplicate in *D. melanogaster*, which could have evolved
689 transition-protein-like roles.

690 While our study cannot establish whether the movement of the *atlas* protein-coding
691 sequence off of the X chromosome onto an autosome early in the evolution of the *melanogaster*
692 group (Fig. 7) affected the gene's essentiality, such movement remains noteworthy. Prior work
693 has found a significant dearth of testis-expressed genes on the X chromosome in *Drosophila*
694 [69, 91-93] and other species [94, 95]. Furthermore, *Drosophila* exhibit suppression of X-linked
695 testis-expressed genes, and transfer of such genes from the X chromosome to autosomal loci
696 results in higher expression levels [96-98]. One of several proposed mechanisms for both the
697 paucity of X-linked testis-expressed genes and the suppression of their expression is meiotic
698 sex chromosome inactivation (MSCI), in which the X chromosome becomes transcriptionally
699 silenced earlier than autosomes [99-104]. Thus, genes that affect meiotic or post-meiotic
700 processes, as *atlas* does, could exert beneficial effects more strongly and/or for a longer period
701 of time if they become encoded autosomally. While the *atlas* protein-coding sequence appears

702 to show male-specific expression regardless of its chromosomal location, it is possible that the
703 movement of *atlas* to chromosome 2 allowed it to evolve a broader or different expression
704 pattern that expanded or modified its role in spermiogenesis. The complex molecular bases of
705 both X suppression and "escape" from the X chromosome in *Drosophila* continue to be actively
706 investigated and debated [102-108], but continued research in this area might inform further
707 interrogation of the forces driving *atlas* off of the X chromosome.

708 The movement of the *atlas* protein-coding sequence to chromosome 2 also created the
709 two-exon gene observed in *D. melanogaster*, in which the longer second exon appears to be
710 entirely non-coding. This second exon is highly conserved across the genus in both sequence
711 and genomic location, and it shows male-specific expression in several species that lack the
712 protein-coding sequence upstream (Fig. 7 and Fig. S8). These patterns of conservation
713 suggest that the second exon might originally have been a non-coding RNA, a class of molecule
714 whose importance in *Drosophila* male reproduction has recently become recognized [109, 110].
715 While previous examples of functional ncRNAs in spermatogenesis have generally acted in
716 *trans* to regulate other genes or affect the functions of other proteins, it is also possible that the
717 long 3' UTR of *atlas* in *D. melanogaster* could affect the translation of *atlas* transcripts. Many
718 genes functioning in spermatid differentiation are transcribed early in spermatogenesis but
719 translationally repressed until later in spermiogenesis, a phenomenon that relies on various
720 forms of post-transcriptional regulation [111, 112]. Future studies of the *atlas* protein-coding
721 sequence in the absence of its 3' UTR, the expression patterns of Atlas protein in species in
722 which it is encoded from the X chromosome, or the genetic ablation of the conserved region in
723 species lacking the protein-coding sequence will provide additional insights.

724 A final issue raised by our results is the exact timing and mechanism of origin for the
725 *atlas* protein-coding sequence. The bioinformatic screen [19] that identified *atlas* and the other
726 genes tested in Fig. 1 was designed to identify both "*de novo*" genes, defined as protein-coding
727 regions in *Drosophila* that had recognizable, but non-ORF-maintaining, TBLASTN hits in

728 outgroup species, and “putative *de novo*” genes, which had no TBLASTN hits in outgroup
729 species. (Importantly, the screen also eliminated any protein with an identifiable protein
730 domain, thus reducing the chances of identifying divergent members of gene families.) The vast
731 majority of the genes we tested with RNAi, including *atlas*, fell into the putative *de novo*
732 category. The bioinformatic screen’s criteria were reasonable for a high-throughput analysis,
733 but BLAST-based methods have known limitations for detecting orthologous sequences in
734 diverged species [74, 113, 114]. The lack of identifiable *atlas* protein-coding genes in several
735 *Drosophila* species (e.g., *D. pseudoobscura* and *D. willistoni*) is unlikely to be due to BLAST
736 homology detection failure, and extensive synteny-based searches confirmed the gene’s
737 absence (Fig. S9). BLAST and synteny-based searches for orthologs in non-*Drosophila*
738 species also did not detect an ortholog, though BLAST searches are not predicted to have
739 adequate sensitivity for a protein of this size and evolutionary rate, at this level of species
740 divergence [74]. Hence, in addition to using synteny to search for orthologs, we used HMMER,
741 which employs hidden Markov models and builds a sequence profile of the target protein using
742 information from multiple orthologs. Since HMMER also did not detect orthologs outside of
743 *Drosophila*, we hypothesize that *atlas* evolved *de novo* at the base of the genus. However,
744 since we remain unable to identify the non-protein-coding sequence from which *atlas* arose, we
745 continue to refer to *atlas* as a putative *de novo* gene [5].

746 Overall, we find that while many putative *de novo* evolved genes are expressed in the *D.*
747 *melanogaster* testes, few have major, non-redundant effects on fertility. However, several such
748 genes have evolved critical roles at distinct stages of spermatogenesis and sperm function. We
749 showed previously that the putative *de novo* gene *saturn* is required for maximal sperm
750 production, as well as for the ability of transferred sperm to migrate successfully to sperm
751 storage organs in females [46]. Another putative *de novo* gene, *goddard*, is required for sperm
752 production and encodes a cytoplasmic protein that appears to localize to elongating axonemes
753 [20, 46]. Loss of *goddard* impairs the individualization of spermatid bundles [20], thus exerting

754 an effect that appears to be upstream of those observed for *saturn* and *atlas*. Here, we report
755 another novel function for a putative *de novo* gene: encoding an essential transition protein that
756 is necessary for proper nuclear condensation in spermiogenesis. Taken together, these results
757 demonstrate that while many *de novo* genes may play subtle roles or share functional
758 redundancy with other genes, *de novo* genes can also become essential players in complex
759 cellular processes that mediate successful reproduction.
760

761 **Materials and Methods**

762 *RNA Interference Screen*

763 *De novo* and putative *de novo* genes inferred to be no older than the *Drosophila* genus
764 were identified previously [19]. We filtered these genes with publicly available RNA-seq data
765 [115] to identify those expressed predominantly in the testes [>50% of RPKM sum deriving from
766 the testes from ModENCODE data; ref. 115], giving a total of 96 genes. To assess each of
767 these candidates for effects on male fertility, we induced knockdown in the male germline by
768 crossing UAS-RNAi flies to Bam-GAL4, UAS-Dicer2 flies [46, 116]. Control flies were generated
769 by crossing the attP-containing genetic background into which UAS-RNAi was inserted to the
770 same GAL4 line. Flies carrying UAS-RNAi were of two types. Roughly half of the genes had
771 publicly available lines from the Vienna *Drosophila* Resource Center [117] or the Transgenic
772 RNAi Project [50]. For the other genes, no publicly available RNAi stock was available, so we
773 constructed TRiP-style stocks in the pValium20 vector as previously described [85]. These
774 constructs were integrated into an AttP site in stock BL 25709 (y^1 v^1 P{nos-phiC31\int.NLS}X;
775 P{CaryP}attP40) from the Bloomington *Drosophila* Stock Center (injections by Genetivision;
776 Houston, TX, USA) and crossed into a *y v* background to screen for v^+. We attempted at least
777 two rounds of transgenic production for each gene. In total, we were able to obtain and test
778 RNAi lines for 57 of the 96 identified genes. Table S3 shows all RNAi lines used and lists the
779 short hairpin sequences cloned for the TRiP lines we constructed.

