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 show 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.

Introduction

The evolution of new genes is integral to the extensive genotypic and phenotypic diversity observed across species. The best-characterized mechanism of novel gene emergence is gene duplication (Zhang 2003; Lipinski et al. 2011); however, rapid expansion in high-quality genomic resources has provided mounting evidence of lineage-specific sequences and the existence of alternative modes of new gene origination. One such mechanism is *de novo* evolution, the birth of new genes from previously non-genic or intronic regions, which is now a widely acknowledged source of protein-coding and RNA genes (McLysaght and Hurst 2016; Bornberg-Bauer and Schmitz 2017; Van Oss and Carvunis 2019). Although *de novo* origination was once considered an unlikely event, catalogs of *de novo* genes have now been published for an expansive range of species (Zhao et al. 2014; Guerzoni and McLysaght 2016; Li et al. 2016; Ruiz-Orera et al. 2016; Lu et al. 2017; Zhang et al. 2019; Chamakura et al. 2020; Puntambekar et al. 2020). Multiple models explain how protein-coding *de novo* genes may acquire both an open reading frame (ORF) and regulatory sequences permitting transcription (Carvunis et al. 2012; Schlötterer 2015; Wilson et al. 2017; Schmitz et al. 2018). Interrogation of the biochemical and biophysical properties of the proteins encoded by *de novo* genes has offered initial insight into the mechanisms of emergence and functional potential of these genes (Schmitz et al. 2018; Vakirlis et al. 2018; Heames et al. 2020; Lange et al. 2021).

The capacity of protein-coding *de novo* genes to evolve important functions is a topic of interest from evolutionary, physiological and molecular perspectives (Keeling et al. 2019). In the last couple of decades, the products of *de novo* genes have been shown to play diverse roles in a variety of organisms. For example, *de novo* genes function in fundamental molecular processes in yeast, such as *BSC4*, a gene implicated in DNA repair, and *MDF1,* which mediates crosstalk between reproduction and growth (Cai et al. 2008; Li et al. 2014). *De novo* genes also evolve roles in organismal responses to disease and changing environmental factors. A putatively *de novo* evolved gene in rice regulates the plant's pathogen resistance response to strains causing bacterial blight (Xiao et al. 2009). Antifreeze glycoprotein genes, essential for survival in frigid ocean temperatures, evolved *de novo* in the ancestor of Arctic codfishes to coincide with cooling oceans in the Northern Hemisphere (Baalsrud et al. 2018; Zhuang et al. 2019). *De novo* genes are additionally implicated in the development and physiology of mammals. In house mice, a *de novo* evolved gene expressed in the oviduct functions in female fertility by regulating pregnancy cycles (Xie et al. 2019). A *de novo* gene found in humans and chimpanzees regulates the oncogenesis and growth of neuroblastoma, revealing the relevance of novel genes to human disease (Suenaga et al. 2014). These studies have started to demonstrate the significance of *de novo* genes, thereby challenging previous assumptions that only ancient, highly conserved genes can be essential.

Across multicellular animals, male reproductive tissues serve as hubs for new gene emergence via numerous mechanisms, including *de novo* evolution (Marques et al. 2005; Levine et al. 2006; Begun et al. 2007; Kaessmann 2010; Baker et al. 2012; Cui et al. 2015; Ruiz-Orera et al. 2016). Proposed causes of this "out of the testis" phenomenon include the high level of promiscuous transcription in testis cells (Soumillon et al. 2013; Necsulea and Kaessmann 2014), the relative simplicity of promoter regions driving expression in the testis (Sorourian et al. 2014), and preferential retention of novel genes with male-biased expression (Palmieri et al. 2014). Sexual selection also drives rapid evolution of reproductive proteins (Wilburn and Swanson 2016) and could drive the emergence of new genes as a mechanism of improving male reproductive ability (Levine et al. 2006). The testis-biased expression of novel genes, combined with growing evidence for new genes acting across a variety of tissue contexts, suggests that many novel genes may function in male reproduction. For example, a pair of young duplicate genes in *Drosophila, apollo* and *artemis,* are essential for male and female fertility, respectively (VanKuren and Long 2018). Continued efforts to identify and

characterize testis-expressed novel genes are consequently critical for understanding the genetic basis of male reproductive phenotypes.

Drosophila serves as an ideal system for interrogating the prevalence, sequence attributes, expression patterns, and functions of testis-expressed *de novo* genes. The availability of well-annotated genomes for numerous *Drosophila* species, the tractability of flies to molecular genetics techniques, and our thorough understanding of *Drosophila* reproductive processes facilitate comprehensive analyses of novel fly reproductive proteins (Demarco et al. 2014; Hales et al. 2015). As observed in other biological systems, *Drosophila de novo* genes retained by selection demonstrate enriched expression in the testis (Levine et al. 2006; Begun et al. 2007; Zhao et al. 2014; Heames et al. 2020). The expression patterns of emerging *de novo* genes in the *Drosophila* testis were recently analyzed at single cell resolution (Witt et al. 2019), thereby providing insight into the dynamics of novel gene expression throughout spermatogenesis. In addition to bioinformatic screens that have started to identify *de novo* genes and large-scale expression analyses of testis-expressed genes, RNAi (Reinhardt et al. 2013) and CRISPR/Cas9-based (Kondo et al. 2017) functional screens have identified putative, testis-expressed *de novo* genes required for fertility. However, a need remains for in-depth experimental and evolutionary characterization of the genes identified in such screens. Detailed examination of the function of *de novo* proteins will enable us to understand how these proteins might integrate into existing gene networks and become essential.

We previously conducted a pilot functional screen of *de novo* genes with testis-enriched expression in *D. melanogaster* and identified two novel genes, *goddard* and *saturn*, that are required for full fertility (Gubala et al. 2017). *Goddard* knockdown males failed to produce any sperm. *Saturn* knockdown males produced fewer sperm, which were inefficient at migrating to female sperm storage organs. Subsequent characterization of Goddard using null deletion alleles and a biochemically tagged rescue construct showed that the protein localizes to elongating sperm axonemes and that, in its absence, individualization complexes associate less efficiently with spermatid nuclei and do not successfully progress along sperm tails (Lange et al. 2021). These data suggested that putative *de novo* genes can evolve essential roles in a rapidly evolving reproductive process, spermatogenesis.

Here, we expand this functional screen by evaluating whether any of 42 putative *de novo* genes that show testis-enriched expression in *D. melanogaster* are required for male fertility. We identified one gene, which we named *atlas,* whose knockdown or knockout results in nearly complete male sterility. We show that *atlas* encodes a transition protein that facilitates spermatid chromatin condensation. The *atlas* gene in *D. melanogaster* arose when a likely *de novo* evolved protein-coding sequence moved off of the X chromosome and was inserted upstream of a well-conserved non-coding RNA. While the *atlas* protein-coding sequence has undergone multiple, independent gene loss events since its apparent origin at the base of the *Drosophila* genus, the gene has evolved a critical function in *D. melanogaster*. These results underscore the importance of detailed functional and evolutionary characterization in understanding the origins of new protein-coding genes and the selective forces that affect their subsequent evolution.

Results

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

A previous pilot screen of 11 putative *de novo* evolved, testis-expressed genes identified two genes that are critical for male fertility in *Drosophila melanogaster* (Gubala et al. 2017). This result, and other recent work (e.g., Abrusán 2013; Zhang et al. 2015; VanKuren and Long 2018), suggested that lineage-specific, newly evolved genes can rapidly become important for fertility, perhaps by gaining interactions with existing protein networks. To determine more comprehensively the frequency with which potential *de novo* evolved genes become essential for fertility, we identified *de novo* or putative *de novo* evolved genes with testis-biased expression. A previous computational analysis identified genes that are detectable only within the *Drosophila* genus, lack identifiable protein domains, and show no homology to other known proteins through BLASTP and TBLASTN searches (Heames et al. 2020). We filtered these genes to identify those expressed exclusively or predominantly in the testis, a common site of *de novo* gene expression in animal species (Begun et al. 2007; Wu et al. 2011; Palmieri et al. 2014; Xie et al. 2019). This resulted in a set of 96 target genes.

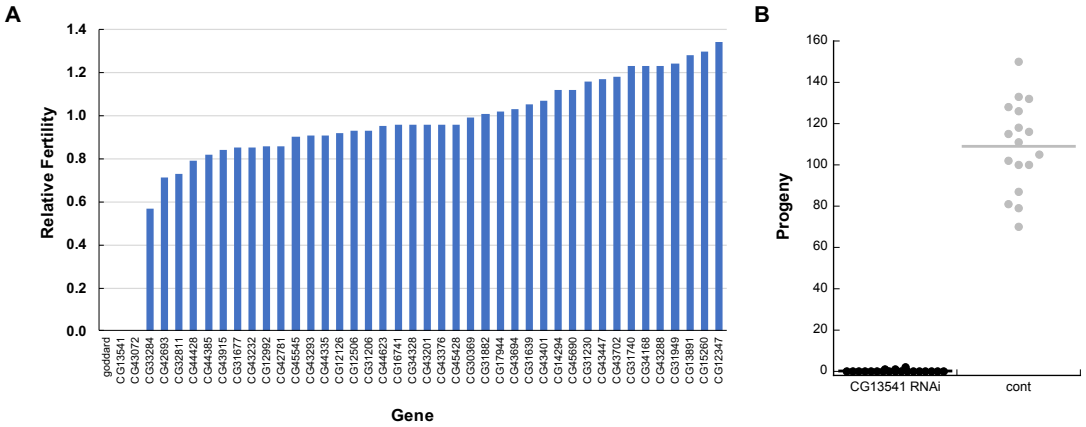
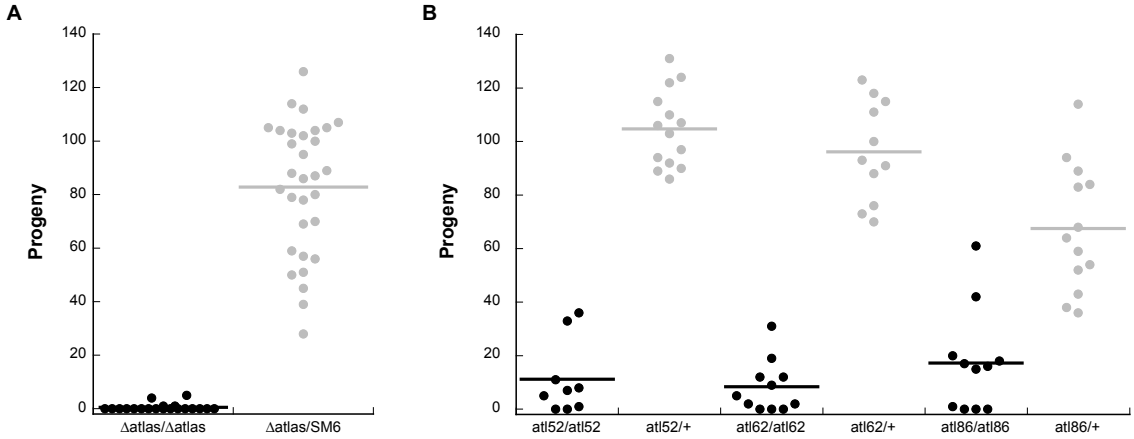


Figure 1. An RNAi screen of putative *de novo* genes identifies *CG13541* as a major contributor to *Drosophila melanogaster* male fertility. A) All RNAi lines that showed at least partial knockdown of the target gene were screened in group fertility assays (see Materials and Methods). Relative fertility was calculated by dividing the average number of progeny produced per female mated to knockdown males by the average number of progeny produced per female mated to control males in a contemporaneous experiment. Relative fertility measurements lack error bars because each gene was tested in only 1-2 replicates. Knockdown of *goddard* was used as a positive control. B) A single-mating, single-pair fertility assay confirms the observed defect when males are knocked down for *CG13541*, as knockdown males showed significantly reduced fertility (control fertility (mean ± SEM): 109.0 ± 5.3; knockdown fertility: 0.2 ± 0.1; two-sample *t*-test assuming unequal variances, $p = 5.6 \times 10^{-13}$).

We used testis-specific RNA interference to screen these genes for roles in male fertility. We obtained RNAi lines from the Vienna *Drosophila* Resource Center (VDRC) and the Transgenic RNAi Project (TRiP) and constructed additional lines using the TRiP-style pValium20 vector (Ni et al. 2011), which is optimized for male germline expression. We obtained and tested an RNAi line for 57 genes and induced knockdown for each with the *Bam*-GAL4 driver, which is expressed in the male germline, and enhanced with a copy of UAS-*Dicer2*. RT-PCR confirmed at least partial knockdown in lines representing 42 genes (see example in Fig. S1). We then screened knockdown males for fertility by allowing groups of 7 knockdown males to mate with 5 wild-type females for 2 days. Progeny counts were standardized to the number of progeny produced by concurrently mated groups of 7 control

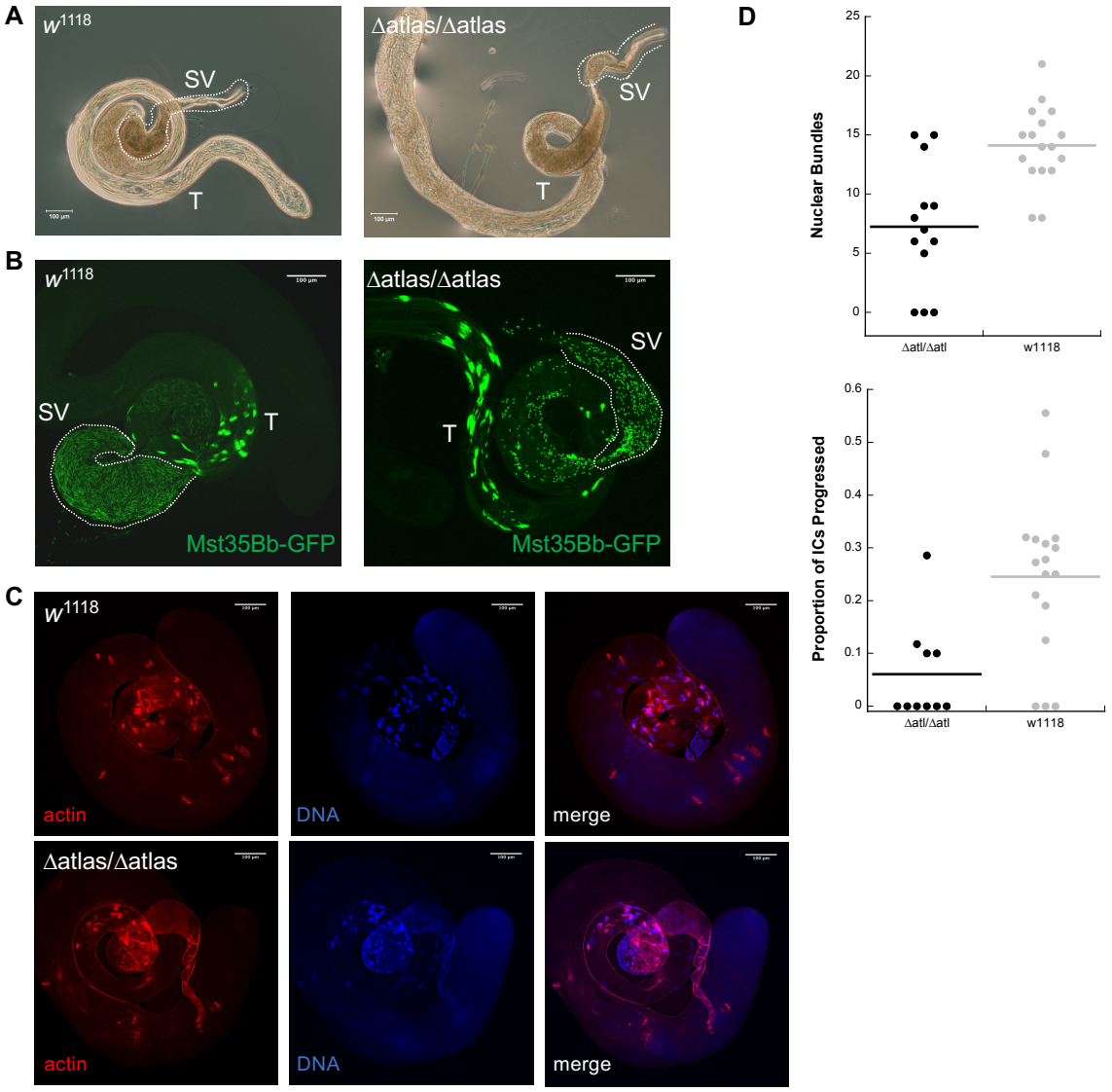
177 males and 5 wild-type females. The results are shown in Fig. 1A. This initial screen identified
178 *CG13541*, whose knockdown severely reduced male fertility. We confirmed the result for
179 *CG13541* by performing single-pair mating fertility assays (Fig. 1B). Knockdown of *CG43072*
180 and *CG33284* also showed consistent fertility defects, but these results did not replicate upon
181 testing with CRISPR-generated null alleles for each gene, so we do not consider them further
182 here. Consistent with our previous convention of naming testis-expressed genes after American
183 rocketry (Gubala et al. 2017), we will from here on refer to *CG13541* as *atlas*.
184
185 *CRISPR-mediated gene mutation validates* atlas *RNAi results*
186
187 We validated the observed fertility defect by using CRISPR/Cas9-based genome editing
188 to construct putative loss-of-function alleles for *atlas* (Fig. S2). The principal allele we used for
189 validation and further functional studies (described below) was a null allele that completely
190 deleted *atlas* from the genome. This allele was generated by targeting each end of the locus
191 with a gRNA. We made three additional frameshift alleles by inducing double-stranded breaks
192 at a gRNA target site just downstream of the *atlas* start codon, which induced non-homologous
193 end joining. Males homozygous for the *atlas* deletion allele have the same fertility defect as
194 knockdown males (Fig. 2A). Males homozygous for any of three frameshift alleles showed
195 significantly reduced, but non-zero, fertility (Fig. 2B). It is possible that residual *atlas* function
196 may be present in these animals, perhaps due to translation initiation at a downstream start
197 codon to create a shorter protein with partial function. Finally, we constructed a genomic rescue
198 construct carrying both the *atlas* coding region and its native regulatory sequences. *Atlas* null
199 males that carried a single copy of the rescue construct had fully restored fertility (Fig. S3).
200 Overall, these data demonstrate that *atlas* loss, and not an RNAi or CRISPR off-target, causes
201 nearly complete male sterility.
202