780 We initially screened males knocked down for each candidate gene for major fertility
781 defects by crossing groups of 7 knockdown or control males to 5 virgin Canton S females,
782 letting the adults lay eggs for ~48 hours, and then discarding adults and quantifying the resulting
783 progeny by counting the pupal cases, as previously described [46]. To assess the degree of
784 knockdown achieved, 10 whole males of each line were homogenized in TRIzol reagent (Life
785 Technologies, Carlsbad, CA). RNA isolation, DNAse treatment, cDNA synthesis and semi-
786 quantitative RT-PCR with gene-specific primers were performed as previously described;

787 amplification of *RpL32* was used as a positive control [46]. We evaluated knockdown efficiency
788 by agarose gel electrophoresis of RT-PCR products as one of four levels: “complete” if no
789 product from the knockdown cDNA sample was visible via agarose gel electrophoresis; “near
790 complete” for a very faint knockdown product that was also much less abundant than the control
791 product; “partial” for a more robust knockdown product that was still visibly less intense than
792 control; and “not knocked down” if the product intensity for the knockdown sample equaled or
793 exceeded that of the control. Any gene that did not show at least partial knockdown was
794 discarded from further analysis, leaving a total of 42 genes successfully screened. Table S3
795 shows the degree of knockdown achieved for each line.
796
797 *CRISPR Genome Editing*
798 To validate RNAi results for *atlas*, *CG43072* and *CG33284*, we used CRISPR/Cas9
799 genome-editing to generate null alleles that could be used for further analysis, as described
800 previously [20]. Briefly, our general strategy was to design gRNAs in the pU6.3 vector
801 (Drosophila Genome Resource Center (DGRC) #1362) that targeted each end of a locus.
802 These plasmids, along with plasmids encoding gRNAs that targeted the *w*+ locus, were co-
803 injected by Rainbow Transgenics (Camarillo, CA) into embryos laid by *vasa*-Cas9 females in a
804 *w*+ background, Bloomington stock #51323 [118]. G_0 animals were crossed to *w*- flies, and
805 members of G_1 broods with a higher-than-expected fraction of *w*- progeny were individually
806 crossed to an appropriate balancer line and then PCR-screened for the desired deletion of the
807 targeted locus.
808 We also constructed three frameshift, expected loss-of-function alleles for *atlas* by using
809 CRISPR to induce non-homologous end joining at a single PAM site just downstream of the
810 *atlas* start codon. *Vasa*-Cas9 embryos were co-injected and screened for w- progeny as
811 described above. We then used squish preps to isolate DNA from G1 flies and used a PCR-
812 RFLP assay to detect mutations. PCR products spanning the gRNA-targeted site were digested

813 with *Bfa*I (New England Biolabs (NEB), Ipswich, MA); undigested products in which the
814 expected *Bfa*I site was lost indicated a mutation, which was balanced and then confirmed by
815 PCR and sequencing of homozygous mutant lines.

816 We used scarless CRISPR editing and homology-directed repair (HDR) to insert the
817 GFP protein-coding sequence in-frame at the end of the *atlas* protein sequence (see Fig. S6;
818 (https://flycrispr.org/scarless-gene-editing/) [66-68]. We first generated an *atlas*-GFP DNA
819 construct by cloning the *atlas* protein-coding sequence into pENTR and using LR Clonase II
820 (Thermo Fisher Scientific, Waltham, MA) to recombine the sequence with pTWG (DGRC #1076;
821 T. Murphy), generating a C-terminally tagged *atlas*:GFP construct. We amplified the *atlas*
822 fragment from *vasa*-Cas9 strain #51323 genomic DNA. Once *atlas*-GFP was obtained in a
823 plasmid, we amplified it with primers that contained 5' *Esp*3I sites and overhangs designed for
824 Golden Gate Assembly (GGA) and that, in the case of the reverse primer, also added on 42
825 nucleotides downstream of the *atlas* stop codon to reach a PAM site identified by FlyCRISPR
826 TargetFinder [119] as being optimal for Cas9/gRNA recognition and cleavage. The primer also
827 introduced a mutation in the PAM site so that insertion of the designed piece of DNA into the
828 genome *in vivo* would not be subject to re-cutting. We also used the NEB Q5 Site-Directed
829 Mutagenesis kit to introduce a silent mutation into the *atlas* protein-coding sequence to
830 eliminate an internal *Esp*3I site. The resulting construct was used as the "left" homology arm for
831 homology-directed repair (HDR) editing. We constructed a "right" homology arm by using NEB
832 Q5 PCR to amplify a 982-bp fragment downstream of the PAM site, using primers modified to
833 contain *Esp*3I sites and overhangs compatible with GGA. We performed GGA by combining
834 these left and right arms, a plasmid containing a PiggyBac transposase-excisable 3xP3-dsRed
835 flanked by *Esp*3I sites, and backbone plasmid pXZ13, with *Esp*3I and T4 DNA ligase (NEB). A
836 combination of colony PCR, restriction digestion and sequencing identified properly assembled
837 plasmids suitable for HDR.

838 *Vasa*-Cas9 embryos were co-injected with the assembled plasmid and a pU6.3 plasmid
839 encoding a gRNA targeting the region just downstream of the *atlas* stop codon. G0 flies were
840 crossed to w^{1118} adults, and G1 flies were screened for red fluorescent eyes using the
841 NIGHTSEA system (NIGHTSEA LLC, Lexington, MA). Six balanced lines from two independent
842 G1 broods were established. To remove the dsRed from the *atlas* locus, we crossed these lines
843 to a PiggyBac transposase line (BDSC #8285) and then selected against pBac and dsRed in
844 the following generation. PCR and sequencing confirmed the expected "scarless" insertion of
845 GFP at the *atlas* locus.
846
847 atlas *Genomic Rescue Line*
848 We constructed an HA-tagged *atlas* rescue line that contained the *atlas* gene flanked by
849 1345 bp of sequence upstream of the start codon (but excluding the coding sequence of
850 upstream gene *CG3124*) and 3000 bp of sequence downstream of the stop codon (including the
851 full 3' UTR) as follows. Genomic sequences were PCR amplified using Q5 High fidelity
852 Polymerase (NEB), purified Canton S genomic DNA (Gentra Puregene Tissue Kit, Qiagen,
853 Germantown, MD), and the atlas rescue F1/R1 and atlas rescue F3/R3 primer sets (see Table
854 S4). The 3x-HA tag was likewise amplified from pTWH (DGRC 1100; T. Murphy) using atlas
855 rescue F2/R2 primers. These DNA fragments were subsequently assembled into a XbaI/AscI-
856 linearized w+attB plasmid (Addgene, Watertown, MA, plasmid 30326, deposited by J.
857 Sekelsky). The assembled construct was then phiC31 integrated into the PBac{y^+-attP-
858 3B}VK00037 (Bloomington *Drosophila* Stock Center (BDSC) stock #24872) docking site
859 (Rainbow Transgenics) and crossed into the *atlas* null background to assess rescue.
860
861 *Fertility Assays and Sperm Visualization*
862 To validate the finding of reduced fertility for *atlas* knockdown males in the group fertility
863 assay described above, we performed single-pair fertility assays in which knockdown or mutant

864 males or their controls were mated individually to Canton S virgin females. Based on previous
865 experience analyzing genes that resulted in sterility or near-sterility [20, 46], we designed
866 assays with *N* = 20-30 flies per male genotype. Matings were observed, and males were
867 discarded after copulation. Females were allowed to lay eggs into the vials for 4 days and then
868 discarded. Pupal cases were counted as a measure of fertility. Crosses to generate and
869 mating assays involving RNAi flies were maintained at 25º to optimize knockdown. Before all
870 assays, flies were reared to sexual maturity (3-7 days) in single-sex groups on cornmeal-
871 molasses food supplemented with dry yeast grains [46].

872 To assess the level of fertility conferred by the *atlas*-GFP allele, we crossed *atlas*-GFP
873 and w^{1118} flies to Δ*atlas*/SM6. Males with genotypes atlas-GFP/Δ*atlas* and +/Δ*atlas* were
874 compared using the single-pair fertility assay described above.

875 To observe the production of sperm in knockdown or mutant males, we introduced the
876 Mst35Bb-GFP marker into these males, which labels mature sperm and late-stage spermatid
877 nuclei with GFP [51]. Samples were prepared, imaged and analyzed as described previously
878 [46].

879

880 *Atlas-GFP Ectopic Expression*

881 We used the Gateway cloning system (Thermo) to construct an *atlas*-GFP transgene
882 expressed under UAS control (primers in Table S4). The *atlas* protein-coding sequence in
883 pENTR was recombined with pTWG (*Drosophila* Genomics Resource Center, T. Murphy) as
884 described above. The resulting plasmid was then inserted into w- flies using P-element-
885 mediated transposition (Rainbow Transgenics), w+ G1s were selected, and several
886 independent insertions were balanced. We crossed male UAS-*atlas*:GFP flies to females from
887 two different driver lines: *tubulin*-GAL4 (to drive ubiquitous expression) and *Bam*-GAL4 (to drive
888 expression in the early germline). We dissected larval salivary glands of the *tub*>*atlas*:GFP
889 males, since these cells are exceptionally large and ideal for visualizing subcellular localization.

890 We then dissected the testes of *Bam>atlas*:GFP males to evaluate whether the localization
891 pattern observed in the salivary gland was consistent in testis tissue, albeit not the same cells in
892 which endogenous atlas appears to be expressed. Protein localization was visualized by
893 fluorescence confocal microscopy on a Leica SP5 microscope (Leica Microsystems, Wetzlar,
894 Germany) and images were captured with LASAF as described previously [20].
895
896 *Imaging Spermatogenesis and Spermatid Nuclear Condensation*
897 We used phase-contrast microscopy to examine the stages of spermatogenesis in whole
898 mount testes [120]. To assess the processes of nuclear condensation and individualization of
899 64-cell cysts of spermatids in the post-meiotic stages of spermatogenesis, we used
900 fluorescence and confocal microscopy to visualize actin-based individualization complexes and
901 nuclei. Samples were processed, and actin and nuclear DNA were visualized with TRITC-
902 phalloidin (Molecular Probes, Eugene, OR) and TOPRO-3 iodine (Thermo), respectively, as
903 described previously [20]. The final stages of nuclear condensation were visualized with the
904 Mst35Bb-GFP marker described above, as well as an equivalent marker, Mst35Bb-dsRed [51].
905 Earlier nuclear stages were visualized with histone H2AvD-RFP (BDSC stock #23651), which is
906 present in round spermatid nuclei and the earliest stages of nuclear elongation [53, 121].
907 Images with H2AvD-RFP were obtained with epifluorescence microscopy, since we lack an
908 appropriate confocal laser for RFP.
909 To examine spermatid nuclei at various stages of condensation, we visualized nuclear
910 bundles using TOPRO-3. Testes of newly eclosed (<1 day old) *atlas* null and control males
911 were dissected in PBS. Testes were then transferred to a droplet of 2% paraformaldehyde on
912 poly-L-lysine treated glass slides and were gently shredded in the post-meiotic region to release
913 sperm bundles. Testes were gently squashed beneath coverslips coated in Sigmacote (Sigma
914 Aldrich, St. Louis, MO). We then froze slides in liquid nitrogen for a few seconds and popped off
915 of the siliconized coverslip with a razor. Slides were incubated in Coplin jars filled with 95%

916 ethanol at -20°C for 30 minutes and then mounted in VECTASHIELD (Vector Laboratories,
917 Burlingame, CA). Nuclear staging was performed by examining the shape of the nuclei. Early
918 and late canoe stages of condensation were distinguished by the absence or presence of
919 Mst35Bb-GFP, respectively. Elongated and late canoe stages were distinguished by the
920 presence or absence, respectively, of vesicles of GFP-tagged nuclear proteins (Atlas-GFP or
921 Mst35Bb-GFP) located basal to the nuclei. Examples of stages are given in Fig. 4. Confocal
922 stacks were taken on a Leica SP5 microscope, images were captured by LASAF, and ImageJ
923 was used to flatten stacks into a single, two-dimensional image. All intact nuclear bundles were
924 counted for each dissection.

925 For the experiments measuring nuclear condensation stage (Table S1), a sample size of
926 N = 10 for each genotype was selected based on the magnitude of the *atlas* null phenotype and
927 the consistent differences observed in previous dissections of these genotypes with Mst35Bb-
928 GFP. Likewise, for the IC-nuclear bundle association and IC progression analysis (Fig. 3C-D),
929 we selected sample sizes of N = ~15 per genotype based on pilot experiments showing that
930 aberrant actin phenotypes were highly consistent in null testes and previous experience with
931 such quantification [20].

932

933 *Evolutionary and Gene Expression Analysis of* atlas

934 We searched for orthologs of the *D. melanogaster* Atlas protein in the original 12
935 sequenced *Drosophila* species with BLASTP searches in FlyBase [122]. We also used
936 TBLASTN searches to identify orthologs in species lacking complete protein annotations. We
937 identified syntenic regions for each species by looking for conserved neighboring genes, such
938 as *ord* and *CG3124*. In addition to analyzing the *atlas* coding region, we conducted separate
939 BLASTN searches for the sequence of the *D. melanogaster* 3'UTR across *Drosophila* species
940 since it has a different conservation pattern than the coding sequence.