203
204
205 **Figure 2. CRISPR-generated deletion and frameshift alleles of *atlas* confirm the gene's**
206 **requirement for male fertility.** A) Single-pair fertility assay for males homozygous for the null (Δ*atlas*)
207 allele or heterozygous controls (Δ*atlas*/SM6). Flies homozygous for the deletion had significantly reduced
208 fertility (control fertility: 82.9 ± 4.5; null fertility: 0.3 ± 0.6; two-sample t-test assuming unequal variances, p
209 $= 5.4 \times 10^{-18}$). B) Single-pair fertility assays for males homozygous or heterozygous for three frameshift
210 alleles of *atlas* generated by imprecise non-homologous end joining at a CRSIRP/Cas9 target site just
211 downstream of the start codon: $atlas^{52}$ (control fertility: 104.7 ± 3.8, mutant fertility: 11.2 ± 4.6; two-sample
212 *t*-test assuming unequal variances: $p = 8 \times 10^{-12}$), $atlas^{62}$ (control fertility: 96.2 ± 5.7; mutant fertility: 8.4 ±
213 2.9; two-sample *t*-test assuming unequal variances: $p = 6.1 \times 10^{-10}$) and $atlas^{86}$ (control fertility: 67.5 ± 6.6;
214 mutant fertility: 17.3 ± 5.8; two-sample *t*-test assuming unequal variances: $p = 9.5 \times 10^{-6}$).
215

216 Atlas *is required for proper spermatid nuclear condensation*
217
218 We next examined how *atlas* loss of function impacted male fertility at the cellular level.
219 Dissection and phase-contrast imaging of *atlas* deletion null or knockdown male reproductive
220 tracts revealed that while the pre-meiotic and meiotic stages of spermatogenesis appeared
221 normal, sperm accumulated at the basal end of the testes, rather than in the seminal vesicles
222 (SVs), over the first week of adulthood (Fig. 3A and Fig. S4). To further characterize the fertility
223 defects in the absence of *atlas*, we examined the Mst35Bb-GFP ("protamine"-GFP) marker in
224 null or knockdown backgrounds (Manier et al. 2010). Mst35Bb encodes one of two protamine-
225 like proteins (highly similar paralogs of each other) that bind DNA in mature sperm. Its GFP
226 fusion construct thus allows visualization of nuclei in late stage spermiogenesis and mature
227 sperm. Consistent with the observed conglomeration of sperm tails at the basal testes, SVs
228 from either *atlas* null or knockdown males contain fewer mature sperm (Fig. 3B, S4C). The
229 nuclei of sperm from null males also appeared wider and less elongated than those of controls.
230 Together, these data suggest that *atlas* is required after meiosis, as developing spermatids take
231 on their final structures.
232 We next examined two post-meiotic processes: individualization of 64-cell spermatid
233 cysts into mature sperm, and spermatid nuclear condensation. Individualization initiates when
234 an actin-rich individualization complex (IC) associates with the bundle of spermatid nuclei. The
235 IC then proceeds down the sperm tails, expelling cytoplasmic waste and remodeling cell
236 membranes to form 64 individual sperm. We visualized this process in males 0-1 days old,
237 when spermatogenesis occurs at high levels, by staining whole mount testes for actin (Fig. 3C-
238 D). Although ICs could associate with nuclear bundles present at the basal end of the testes in
239 both control and *atlas* null males, we observed significantly fewer nuclear bundle-associated ICs
240 in nulls (Fig. 3C-D). While control testes typically had several ICs progressing down sperm tails,
241 we saw a significantly reduced proportion of progressed bundles in nulls (Fig. 3C-D). In some
242 null testes, we also observed individual investment cones dissociated away from progressing
243 ICs (Fig. 3C).
244 The ability of ICs to assemble at nuclear bundles and progress down sperm tails may be
245 reduced if nuclear condensation is aberrant (reviewed in Steinhauer 2015). During *Drosophila*
246 spermiogenesis, round spermatid nuclei undergo a series of stepwise, morphological changes
247 that are the product of two distinct, but related processes: changes in the chromatin packaging
248 of DNA, and changes in nuclear shape (Rathke et al. 2007; Fabian and Brill 2012; Rathke et al.
249 2014). The end result is thin, condensed nuclei. We quantified this process in testes dissected
250 from newly eclosed wild-type and *atlas* null males expressing Mst35Bb-GFP, which marks the
251 final stages of condensation. We shredded the post-meiotic region of the testes in the presence
252 of a fixative and counted the number of nuclear bundles that exhibited each of five stages of
253 condensation (Rathke et al. 2007): round nuclei, early canoe-stage (unmarked with Mst35Bb-
254 GFP), late canoe-stage (marked with Mst35Bb-GFP), elongated nuclei, and fully condensed
255 nuclei (Fig. 4). Condensation of the nuclear bundles in *atlas* null testes progressed at similar
256 rates to controls through the late canoe stage (Table S1). However, in *atlas* null males, all
257 nuclear bundles that progressed past the canoe stage (which included ~60% [range: 26-100%]
258 of all observed bundles) showed an aberrant "curled" phenotype (Fig. 4; Table S1). These data
259 suggest that Atlas protein is required during the later stages of nuclear condensation and are
260 consistent with the idea that the loss of *atlas* affects nuclear shape in a way that reduces IC
261 assembly and sperm individualization (see Fig. 3C).
262

263



264
265 **Figure 3. Cytological investigations of the *atlas* mutant fertility defect.** A) Phase contrast
266 microscopy of male reproductive tracts dissected from 7-day-old, unmated control (w^{1118}) or *atlas* null
267 males. Control males show the expected accumulation of sperm in the seminal vesicle (SV), which
268 appears here as a darker brown shading, while null male have an aberrant accumulation of sperm tails at
269 the basal end of the testis (T). B) Visualization of Mst35Bb-GFP in 4-day old control and *atlas* null testes.
270 While Mst35Bb is expressed in spermatid nuclei in the absence of *atlas*, the nuclei appear shorter and
271 much less numerous in the outlined SV. C) Representative images from phalloidin staining of w^{1118} and
272 *atlas* null testes used to assess the association of individualization complexes (ICs) with nuclear bundles
273 and the progression of ICs down the length of sperm tails. D) At top, number of nuclear bundles with ICs
274 associated in control and *atlas* null testes. Significantly more ICs were observed in control testes (control:
275 $N = 17$, median = 14; mutant: $N = 13$, median = 7; Wilcoxon rank-sum test W = 34, $p = 0.0014$). At
276 bottom, proportion of all observed ICs that were intact and that had progressed away from nuclear
277 bundles. Three mutant testes with no observed ICs were excluded from the analysis. A significantly
278 higher proportion of ICs progressed in control testes (control: $N = 17$, median = 0.27; control: $N = 10$,
279 median = 0; Wilcoxon rank-sum test W = 28, $p = 0.0038$).
280
281