941 To test for sex-specific expression biases for both the ORF and the 3' UTR sequences,
942 we used adult male- and female-specific RNA-seq data from numerous *Drosophila* species
943 accessed through the Genomics Education Partnership version of the UCSC Genome Browser
944 (http://gander.wustl.edu/) and initially collected by Brown et al. [115] and Chen et al. [123] . We
945 also confirmed these findings experimentally in several species by performing RT-PCR on
946 cDNA isolated from whole males and whole females, as previously described [46].
947 To search for *atlas* orthologs in non-*Drosophila* Dipterans, we obtained from ENSEMBL
948 Metazoa the genomes of *Musca domestica, Glossina morsitans, Lucilia cuprina, Aedes aegypti,*
949 *Anopheles darlingi, Anopheles gambiae, Culex quinquefasciatus* and *Mayetiola destructor*. We
950 performed a synteny search (summarized in Fig. S11) in each species by identifying the nearest
951 neighbors of *atlas* in the *D. ananassae* and *D. virilis* genomes that had an identifiable homolog
952 in each species. In all cases, the homologs of the nearest neighbors on each side of *atlas* were
953 found on different contigs, suggesting synteny breakdown. We obtained up to 1 Mb of
954 sequence on each side of each identified homolog and queried it with BLASTN, TBLASTN, and
955 Exonerate [124] for regions with significantly similarity to any portion of the Atlas protein or
956 cDNA sequences. No significant hits, and no hits better than what could be found in other parts
957 of the genome, were found. Finally, we used HMMER to search for orthologs in all annotated
958 proteomes and all metazoan proteomes. We first queried the database with Atlas from either *D.*
959 *melanogaster* or *D. virilis* and accepted hits that fell below an e-value cutoff of 0.01 and a
960 minimum hit length of 3%. These hits were then included iteratively in subsequent searches
961 until no new significant hits were found.
962 We analyzed the molecular evolution of the *atlas* protein-coding sequence by obtaining
963 orthologous protein-coding sequences from *melanogaster* group species. (Analysis out of this
964 group was not performed due to high sequence divergence and poor alignment quality.) We
965 used BLASTP to identify these sequences from GenBank and then extracted the coding DNA
966 sequence for each. Sequences were aligned, checked for recombination, used to construct a

967 gene tree, and analyzed with the PAML sites test as described previously [78], except that
968 alignment positions that included gaps were masked from the PAML analysis. We initially
969 analyzed a set of 13 species (*melanogaster*, *simulans*, *sechellia*, *yakuba*, *erecta*, *suzukii*,
970 *takahashii*, *biarmipes*, *rhopaloa*, *ficusphila*, *elegans*, *eugracilis* and *ananassae*); we excluded an
971 ortholog detected in *D. bipectinata* due to poor alignment. This initial analysis detected a class
972 of sites with significant evidence of positive selection, but closer inspection of the alignment
973 revealed that the site with the strongest evidence of selection, corresponding to *D.*
974 *melanogaster* residue 31R, may have been driven by a questionable alignment due to an
975 insertion in that region that was unique to *D. takahashii*. The reported results come from an
976 analysis that excluded *D. takahashii*, which produced a more reliable alignment and showed no
977 evidence for any sites under positive selection.
978

Acknowledgements

We thank Dr. Alexis Hill for assistance with developing the *atlas*-GFP transgene; Gynesis Vance, Elvis Perez, Ishanpepe Jagusah, Emily Gualdino, and students in the Fall 2016 and Fall 2017 sections of Biology 261 Lab at Holy Cross for assistance with RNAi fertility screens; Dr. Rob Bellin for assistance with confocal microscopy; Dr. Justin McAlister for use of the NIGHTSEA system; the Bloomington Stock Center and the Vienna *Drosophila* Resource Center for fly strains; and, Dr. Mariana Wolfner and members of the Findlay and Bornberg-Bauer labs for productive discussions about the project and the manuscript. This work was supported by NSF CAREER Award #1652013 (to GDF) and a Humboldt Fellowship (to AG). GDF and EBB are also grateful to the Evolution Think Tank at the University of Münster, which funded GDF as a visiting fellow for in-person collaboration.

991 **References**
992
993 1. Zhang J. Evolution by gene duplication: an update. Trends Ecol Evol. 2003;18(6):292-8.
994 doi: https://doi.org/10.1016/S0169-5347(03)00033-8.
995 2. Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. High
996 spontaneous rate of gene duplication in *Caenorhabditis elegans*. Curr Biol. 2011;21(4):306-10.
997 doi: 10.1016/j.cub.2011.01.026.
998 3. Van Oss SB, Carvunis A-R. *De novo* gene birth. PLoS Genet. 2019;15(5):e1008160. doi:
999 10.1371/journal.pgen.1008160.
1000 4. Bornberg-Bauer E, Schmitz JF. Fact or fiction: Updates on how protein-coding genes
1001 might emerge *de novo* from previously non-coding DNA. F1000Res. 2017;6:57. doi:
1002 10.12688/f1000research.10079.1.
1003 5. McLysaght A, Hurst LD. Open questions in the study of *de novo* genes: what, how and
1004 why. Nat Rev Genet. 2016;17:567.
1005 6. Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, et al. On the origin of *de novo*
1006 genes in *Arabidopsis thaliana* populations. Genome Biol Evol. 2016;8(7):2190-202. doi:
1007 10.1093/gbe/evw164.
1008 7. Guerzoni D, McLysaght A. *De novo* genes arise at a slow but steady rate along the
1009 primate lineage and have been subject to incomplete lineage sorting. Genome Biol Evol.
1010 2016;8(4):1222-32. doi: 10.1093/gbe/evw074.
1011 8. Lu T-C, Leu J-Y, Lin W-C. A comprehensive analysis of transcript-supported de novo
1012 genes in *Saccharomyces sensu stricto* yeasts. Mol Biol Evol. 2017;34(11):2823-38. doi:
1013 10.1093/molbev/msx210.
1014 9. Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, et al. Rapid evolution of protein
1015 diversity by de novo origination in *Oryza*. Nat Ecol Evol. 2019;3(4):679-90. doi: 10.1038/s41559-
1016 019-0822-5.
1017 10. Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of *de novo* genes in
1018 *Drosophila melanogaster* populations. Science. 2014;343(6172):769-72. doi:
1019 10.1126/science.1248286.
1020 11. Puntambekar S, Newhouse R, Navas JSM, Chauhan R, Vernaz G, Willis T, et al.
1021 Evolutionary divergence of novel open reading frames in cichlids speciation. Sci Rep.
1022 2020;10(1). doi: 10.1038/s41598-020-78555-0.
1023 12. Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, et al.
1024 Origins of *de novo* genes in human and chimpanzee. PLoS Genet. 2016;11(12):e1005721.
1025 13. Chamakura KR, Tran JS, O'Leary C, Lisciandro HG, Antillon SF, Garza KD, et al. Rapid
1026 de novo evolution of lysis genes in single-stranded RNA phages. Nat Commun. 2020;11(1). doi:
1027 10.1038/s41467-020-19860-0.
1028 14. Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al.
1029 Proto-genes and *de novo* gene birth. Nature. 2012;487:370-4.
1030 15. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted
1031 by the preadaptation hypothesis of *de novo* gene birth. Nat Ecol Evol. 2017;1(6):0146.
1032 16. Schlötterer C. Genes from scratch – the evolutionary fate of *de novo* genes. Trends
1033 Genet. 2015;31(4):215-9. doi: 10.1016/j.tig.2015.02.007.
1034 17. Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient *de novo* genes can evolve from
1035 frozen accidents that escaped rapid transcript turnover. Nat Ecol Evol. 2018;2(10):1626-32. doi:
1036 10.1038/s41559-018-0639-7.
1037 18. Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A molecular
1038 portrait of de novo genes in yeasts. Mol Biol Evol. 2018;35(3):631-45. doi:
1039 10.1093/molbev/msx315.

19. Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving de novo genes drives protein-coding novelty in *Drosophila*. J Mol Evol. 2020;88(4):382-98. doi: 10.1007/s00239-020-09939-z.
20. Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, et al. Structural and functional characterization of a putative de novo gene in *Drosophila*. Nat Commun. 2021;12:1667. doi: 10.1101/2021.01.18.427054.
21. Keeling DM, Garza P, Nartey CM, Carvunis A-R. The meanings of 'function' in biology and the problematic case of de novo gene emergence. eLife. 2019;8:e47014. doi: 10.7554/elife.47014.
22. Cai J, Zhao R, Jiang H, Wang W. *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. Genetics. 2008;179(1):487-96. doi: 10.1534/genetics.107.084491.
23. Li D, Yan Z, Lu L, Jiang H, Wang W. Pleiotropy of the *de novo*-originated gene *MDF1*. Sci Rep. 2014;4(1):7280. doi: 10.1038/srep07280.
24. Xiao W, Liu H, Li Y, Li X, Xu C, Long M, et al. A rice gene of *de novo* origin negatively regulates pathogen-induced defense response. PLoS One. 2009;4(2):e4603. doi: 10.1371/journal.pone.0004603.
25. Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, et al. *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. Mol Biol Evol. 2018;35(3):593-606. doi: 10.1093/molbev/msx311.
26. Zhuang X, Yang C, Murphy KR, Cheng CHC. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proc Natl Acad Sci U S A. 2019;116(10):4400-5. doi: 10.1073/pnas.1817138116.
27. Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, et al. A de novo evolved gene in the house mouse regulates female pregnancy cycles. eLife. 2019;8. doi: 10.7554/elife.44392.
28. Suenaga Y, Islam SMR, Alagu J, Kaneko Y, Kato M, Tanaka Y, et al. NCYM, a *cis*-antisense gene of MYCN, encodes a *de novo* evolved protein that inhibits GSK3β resulting in the stabilization of MYCN in human neuroblastomas. PLoS Genet. 2014;10(1):e1003996. doi: 10.1371/journal.pgen.1003996.
29. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 2005;3(11):e357.
30. Baker RH, Narechania A, Johns PM, Wilkinson GS. Gene duplication, tissue-specific gene expression and sexual conflict in stalk-eyed flies (Diopsidae). Philos Trans R Soc Lond B Biol Sci. 2012;367(1600):2357-75. doi: 10.1098/rstb.2011.0287.
31. Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. Mol Plant. 2015;8(6):935-45. doi: 10.1016/j.molp.2014.12.008; 20
10.1016/j.molp.2014.12.008.
32. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20(10):1313-26. doi: 10.1101/gr.101386.109.
33. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proc Natl Acad Sci U S A. 2006;103(26):9935-9. doi: 10.1073/pnas.0509809103.
34. Begun DJ, Lindfors HA, Kern AD, Jones CD. Evidence for *de Novo* Evolution of Testis-Expressed Genes in the *Drosophila yakuba/Drosophila erecta* Clade. Genetics. 2007;176(2):1131-7. doi: 10.1534/genetics.106.069245.
35. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. Nat Rev Genet. 2014;15(11):734-48. doi: 10.1038/nrg3802.
36. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. Cell Rep. 2013;3(6):2179-90. doi: 10.1016/j.celrep.2013.05.031.

1091 37. Sorourian M, Kunte MM, Domingues S, Gallach M, Özdil F, Río J, et al. Relocation
1092 facilitates the acquisition of short *cis*-regulatory regions that drive the expression of retrogenes
1093 during spermatogenesis in *Drosophila*. Mol Biol Evol. 2014;31(8):2170-80. doi:
1094 10.1093/molbev/msu168.
1095 38. Palmieri N, Kosiol C, Schlötterer C. The life cycle of *Drosophila* orphan genes. eLife.
1096 2014;3:e01311. doi: 10.7554/eLife.01311.
1097 39. Wilburn DB, Swanson WJ. From molecules to mating: Rapid evolution and biochemical
1098 studies of reproductive proteins. J Proteomics. 2016;135:12-25. doi:
1099 10.1016/j.jprot.2015.06.007.
1100 40. VanKuren NW, Long M. Gene duplicates resolving sexual conflict rapidly evolved
1101 essential gametogenesis functions. Nat Ecol Evol. 2018;2(4):705-12. doi: 10.1038/s41559-018-
1102 0471-0.
1103 41. Hales KG, Korey CA, Larracuente AM, Roberts DM. Genetics on the fly: a primer on the
1104 *Drosophila* model system. Genetics. 2015;201(3):815-42. doi: 10.1534/genetics.115.183392.
1105 42. Demarco RS, Eikenes TH, Haglund K, Jones DL. Investigating spermatogenesis in
1106 *Drosophila melanogaster*. Methods. 2014;68(1):218-27. doi: 10.1016/j.ymeth.2014.04.020.
1107 43. Witt E, Benjamin S, Svetec N, Zhao L. Testis single-cell RNA-seq reveals the dynamics
1108 of de novo gene transcription and germline mutational bias in *Drosophila*. eLife. 2019;8. doi:
1109 10.7554/elife.47138.
1110 44. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. *De novo* ORFs in
1111 Drosophila are important to organismal fitness and evolved rapidly from previously non-coding
1112 sequences. PLoS Genet. 2013;9(10):e1003860. doi: 10.1371/journal.pgen.1003860.
1113 45. Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan L, Pang N, et al. New
1114 genes often acquire male-specific functions but rarely become essential in *Drosophila*. Genes
1115 Dev. 2017;31(18):1841-6. doi: 10.1101/gad.303131.117.
1116 46. Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, et al. The
1117 *goddard* and *saturn* genes are essential for *Drosophila* male fertility and may have arisen *de*
1118 *novo*. Mol Biol Evol. 2017;34(5):1066-82. doi: 10.1093/molbev/msx057.
1119 47. Abrusán G. Integration of new genes into cellular networks, and their structural
1120 maturation. Genetics. 2013;195(4):1407-17. doi: 10.1534/genetics.113.152256.
1121 48. Zhang W, Landback P, Gschwend AR, Shen B, Long M. New genes drive the evolution
1122 of gene interaction networks in the human and mouse genomes. Genome Biol. 2015;16(1). doi:
1123 10.1186/s13059-015-0772-4.
1124 49. Wu DD, Irwin DM, Zhang YP. De novo origin of human protein-coding genes. PLoS
1125 Genet. 2011;7(11). doi: 10.1371/journal.pgen.1002379.
1126 50. Ni J-Q, Zhou R, Czech B, Liu L-P, Holderbaum L, Yang-Zhou D, et al. A genome-scale
1127 shRNA resource for transgenic RNAi in *Drosophila*. Nat Methods. 2011;8(5):405-7. doi:
1128 10.1038/nmeth.1592.
1129 51. Manier MK, Belote JM, Berben KS, Novikov D, Stuart WT, Pitnick S. Resolving
1130 mechanisms of competitive fertilization success in *Drosophila melanogaster*. Science.
1131 2010;328(5976):354-7. doi: 10.1126/science.1187096.
1132 52. Steinhauer J. Separating from the pack: molecular mechanisms of *Drosophila* spermatid
1133 individualization. Spermatogenesis. 2015;5(2):e1041345. doi:
1134 10.1080/21565562.2015.1041345.
1135 53. Rathke C, Baarends WM, Jayaramaiah-Raja S, Bartkuhn M, Renkawitz R, Renkawitz-
1136 Pohl R. Transition from a nucleosome-based to a protamine-based chromatin configuration
1137 during spermiogenesis in *Drosophila*. J Cell Sci. 2007;120(9):1689-700. doi:
1138 10.1242/jcs.004663.
1139 54. Fabian L, Brill JA. Drosophila spermiogenesis. Spermatogenesis. 2012;2(3):197-212.
1140 doi: 10.4161/spmg.21798.