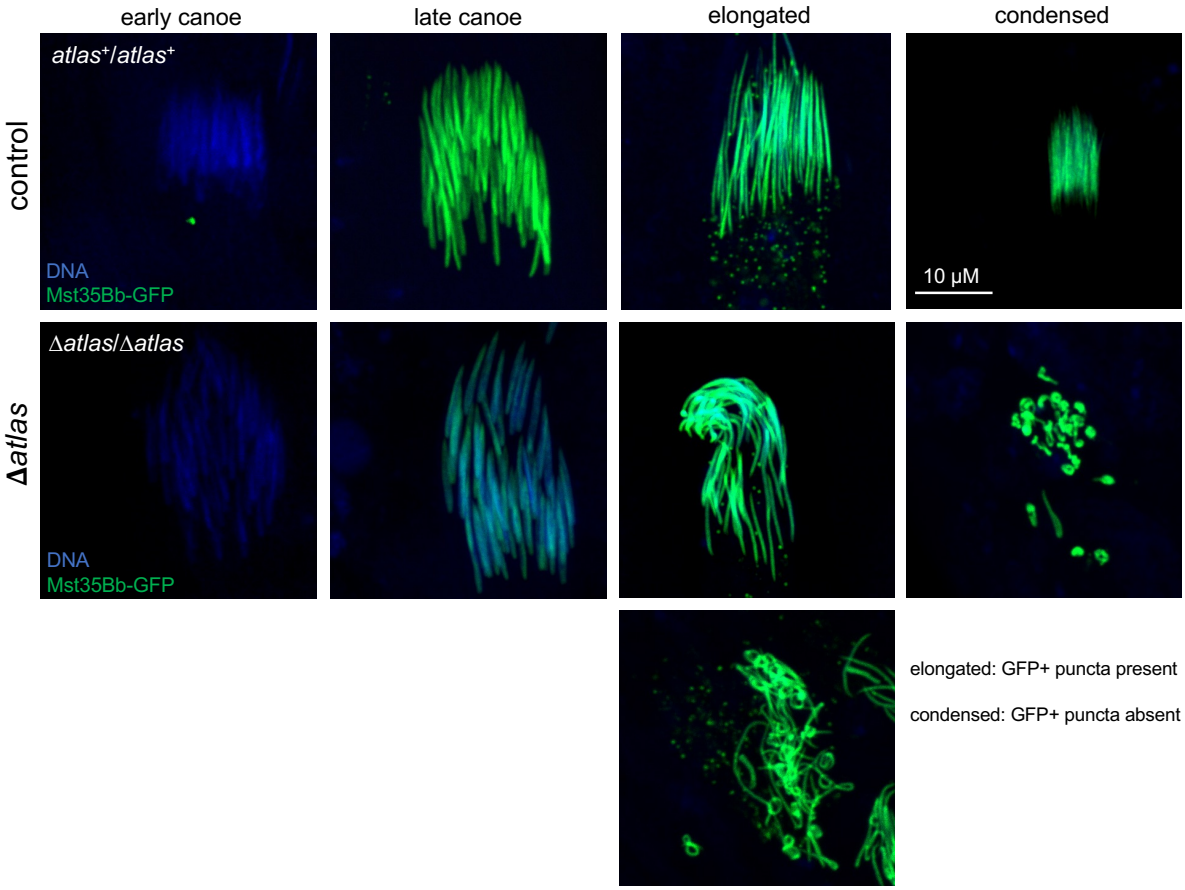


Figure 4. *Atlas* null males show aberrant nuclear shaping at and beyond the elongated stage of spermatid nuclear condensation. Early and late canoe stages were distinguished by the absence or presence of Mst35Bb-GFP, respectively. Elongated and condensed stages were distinguished by the presence or absence of GFP-positive puncta, respectively. 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 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 (Jayaramaiah Raja and Renkawitz-Pohl 2005; Rathke et al. 2007; Kanippayoor et al. 2013), 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 (Dorus et al. 2008; Alvi et al. 2013; Rathke et al. 2014; Eren-Ghiani et al. 2015; Gärtner et al. 2015; Alvi et al. 2016; Gärtner et al. 2019), suggesting that this variety of chromatin binding proteins could have originated through gene duplication and divergence. Consistent with its putative *de novo* origin, Atlas lacks a detectable HMG-box domain. However, Atlas is otherwise similar to these other sperm chromatin binding proteins: the ~20 kDa protein has a highly basic predicted isoelectric point of 10.7, and its primary sequence contains the sequence KRDK, which matches the canonical consensus sequence for nuclear import, K(K/R)X(K/R) (Lange et al. 2007). To test the hypothesis that Atlas is nuclear localized, and could thus bind DNA, we generated an *atlas-GFP* transgene under UAS control

306 and expressed it ubiquitously using *tubulin*-GAL4 and in the early male germline using *Bam*-
307 GAL4. In both larval salivary glands and early male germline cells, atlas-GFP appeared to be
308 nuclear localized (Fig. S5).
309 While these results were consistent with Atlas protein localizing to the nucleus, they did
310 not allow us to visualize Atlas in the cells in which it is normally expressed. To do so, we used
311 CRISPR/Cas9-induced homology directed repair (https://flycrispr.org/scarless-gene-editing/)
312 (Bruckner et al. 2017; Bier et al. 2018; Hill et al. 2019) to create an *atlas*-GFP fusion at the
313 endogenous *atlas* locus (see Fig. S6 and Materials and Methods). We first confirmed the
314 functionality of the knock-in allele by showing that males with the *atlas* locus genotype *atlas*-
315 GFP/Δ*atlas* had equivalent fertility to males of genotype *atlas*+/Δ*atlas* (Fig. 5A). We then
316 visualized Atlas-GFP fusion protein in whole-mount testes in conjunction with phalloidin-stained
317 actin (Fig. 5C). Atlas-GFP was absent from seminal vesicles, consistent with its absence from
318 the proteome of mature *D. melanogaster* sperm (Dorus et al. 2006; Wasbrough et al. 2010).
319 Instead, Atlas-GFP colocalized with condensing nuclear bundles near the basal end of the
320 testes (Fig. 5C). Actin-based ICs were also observed in the basal testes, but generally did not
321 co-localize with Atlas-GFP, suggesting that Atlas-GFP is present in condensing nuclei before IC
322 association (Fig. 5C). This result, taken together with the aberrant nuclear condensation in the
323 absence of *atlas* (Fig. 4), is consistent with the idea that Atlas is a transition protein. Transition
324 proteins are chromatin components that act transiently during spermatid nuclear condensation.
325 A series of transition proteins first replace histones as the primary DNA binding proteins in the
326 nucleus and then give way to protamines, the proteins that package chromatin in mature sperm
327 (Rathke et al. 2007; Rathke et al. 2014; Gärtner et al. 2015).
328 To further elucidate the role of *atlas* in nuclear condensation, we next examined Atlas-
329 GFP localization in the presence of either an early spermatid nuclear marker, histone H2Av-
330 RFP (Schuh et al. 2007; Rathke et al. 2014), or Mst35Bb-dsRed (Manier et al. 2010), a marker
331 of nuclei from the late canoe stage through final condensation. Atlas-GFP showed no co-
332 localization with H2Av-RFP, suggesting that Atlas functions after histone removal (Fig. 5B). In
333 contrast, some GFP-positive bundles co-localized with Mst35Bb-dsRed, but others did not (Fig.
334 5D). These data suggest that Atlas may be one of the final transition proteins used in nuclear
335 condensation before the chromatin becomes fully condensed with protamines.
336

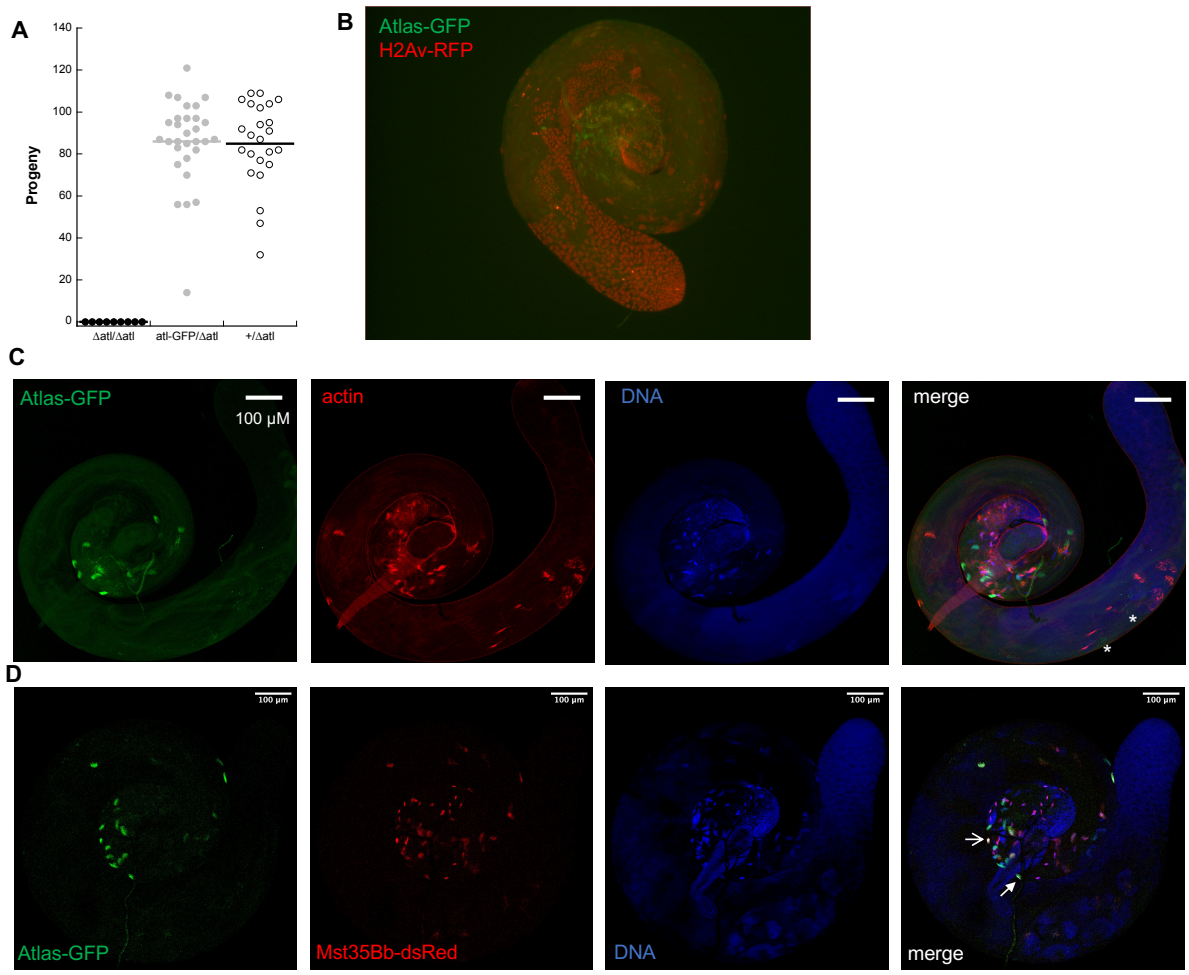


Figure 5. An *atlas*-GFP allele generated at the endogenous *atlas* locus rescues the fertility defect of *atlas* null flies and encodes a protein that localizes to condensing spermatid nuclei. A) A single copy of the *atlas*-GFP allele completely rescues the fertility defect caused by the Δ*atlas* allele and shows equivalent fertility 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 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. Some Atlas-GFP is observed near progressing individualization complexes toward the apical testis in the merged image (marked with asterisks; see also Fig. 6D). D) Atlas-GFP partially colocalizes with Mst35Bb-dsRed, a marker of the final stage of nuclear condensation. 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.

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 day of eclosion. Based on the stage of the defect in *atlas* null males (Fig. 3-4) and the pattern of Atlas-GFP-positive bundles in whole-mount testes (Fig. 5), we hypothesized that Atlas-GFP would localize to the later stages of nuclear condensation. Consistent with this hypothesis, we

did not detect Atlas-GFP in round or early canoe stage bundles (Fig. 6A-B). Atlas-GFP co-localized with DNA in late canoe stage bundles (Fig. 6C). Interestingly, when nuclei elongated further, GFP was detected not in the nucleus, but as puncta basal to the nuclei (Fig. 6D; see also Fig. 5C). Since Atlas-GFP is not observed in mature sperm in the SV (Fig. 5C), these data suggest that Atlas may function as a transition protein that facilitates the condensation of spermatid nuclei from histone-based DNA packaging to protamine-like-based DNA packaging (Rathke et al. 2007) and is then removed from nuclei once protamines bind DNA. Indeed, the appearance of Atlas in nuclei during the late canoe stage of condensation is similar to the pattern observed for a previously characterized transition protein, Tpl94D (Rathke et al. 2007). We hypothesize that the failure of *atlas* null sperm to form needle-like nuclei can be explained by the absence of Atlas from the late canoe nucleus. It is also possible that the apparent removal of Atlas-GFP from nuclei (Fig. 6D) represents a mechanism for removing transition proteins from the nucleus after they exert their functions. We observed above that some Mst35Bb-GFP also appears to be removed from the nucleus in puncta during the elongation stage of nuclear condensation (see elongated stage of control nuclear bundles, Fig. 4), even though other Mst35Bb-GFP molecules ultimately package DNA in mature, individualized sperm. This could occur if Mst35Bb-GFP is present in excess of what is needed to package DNA.

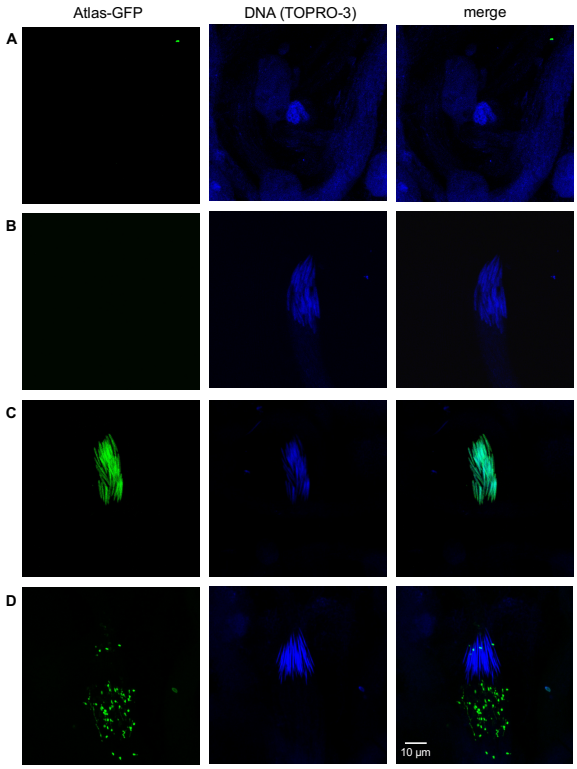
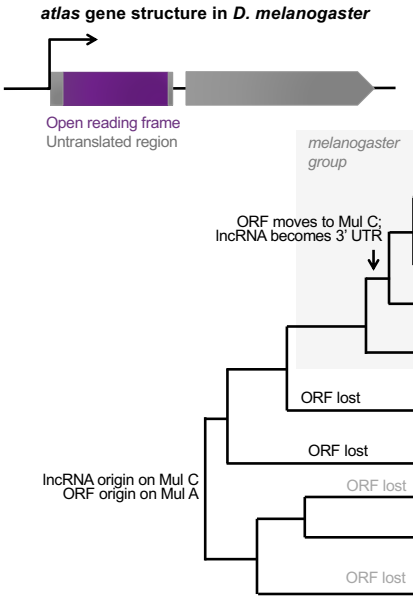


Figure 6. Atlas-GFP is present in late canoe-stage spermatid nuclei and then appears to leave the nucleus in puncta. Staging of condensing spermatid nuclei fixed in paraformaldehyde from *atlas*-GFP males stained with TO-PRO-3 DNA stain. Atlas-GFP is not detectable in (A) round stage or (B) early canoe 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.