1141 55. Rathke C, Baarends WM, Awe S, Renkawitz-Pohl R. Chromatin dynamics during
1142 spermiogenesis. Biochim Biophys Acta. 2014;1839(3):155-68. doi:
1143 10.1016/j.bbagrm.2013.08.004.
1144 56. Kanippayoor RL, Alpern JHM, Moehring AJ. Protamines and spermatogenesis in
1145 *Drosophila* and *Homo sapiens*. Spermatogenesis. 2013;3(2):e24376. doi: 10.4161/spmg.24376.
1146 57. Jayaramaiah Raja S, Renkawitz-Pohl R. Replacement by *Drosophila melanogaster*
1147 protamines and Mst77F of histones during chromatin condensation in late spermatids and role
1148 of Sesame in the removal of these proteins from the male pronucleus. Mol Cell Biol.
1149 2005;25(14):6165-77. doi: 10.1128/mcb.25.14.6165-6177.2005.
1150 58. Gärtner SMK, Rothenbusch S, Buxa MK, Theofel I, Renkawitz R, Rathke C, et al. The
1151 HMG-box-containing proteins tHMG-1 and tHMG-2 interact during the histone-to-protamine
1152 transition in *Drosophila* spermatogenesis. Eur J Cell Biol. 2015;94(1):46-59. doi:
1153 https://doi.org/10.1016/j.ejcb.2014.10.005.
1154 59. Gärtner SMK, Hundertmark T, Nolte H, Theofel I, Eren-Ghiani Z, Tetzner C, et al. Stage-
1155 specific testes proteomics of *Drosophila melanogaster* identifies essential proteins for male
1156 fertility. Eur J Cell Biol. 2019;98(2):103-15. doi: https://doi.org/10.1016/j.ejcb.2019.01.001.
1157 60. Alvi ZA, Chu T-C, Schawaroch V, Klaus AV. Genomic and expression analysis of
1158 transition proteins in *Drosophila*. Spermatogenesis. 2016;5(3):e1178518. doi:
1159 10.1080/21565562.2016.1178518.
1160 61. Alvi ZA, Chu T-C, Schawaroch V, Klaus AV. Protamine-like proteins in 12 sequenced
1161 species of *Drosophila*. Protein Pept Lett. 2013;20(1):17-35. doi: 10.2174/092986613804096847.
1162 62. Dorus S, Freeman ZN, Parker ER, Heath BD, Karr TL. Recent Origins of Sperm Genes
1163 in Drosophila. Mol Biol Evol. 2008;25(10):2157-66. doi: 10.1093/molbev/msn162.
1164 63. Eren-Ghiani Z, Rathke C, Theofel I, Renkawitz-Pohl R. Prtl99C Acts Together with
1165 Protamines and Safeguards Male Fertility in *Drosophila*. Cell Rep. 2015;13(11):2327-35. doi:
1166 10.1016/j.celrep.2015.11.023.
1167 64. Doyen CM, Chalkley GE, Voets O, Bezstarosti K, Demmers JA, Moshkin YM, et al. A
1168 Testis-Specific Chaperone and the Chromatin Remodeler ISWI Mediate Repackaging of the
1169 Paternal Genome. Cell Rep. 2015;13(7):1310-8. doi: 10.1016/j.celrep.2015.10.010.
1170 65. Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH. Classical nuclear
1171 localization signals: definition, function, and interaction with Importin α. J Biol Chem.
1172 2007;282(8):5101-5. doi: 10.1074/jbc.r600026200.
1173 66. Hill A, Jain P, Ben-Shahar Y. The *Drosophila* ERG channel seizure plays a role in the
1174 neuronal homeostatic stress response. PLoS Genet. 2019;15(8):e1008288. doi:
1175 https://doi.org/10.1371/journal.pgen.1008288.
1176 67. Bier E, Harrison MM, O'Connor-Giles KM, Wildonger J. Advances in Engineering the Fly
1177 Genome with the CRISPR-Cas System. Genetics. 2018;208(1):1-18. doi:
1178 10.1534/genetics.117.1113.
1179 68. Bruckner JJ, Zhan H, Gratz SJ, Rao M, Ukken F, Zilberg G, et al. Fife organizes synaptic
1180 vesicles and calcium channels for high-probability neurotransmitter release. J Cell Biol.
1181 2017;216(1):231-46. doi: 10.1083/jcb.201601098.
1182 69. Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. Genomic and
1183 functional evolution of the *Drosophila melanogaster* sperm proteome. Nat Genet.
1184 2006;38(12):1440-5. doi: 10.1038/ng1915.
1185 70. Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, et al. The
1186 *Drosophila melanogaster* sperm proteome-II (DmSP-II). J Proteomics. 2010;73(11):2171-85.
1187 doi: 10.1016/j.jprot.2010.09.002.
1188 71. Schuh M, Lehner CF, Heidmann S. Incorporation of *Drosophila* CID/CENP-A and CENP-
1189 C into Centromeres during Early Embryonic Anaphase. Curr Biol. 2007;17(3):237-43. doi:
1190 10.1016/j.cub.2006.11.051.

1191 72. Rele CP, Sandlin KM, Leung W, Reed LK. Manual annotation of genes within *Drosophila*
1192 Species: the Genomics Education Partnership protocol. bioRxiv [Preprint]. 2021 [posted 2020
1193 Dec 11; revised 2020 Dec 12; revised 2020 Dec 14; revised 2020 Dec 15; revised 2021 Jan 27;
1194 revised 2021 May 4; revised 2021 May 21; revised 2021 June 5;
1195 cited 2021 Aug 10]. Available from:
1196 https://www.biorxiv.org/content/10.1101/2020.12.10.420521v8.article-info doi:
1197 https://doi.org/10.1101/2020.12.10.420521
1198 73. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, et al.
1199 Polytene chromosomal maps of 11 Drosophila species: the order of genomic scaffolds inferred
1200 from genetic and physical maps. Genetics. 2008;179(3):1601-55. doi:
1201 10.1534/genetics.107.086074.
1202 74. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be
1203 explained by homology detection failure. PLoS Biol. 2020;18(11):e3000862. doi:
1204 10.1371/journal.pbio.3000862.
1205 75. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018
1206 update. Nucleic Acids Res. 2018;46(W1):W200-W4. doi: 10.1093/nar/gky448.
1207 76. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al.
1208 OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant,
1209 archaeal, bacterial and viral orthologs. Nucleic Acids Res. 2017;45(D1):D744-D9. doi:
1210 10.1093/nar/gkw1119.
1211 77. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol.
1212 2007;24(8):1586-91. doi: 10.1093/molbev/msm088.
1213 78. McGeary MK, Findlay GD. Molecular evolution of the sex peptide network in *Drosophila*.
1214 J Evol Biol. 2020;33(5):629-41. doi: 10.1111/jeb.13597.
1215 79. Hundertmark T, Gärtner SMK, Rathke C, Renkawitz-Pohl R. Nejire/dCBP-mediated
1216 histone H3 acetylation during spermatogenesis is essential for male fertility in *Drosophila*
1217 *melanogaster*. PLoS One. 2018;13(9):e0203622. doi: 10.1371/journal.pone.0203622.
1218 80. Rathke C, Barckmann B, Burkhard S, Jayaramaiah-Raja S, Roote J, Renkawitz-Pohl R.
1219 Distinct functions of Mst77F and protamines in nuclear shaping and chromatin condensation
1220 during *Drosophila* spermiogenesis. Eur J Cell Biol. 2010;89(4):326-38. doi:
1221 10.1016/j.ejcb.2009.09.001.
1222 81. Kimura S, Loppin B. The *Drosophila* chromosomal protein Mst77F is processed to
1223 generate an essential component of mature sperm chromatin. Open Biol. 2016;6(11):160207.
1224 doi: 10.1098/rsob.160207.
1225 82. Tirmarche S, Kimura S, Sapey-Triomphe L, Sullivan W, Landmann F, Loppin B.
1226 *Drosophila* protamine-like Mst35Ba and Mst35Bb are required for proper sperm nuclear
1227 morphology but are dispensable for male fertility. G3 (Bethesda). 2014;4(11):2241-5. doi:
1228 10.1534/g3.114.012724.
1229 83. Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, et al. De novo
1230 emergence of adaptive membrane proteins from thymine-rich genomic sequences. Nat
1231 Commun. 2020;11(1). doi: 10.1038/s41467-020-14500-z.
1232 84. Ravi Ram K, Wolfner MF. Sustained post-mating response in *Drosophila melanogaster*
1233 requires multiple seminal fluid proteins. PLoS Genet. 2007;3(12):e238. doi:
1234 10.1371/journal.pgen.0030238.
1235 85. Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. Evolutionary rate
1236 covariation identifies new members of a protein network required for *Drosophila melanogaster*
1237 female post-mating responses. PLoS Genet. 2014;10(1):e1004108.
1238 86. Schnakenberg SL, Matias WR, Siegal ML. Sperm-storage defects and live birth in
1239 *Drosophila* females lacking spermathecal secretory cells. PLoS Biol. 2011;9(11):e1001192. doi:
1240 10.1371/journal.pbio.1001192.

1241 87. Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, et al. Functional
1242 evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition.
1243 Proc Natl Acad Sci U S A. 2012;109(6):2043-8. doi: 10.1073/pnas.1121327109.
1244 88. Civetta A, Finn S. Do candidate genes mediating conspecific sperm precedence affect
1245 sperm competitive ability within species? A test case in *Drosophila*. G3 (Bethesda).
1246 2014;4(9):1701-7. doi: 10.1534/g3.114.012476.
1247 89. Wong A, Albright SN, Giebel JD, Ram KR, Ji S, Fiumera AC, et al. A role for Acp29AB, a
1248 predicted seminal fluid lectin, in female sperm storage in *Drosophila melanogaster*. Genetics.
1249 2008;180(2):921-31. doi: 10.1534/genetics.108.092106.
1250 90. Larkin A, Marygold SJ, Antonazzo G, Attrill H, Gilberto, Garapati PV, et al. FlyBase:
1251 updates to the *Drosophila melanogaster* knowledge base. Nucleic Acids Res.
1252 2021;49(D1):D899-D907. doi: 10.1093/nar/gkaa1026.
1253 91. Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, et al. Paucity of genes
1254 on the *Drosophila* X chromosome showing male-biased expression. Science.
1255 2003;299(5607):697-700. doi: 10.1126/science.1079190.
1256 92. Parisi M, Nuttall R, Edwards P, Minor J, Naiman D, Lü J, et al. A survey of ovary-, testis-,
1257 and soma-biased gene expression in *Drosophila melanogaster* adults. Genome Biol.
1258 2004;5(6):R40. doi: 10.1186/gb-2004-5-6-r40.
1259 93. Vibranovski MD, Zhang Y, Long M. General gene movement off the X chromosome in
1260 the *Drosophila* genus. Genome Res. 2009;19(5):897-903. doi: 10.1101/gr.088609.108.
1261 94. Emerson JJ, Kaessmann H, Betrán E, Long M. Extensive gene traffic on the mammalian
1262 X chromosome. Science. 2004;303(5657):537-40. doi: 10.1126/science.1090042.
1263 95. Reinke V, San Gil I, Ward S, Kazmer K. Genome-wide germline-enriched and sex-
1264 biased expression profiles in *Caenorhabditis elegans*. Development. 2004;131(2):311-23. doi:
1265 10.1242/dev.00914.
1266 96. Argyridou E, Huylmans AK, Königer A, Parsch J. X-linkage is not a general inhibitor of
1267 tissue-specific gene expression in *Drosophila melanogaster*. Heredity (Edinb). 2017;119(1):27-
1268 34. doi: 10.1038/hdy.2017.12.
1269 97. Kemkemer C, Catalán A, Parsch J. 'Escaping' the X chromosome leads to increased
1270 gene expression in the male germline of *Drosophila melanogaster*. Heredity (Edinb).
1271 2014;112(2):149-55. doi: 10.1038/hdy.2013.86.
1272 98. Argyridou E, Parsch J. Regulation of the X chromosome in the germline and soma of
1273 *Drosophila melanogaster* males. Genes (Basel). 2018;9(5):242. doi: 10.3390/genes9050242.
1274 99. Vibranovski MD, Lopes HF, Karr TL, Long M. Stage-specific expression profiling of
1275 *Drosophila* spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic
1276 relocation of testis-expressed genes. PLoS Genet. 2009;5(11):e1000731. doi:
1277 10.1371/journal.pgen.1000731.
1278 100. Zhang YE, Vibranovski MD, Krinsky BH, Long M. Age-dependent chromosomal
1279 distribution of male-biased genes in *Drosophila*. Genome Res. 2010;20(11):1526-33. doi:
1280 10.1101/gr.107334.110.
1281 101. Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. Chromosomal
1282 redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the
1283 X chromosome. PLOS Biol. 2010;8(10):e1000494. doi: 10.1371/journal.pbio.1000494.
1284 102. Gao G, Vibranovski MD, Zhang L, Li Z, Liu M, Zhang YE, et al. A long-term
1285 demasculinization of X-linked intergenic noncoding RNAs in *Drosophila melanogaster*. Genome
1286 Res. 2014;24(4):629-38. doi: 10.1101/gr.165837.113.
1287 103. Vibranovski MD, Zhang YE, Kemkemer C, Lopes HF, Karr TL, Long M. Re-analysis of
1288 the larval testis data on meiotic sex chromosome inactivation revealed evidence for tissue-
1289 specific gene expression related to the drosophila X chromosome. BMC Biol. 2012;10(1):49.
1290 doi: 10.1186/1741-7007-10-49.