Evolutionary origins of atlas

To better understand the evolutionary origin of *atlas* and its evolution since emergence, we used a combination of BLAST- and synteny-based approaches to identify *atlas* orthologs throughout the genus (Gubala et al. 2017; Rele et al. 2020). One notable feature of this two-exon gene is that the protein-coding region (519 nucleotides) is contained entirely within the first exon (622 nt); the longer, second exon (910 nt) appears to be entirely non-coding (Fig. 7). Surprisingly, the second exon is more widely conserved. BLASTN detected significant matches to this region (range of hit length: 185-864 nt) in the same genomic location on Muller element C (see Schaeffer et al. 2008 for explanation of Muller elements), as assessed by synteny, in all *Drosophila* species examined, including distantly related species such as *D. virilis* and *D. grimshawi* (Fig. 7 and S7). The protein-coding first exon showed a more limited phylogenetic distribution. In most members of the *melanogaster* group of *Drosophila* (gray box in Fig. 7), this exon is found in a conserved position, adjacent to the non-coding region on the equivalent of *D. melanogaster* chromosome 2R (Table S2). In *D. ananassae*, however, the protein-coding region is found on the X chromosome (Muller element A). A putative ortholog for the protein-coding sequence is detectable by BLASTP in a partially syntenic region on the same Muller element in *D. virilis* (Table S2, Fig. S8). These data suggest that the *atlas* protein-coding sequence initially arose on Muller element A and then moved to Muller element C, giving rise to the gene structure observed in extant *D. melanogaster* and its sister species.



	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. S7 and S9.

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 (Rele et al. 2020). 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 (Weisman et al. 2020); 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. S8 and Table S2. These searches detected no evidence for *atlas* orthologs in the following *Drosophila* species: *obscura*, *miranda, willistoni, hydei, arizonae, mojavensis, navojoa,* and *grimshawi*. Some of these species had unannotated, male-expressed transcripts in the regions syntenic to the Muller A location of *atlas* in *D. ananassae* and *D. virilis*, but when each was compared with BLASTP to the *D. melanogaster* proteome, all matched proteins other than Atlas, suggesting they may be lineage-specific paralogs of other genes (Fig. S8). In sister species *D. pseudoobscura* and *D. persimilis*, we detected a male-expressed transcript predicted to encode a protein with a pI > 10 in the region syntenic to the location of *atlas* in *D. virilis*, but the predicted protein sequences showed no significant BLASTP similarity to *atlas* orthologs (Table S2). While this predicted protein may represent a divergent *atlas* ortholog, the abSENSE method predicts low probabilities of BLASTP detection failure when searching for Atlas protein in these species (0.02 and 0.04, respectively), so we favor the hypothesis of a lineage-specific, newly evolved gene in the region. Conversely, in *D. busckii*, we detected, in the region syntenic to the *D. virilis atlas* locus, a male-expressed transcript predicted to encode a protein with significant BLASTP identity to *D. melanogaster* Atlas (e = 4e-8), but with a predicted pI of 5.1 and a ~50 percent shorter open-reading frame (Table S2). The ortholog status of this predicted protein is also unclear, but because of its dramatically altered size and pI, it is unlikely to have a functional role equivalent to that of *D. melanogaster* Atlas.

To investigate whether the protein-coding region may have reproductive functions in other species, we used sex-specific RNA-seq data from numerous *Drosophila* species curated by the Genomics Education Partnership (Rele et al. 2020; thegep.org) and verified several of these results by RT-PCR (Fig. 7, Fig. S9). In all species in which *atlas* was detected, the protein-coding region was expressed specifically in males regardless of its genomic location (Fig. 7). Interestingly, the non-coding region showed male-specific expression in species lacking an unambiguous, orthologous coding region, such as *D. pseudoobscura* and *D. mojavensis*. Conversely, while *D. yakuba* and *D. erecta* express the protein-coding region robustly, we found no RNA-seq evidence to support expression of the non-coding second exon, in spite of its sequence conservation (Fig. S7). Based on its high level of sequence conservation, consistent genomic location and expression in a variety of species, it is possible that what we now consider to be the 3' untranslated region of *atlas* from *D. melanogaster* was, ancestrally, a non-coding RNA.

The FlyBase database reports two transcript isoforms of *atlas* in *D. melanogaster*: the *atlas*-RA isoform is 986 nucleotides, while the *atlas*-RB isoform is 1528 nt. These isoforms differ in how much of the second, non-coding exon is included in the transcript. We used RT-PCR of whole male cDNA to assess the presence of these isoforms and their relative abundances. Primers designed to amplify a region present in both isoforms produced products that appeared more abundant than primers designed to amplify only the long isoform, even though both primer pairs appeared to amplify genomic DNA with equal efficiency. Based on RT-PCR band intensities and controlling for product size and genomic PCR band intensities, we estimated that the short isoform is about 3-fold more abundant. This difference in abundance is mirrored in available RNA-seq data, which show approximately 3- to 4-fold higher levels of expression in the upstream part of exon 2 (Fig. S7), a pattern that also appears in *D. simulans* and *D. sechellia*. Evaluating the potential significance of this finding awaits functional characterization of the non-coding region.

As we have observed for other putative *de novo* genes with essential male reproductive functions (Gubala et al. 2017), the pattern of *atlas* protein-coding sequence presence/absence across the phylogeny is difficult to explain parsimoniously. If we assume that gene birth events are less frequent than gene deaths, since the latter can occur through many possible mutational events and can happen separately along multiple phylogenetic lineages, our data support the hypothesis of a single origin of the protein-coding sequence at the base of the genus, followed by independent loss events on the lineages leading to *D. grimshawi*, *D. mojavensis* and *D. willistoni*, and potentially also *D. pseudoobscura*/*persimilis* (Table S2). We summarize these findings for 12 representative species of *Drosophila* in Fig. 7. The general patterns of loss do not change when all species of Table S2 are considered, though an additional loss in the *melanogaster* group is likely due to the absence of a detectable ortholog in *D. kikkiwai* and *D. serrata*. As noted above, the pattern of gene loss can also appear due to orthology detection failure (Weisman et al. 2020), for which we tried to account with our additional search methods described above. We also note, however, that the probability of BLASTP-based ortholog detection failure is relatively low for some *Drosophila* species that lack *atlas*, including *D. pseudoobscura* (probability of non-detection due to divergence = 0.02), *D. persimilis* (p = 0.04) and *D. willistoni* (p = 0.06). The probability is higher for other species, *D. mojavensis* (p = 0.33) and *D. grimshawi* (p = 0.66), underscoring the importance of our additional search strategies. Overall, our data support the hypothesis of multiple, independent loss events within *Drosophila*.

AbSENSE produces a 1.00 probability of BLASTP-based atlas ortholog detection failure outside of *Drosophila*, reflecting the protein's short length and relatively rapid divergence (see below). Indeed, the protein-coding and non-coding transcriptomes from each species showed no matches to Atlas protein or cDNA sequences by BLAST. We thus used another synteny-based approach, summarized in Fig. S10, to look for the protein-coding gene in other Dipterans with well-resolved genomes: *Musca domestica*, *Glossina morsitans*, *Lucilia cuprina*, *Aedes aegypti*, *Anopheles darlingi*, *Anopheles gambiae*, *Culex quinquefasciatus* and *Mayetiola destructor*. In none of these species was a putative homolog found in any potential syntenic region.

Recognizing the limitation of even this approach, we also used HMMER (Potter et al. 2018) to search iteratively either all genomes in ENSEMBL, or all metazoan genomes in ENSEMBL, for annotated proteins with identity to Atlas from *D. melanogaster* or *D. virilis*. These searches initially identified significant hits to the Atlas orthologs we identified above from other *Drosophila* species. When these collections of orthologs were used as queries, no further proteins outside of *Drosophila* were a significant match. As a control, we performed the same search strategy with *D. melanogaster* Mst35Bb, a protein whose length, amino acid composition, and function are similar to Atlas. These searches readily identified orthologs throughout Diptera, consistent with predictions of its conservation from the OrthoDB database (Zdobnov et al. 2017). Thus, we conclude that *atlas* is a putative *de novo* evolved gene that is limited phylogenetically to the *Drosophila* genus.