1291 104. Mahadevaraju S, Fear JM, Akeju M, Galletta BJ, Pinheiro MMLS, Avelino CC, et al.
1292 Dynamic sex chromosome expression in *Drosophila* male germ cells. Nat Commun. 2021;12(1).
1293 doi: 10.1038/s41467-021-20897-y.
1294 105. Mikhaylova LM, Nurminsky DI. Lack of global meiotic sex chromosome inactivation, and
1295 paucity of tissue-specific gene expression on the *Drosophila* X chromosome. BMC Biol.
1296 2011;9(1):29. doi: 10.1186/1741-7007-9-29.
1297 106. Landeen EL, Muirhead CA, Wright L, Meiklejohn CD, Presgraves DC. Sex chromosome-
1298 wide transcriptional suppression and compensatory *cis*-regulatory evolution mediate gene
1299 expression in the *Drosophila* male germline. PLoS Biol. 2016;14(7):e1002499. doi:
1300 10.1371/journal.pbio.1002499.
1301 107. Meiklejohn CD, Landeen EL, Cook JM, Kingan SB, Presgraves DC. Sex chromosome-
1302 specific regulation in the *Drosophila* male germline but little evidence for chromosomal dosage
1303 compensation or meiotic inactivation. PLoS Biol. 2011;9(8):e1001126. doi:
1304 10.1371/journal.pbio.1001126.
1305 108. Vibranovski MD. Meiotic Sex Chromosome Inactivation in *Drosophila*. J Genomics.
1306 2014;2:104-17. doi: 10.7150/jgen.8178.
1307 109. Wen K, Yang L, Xiong T, Di C, Ma D, Wu M, et al. Critical roles of long noncoding RNAs
1308 in *Drosophila* Spermatogenesis. Genome Res. 2016;26(9):1233-44. doi:
1309 10.1101/gr.199547.115.
1310 110. Bouska MJ, Bai H. Long noncoding RNA regulation of spermatogenesis via the spectrin
1311 cytoskeleton in *Drosophila*. G3 (Bethesda). 2021. doi: 10.1093/g3journal/jkab080.
1312 111. Lim C, Tarayrah L, Chen X. Transcriptional regulation during Drosophila
1313 spermatogenesis. Spermatogenesis. 2012;2(3):158-66. doi: 10.4161/spmg.21775.
1314 112. White-Cooper H. Molecular mechanisms of gene regulation during Drosophila
1315 spermatogenesis. Reproduction. 2010;139(1):11-21. doi: 10.1530/rep-09-0083.
1316 113. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome
1317 evolution. Mol Biol Evol. 2015;32(1):258-67. doi: 10.1093/molbev/msu286.
1318 114. Moyers BA, Zhang J. Toward reducing phylostratigraphic errors and biases. Genome
1319 Biol Evol. 2018;10(8):2037-48. doi: 10.1093/gbe/evy161.
1320 115. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, et al. Diversity and
1321 dynamics of the *Drosophila* transcriptome. Nature. 2014;512(7515):393-9. doi:
1322 10.1038/nature12962.
1323 116. White-Cooper H. Tissue, cell type and stage-specific ectopic gene expression and RNAi
1324 induction in the Drosophila testis. Spermatogenesis. 2012;2(1):11-22. doi: 10.4161/spmg.19088.
1325 117. Dietzl G, Chen D, Schnorrer F, Su K-C, Barinova Y, Fellner M, et al. A genome-wide
1326 transgenic RNAi library for conditional gene inactivation in *Drosophila*. Nature.
1327 2007;448(7150):151-6. doi: 10.1038/nature05954.
1328 118. Ge DT, Tipping C, Brodsky MH, Zamore PD. Rapid screening for CRISPR-directed
1329 editing of the *Drosophila* genome using *white* coconversion. G3 (Bethesda). 2016;6(10):3197-
1330 206. doi: 10.1534/g3.116.032557.
1331 119. Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, et al.
1332 Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in *Drosophila*.
1333 Genetics. 2014;196(4):961-71. doi: 10.1534/genetics.113.160713.
1334 120. White-Cooper H. Spermatogenesis: analysis of meiosis and morphogenesis. In:
1335 Henderson D, editor. Drosophila Cytogenetics Protocols. Methods Mol Biol. 247. Totowa, NJ:
1336 Humana Press; 2004. p. 45-75.
1337 121. Clarkson M, Saint R. A His2AvDGFP fusion gene complements a lethal His2AvD mutant
1338 allele and provides an in vivo marker for Drosophila chromosome behavior. DNA Cell Biol.
1339 1999;18(6):457-62. doi: 10.1089/104454999315178.

1340 122. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of
1341 genes and genomes on the Drosophila phylogeny. Nature. 2007;450(7167):203-18. doi:
1342 10.1038/nature06341.
1343 123. Chen Z-X, Sturgill D, Qu J, Jiang H, Park S, Boley N, et al. Comparative validation of the
1344 *D. melanogaster* modENCODE transcriptome annotation. Genome Res. 2014;24(7):1209-23.
1345 doi: 10.1101/gr.159384.113.
1346 124. Slater G, Birney E. Automated generation of heuristics for biological sequence
1347 comparison. BMC Bioinformatics. 2005;6(1):31. doi: 10.1186/1471-2105-6-31.
1348