Finally, we used standard tests of molecular evolution to examine the selective pressures that have shaped Atlas protein within the *melanogaster* group. We aligned the *atlas* protein-coding sequences from 12 species and used PAML to ask whether a model (M8_ allowing for positive selection, as well as neutral evolution and purifying selection, explained the data better than models (M7 and M8a) that allowed only neutral evolution and purifying selection (Yang 2007; McGeary and Findlay 2020). These data showed that while the *atlas* protein-coding sequence's rate of evolution was accelerated relative to most *Drosophila* proteins (whole-gene estimated d_N/d_S, ω = 0.41 by PAML model M0), there was no significant evidence for positive selection acting to recurrently diversify a subset of sites within the protein (Table 2).

522 **Table 2. PAML sites tests for positive selection acting on atlas in the *melanogaster***
523 **group.**
524
525 Model M0 (uniform w across all sites): $\omega = 0.41$, ln L = -4032.24, np = 23
526 Model M7 (10 site classes, each with $0 \leq \omega \leq 1$): ln L = -3961.28, np = 24
527 Model M8 (10 site classes as in M7, plus one class with $\omega \geq 1$): ln L = -3960.41, np = 26, ω for
528 extra class of sites = 1.39 (9.1% of sites)
529 Model M8a (10 site classes as in M7, plus one class with $\omega = 1$): ln L = -3961.02, np = 25,
530 16.6% of sites in the $\omega = 1$ class
531
532 M7 vs. M8 likelihood ratio test: $\chi^2 = 1.74$, df = 2, $p = 0.42$
533 M8 vs. M8a likelihood ratio test: $\chi^2 = 1.22$, df = 1, $p = 0.27$
534
535
536

Discussion

Across taxa, many *de novo* evolved genes are expressed in the male reproductive system (Levine et al. 2006; Begun et al. 2007; Cui et al. 2015; Ruiz-Orera et al. 2016). Identifying those genes that have evolved essential roles in reproduction will provide insight into how newly evolved genes integrate with existing cellular networks (Abrusán 2013) and how evolutionary novelties permit adaptation in the face of sexual selection. Here, we screened 42 putatively *de novo* evolved genes for major effects on male *D. melanogaster* reproduction. Our primary screen identified three genes whose knockdown caused an apparent reduction in male fertility. However, subsequent CRISPR-mediated gene deletion revealed that only one of these genes, *atlas*, was truly essential. This result underscores the importance of validating genes identified in RNAi screens through traditional loss-of-function genetics and other approaches.

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

Several lines of evidence suggest that Atlas is a transition protein that facilitates the change from histone-based to protamine-based chromatin packaging in spermatid nuclei. Atlas localizes throughout spermatid nuclei (Fig. 6) and has biochemical properties consistent with direct DNA interaction. The protein appears specifically at the late canoe stage of nuclear compaction (Fig. 5B-D). Its lack of overlap with testis-specific histones (Fig. 5C), partial overlap with Mst35Bb (Fig. 5D), likely removal from needle-stage nuclei (Fig. 6) and absence from mature sperm (Figs. 5-6) are all consistent with the expression profile of a transition protein. Several other transition proteins have been characterized in *D. melanogaster*, including Tpl94D, thmg-1, thmg-2, and Mst84B (Rathke et al. 2007; Gärtner et al. 2015; Gärtner et al. 2019). Collectively, the transition proteins vary in the stage of nuclear condensation at which they first appear and the range of nuclear shapes over which they are found (Hundertmark et al. 2018), but otherwise match Atlas in their biochemical properties, transient roles, and localization throughout the nucleus. Compared to these other transition proteins, Atlas is present over a fairly narrow range of nuclear condensation stages and reaches its peak expression just prior to the onset of individualization. *Atlas* is also the only transition protein gene characterized to date whose removal disrupts fertility, as *Tpl94D*, *thmg-1*, *thmg-2* and *Mst84B* mutants are all fertile (Rathke et al. 2007; Gärtner et al. 2015; Gärtner et al. 2019). This may reflect the relatively later timing of Atlas's expression in spermatid nuclei, reduced functional redundancy between DNA-binding proteins at the later stages of condensation, a potential interaction between Atlas and an essential protamine-like protein, and/or a more stringent requirement for DNA binding at these stages.

Transition proteins give way in spermatid nuclei to protamine-like proteins, which bind DNA in mature sperm and persist through fertilization. In this way, protamine-like proteins function analogously to vertebrate protamines, though they are believed to be evolutionarily independent (Jayaramaiah Raja and Renkawitz-Pohl 2005; Doyen et al. 2015). In *D. melanogaster*, protamine-like proteins include Mst35Ba, Mst35Bb, Prtl99C and Mst77F (Rathke

588 et al. 2010; Eren-Ghiani et al. 2015; Kimura and Loppin 2016). Interestingly, while all
589 characterized protamine-like proteins are present in mature sperm, only some are essential for
590 fertility. Knockouts of *Mst35Ba*, *Mst35Bb*, or both show occasional nuclear shaping defects, but
591 male fertility is normal (Rathke et al. 2010; Tirmarche et al. 2014). In contrast, mutants of
592 *Prtl99C* or *Mst77F* are sterile. Prtl99C and Mst35Ba/b bind condensed DNA independently of
593 each other and contribute additively to the shortening of needle-stage nuclei, but Prtl99C's
594 effect is ~3x greater (Doyen et al. 2015; Eren-Ghiani et al. 2015). This difference is apparently
595 great enough to reduce fertility only in *Prtl99C* mutants. In contrast, *Mst77F* and *Mst35Ba/b*
596 show a genetic interaction, as *Mst35Ba/b* null flies become nearly sterile in an *Mst77F*
597 heterozygous background (Kimura and Loppin 2016). Furthermore, while Mst35Bb-GFP is
598 expressed in *Mst77F* nulls, these flies show deformed spermatid nuclei that do not reach a
599 recognizable needle-like stage. Because *atlas* nulls show considerable phenotypic similarity to
600 *Mst77F* nulls, but not *Prtl99C* nulls, we hypothesize that Atlas may act in a pathway with
601 Mst77F. Our observation of inefficient IC movement down sperm tails in *atlas* null testes is
602 reminiscent of a similar phenotype in *Mst77F* nulls (Kimura and Loppin 2016), providing further
603 evidence that these proteins may act in a common pathway. In both cases, ICs can form at
604 misshapen canoe-stage nuclei, but fail at a subsequent step. While the exact relationship
605 between nuclear abnormalities in the late canoe stage and individualization is not entirely
606 understood, it is possible that nuclear shape and the organization of nuclear bundles impact the
607 ability of IC association and IC progression, as is also observed in mutants of another gene,
608 *dPSMG1*, which controls nuclear shape (Gärtner et al. 2019).
609 Because *de novo* genes emerge from non-coding sequences, they typically encode
610 proteins that are short and lack complex structure (Schlötterer 2015; Van Oss and Carvunis
611 2019). Indeed, expanding the length of the protein-coding region and evolving higher-level
612 protein structures are hypothesized to be among the final stages of new gene evolution
613 (Bornberg-Bauer and Schmitz 2017). In light of these constraints, what kinds of cellular
614 functions might be available to newly evolved proteins? Vakirlis et al. (2020) overexpressed
615 emerging proto-genes in *S. cerevisiae* and found that those encoding proteins with predicted
616 transmembrane (TM) domains were more likely to be adaptive, as assessed by the effect of
617 proto-gene overexpression on growth rate. Such proteins may arise when thymine-rich
618 intergenic regions undergo mutations that allow protein-coding gene birth and expression, since
619 many codons with multiple U nucleotides encode amino acids commonly found in TM domains
620 (Vakirlis et al. 2020). Our imaging data, in addition to the prediction tools employed by Vakirlis
621 et al. (2020), suggest that Atlas does not contain a TM domain. However, just as the amino
622 acid compositional requirements of a TM domain are not overly complex, neither are those of
623 DNA binding proteins. In essence, these proteins must simply be small, have a high
624 concentration of positively charged residues, and contain a nuclear localization signal, which
625 itself requires a small patch of positively charged residues (Lange et al. 2007). Thus, DNA
626 binding proteins may be a relatively easy class of protein to evolve *de novo*.
627 While many putative *de novo* genes are expressed in the *D. melanogaster* testis (Fig. 1A
628 and Heames et al. 2020), *atlas* was the only verified hit from our screen that was essential for
629 male fertility. This result raises two related questions. First, why has selection maintained the
630 expression of the other potential *de novo* genes in our screen? In general, it is common for the
631 knockdown of protein-coding genes expressed in reproductive tissues in *D. melanogaster* to
632 result in no detectable fertility defects (Ravi Ram and Wolfner 2007; Schnakenberg et al. 2011;
633 Findlay et al. 2014; Gubala et al. 2017). One hypothesis to explain this pattern is that while the
634 loss of function of such genes may cause small reductions in fertility that would be subject to
635 strong negative selection in nature, the conditions used to assay such knockdown animals in the
636 primary screens are rarely tailored to detect differences of this magnitude. Another possibility,
637 not mutually exclusive, is that while the genes may be expendable in non-competitive, non-
638 exhaustive mating conditions, their absence may result in lower fitness in sperm competitive

environment, environments in which males mate several times in quick succession, or environments in which sperm must persist in storage for longer intervals or during less optimal conditions (Wong et al. 2008; Yeh et al. 2012; Civetta and Finn 2014).

A second question raised by our finding that *atlas* encodes an essential transition protein is: how might *atlas* have evolved to become essential for fertility in *D. melanogaster*, particularly when other transition proteins appear functionally redundant? Other proteins involved in spermatid chromatin compaction show variable levels of conservation across *Drosophila*. For example, the protamines around which DNA is wrapped in mature sperm (Jayaramaiah Raja and Renkawitz-Pohl 2005) are found across all sequenced *Drosophila* species (Alvi et al. 2013), and orthologs are also reported in FlyBase from other Dipteran and non-Dipteran insects (Larkin et al. 2021). However, transition protein Tpl94D is reported to be restricted to species ranging from *D. melanogaster* to *D. pseudoobscura* (Alvi et al. 2016), as are the related proteins tHMG1 and tHMG2 with high-mobility group domains (Gärtner et al. 2015; Larkin et al. 2021). Results like these suggest that while some protamine-like proteins (i.e., Mst35Ba and Mst35Bb) have consistently been among the final chromatin-packaging proteins, the specific proteins facilitating the transition from histones to protamines have likely varied over evolutionary time. Against this backdrop, and based on our analyses of the protein's presence/absence, biochemical properties, and expression patterns in extant species, we hypothesize that while the Atlas protein likely had some DNA-binding ability and male-specific expression upon its origin, it was only one of several proteins involved in spermatid chromatin compaction. Since *atlas* was lost independently in several lineages after its birth (Fig. 7), Atlas was likely non-essential at its outset, but rather evolved an essential function within the *melanogaster* group of species. Such evolution of essentiality could have occurred because of the loss of a protein with a complementary function and/or changes in the process of spermatogenesis that thrust Atlas into a functionally unique role. It is also worth noting that species that have evidently lost *atlas* might have undergone other compensatory changes in their repertoires of spermatid DNA binding proteins. For example, *D. willistoni* lacks *atlas* but appears to have several additional paralogs of the protamines found only in duplicate in *D. melanogaster*, which could have evolved transition-protein-like roles.

While our study cannot establish whether the movement of the *atlas* protein-coding sequence off of the X chromosome onto an autosome early in the evolution of the *melanogaster* group (Fig. 7) affected the gene's essentiality, such movement remains noteworthy. Prior work has found a significant dearth of testis-expressed genes on the X chromosome in *Drosophila* (Parisi et al. 2003; Parisi et al. 2004; Dorus et al. 2006; Vibranovski, Zhang, et al. 2009) and other species (Emerson et al. 2004; Reinke et al. 2004). Furthermore, *Drosophila* exhibit suppression of X-linked testis-expressed genes, and transfer of such genes from the X chromosome to autosomal loci results in higher expression levels (Kemkemer et al. 2014; Argyridou et al. 2017; Argyridou and Parsch 2018). One of several proposed mechanisms for both the paucity of X-linked testis-expressed genes and the suppression of their expression is meiotic sex chromosome inactivation (MSCI), in which the X chromosome becomes transcriptionally silenced earlier than autosomes (Vibranovski, Lopes, et al. 2009; Zhang, Vibranovski, Krinsky, et al. 2010; Zhang, Vibranovski, Landback, et al. 2010; Vibranovski et al. 2012; Gao et al. 2014; Mahadevaraju et al. 2021). Thus, genes that affect meiotic or post-meiotic processes, as *atlas* does, could exert beneficial effects more strongly and/or for a longer period of time if they become encoded autosomally. While the *atlas* protein-coding sequence appears to show male-specific expression regardless of its chromosomal location, it is possible that the movement of *atlas* to chromosome 2 allowed it to evolve a broader or different expression pattern that expanded or modified its role in spermiogenesis. The complex molecular bases of both X suppression and "escape" from the X chromosome in *Drosophila* continue to be actively investigated and debated (Meiklejohn et al. 2011; Mikhaylova and Nurminsky 2011; Vibranovski et al. 2012; Gao et al. 2014; Vibranovski 2014; Landeen et al.

690 2016; Mahadevaraju et al. 2021), but continued research in this area might inform further
691 interrogation of the forces driving *atlas* off of the X chromosome.
692 The movement of the *atlas* protein-coding sequence to chromosome 2 also created the
693 two-exon gene observed in *D. melanogaster*, in which the longer second exon appears to be
694 entirely non-coding. This second exon is highly conserved across the genus in both sequence
695 and genomic location, and it shows male-specific expression in several species that lack the
696 protein-coding sequence upstream (Fig. 7 and Fig. S7). These patterns of conservation
697 suggest that the second exon might originally have been a non-coding RNA, a class of molecule
698 whose importance in *Drosophila* male reproduction has recently become recognized (Wen et al.
699 2016; Bouska and Bai 2021). While these previous examples of functional ncRNAs in
700 spermatogenesis have generally acted in *trans* to regulate other genes or affect the functions of
701 other proteins, it is also possible that the long 3' UTR of *atlas* in *D. melanogaster* could affect
702 the translation of *atlas* transcripts. Many genes functioning in spermatid differentiation are
703 transcribed early in spermatogenesis but translationally repressed until later in spermiogenesis,
704 a phenomenon that relies on various forms of post-transcriptional regulation (White-Cooper
705 2010; Lim et al. 2012). Future studies of the *atlas* protein-coding sequence in the absence of its
706 3' UTR, the expression patterns of Atlas protein in species in which it is encoded from the X
707 chromosome, or the genetic ablation of the conserved region in species lacking the protein-
708 coding sequence will provide additional insights.
709 A final issue raised by our results is the exact timing and mechanism of origin for the
710 *atlas* protein-coding sequence. The bioinformatic screen (Heames et al. 2020) that identified
711 *atlas* and the other genes tested in Fig. 1 was designed to identify both "*de novo*" genes,
712 defined as protein-coding regions in *Drosophila* that had recognizable, but non-ORF-
713 maintaining, TBLASTN hits in outgroup species, and "putative *de novo*" genes, which had no
714 TBLASTN hits in outgroup species. (Importantly, the screen also eliminated any protein with an
715 identifiable protein domain, thus reducing the chances of identifying divergent members of gene
716 families.) The vast majority of the genes we tested with RNAi, including *atlas*, fell into the
717 putative *de novo* category. The bioinformatic screen's criteria were reasonable for a high-
718 throughput analysis, but BLAST-based methods have known limitations for detecting
719 orthologous sequences in diverged species (Moyers and Zhang 2015, 2018; Weisman et al.
720 2020). The lack of identifiable *atlas* protein-coding genes in several *Drosophila* species (e.g., *D.*
721 *pseudoobscura* and *D. willistoni*) is unlikely to be due to BLAST homology detection failure, and
722 extensive synteny-based searches confirmed the gene's absence (Fig. S9). BLAST and
723 synteny-based searches for orthologs in non-*Drosophila* species also did not detect an ortholog,
724 though BLAST searches are not predicted to have adequate sensitivity, for a protein of this size
725 and evolutionary rate, at this level of species divergence (Weisman et al. 2020). Hence, in
726 addition to using synteny to search for orthologs, we used HMMER, which employs hidden
727 Markov models and builds a sequence profile of the target protein using information from
728 multiple orthologs. Since HMMER also did not detect orthologs outside of *Drosophila*, we
729 hypothesize that *atlas* evolved *de novo* at the base of the genus. However, since we remain
730 unable to identify the non-protein-coding sequence from which *atlas* arose, we continue to refer
731 to *atlas* as a putative *de novo* gene (McLysaght and Hurst 2016).
732 Overall, we find that while many putative *de novo* evolved genes are expressed in the *D.*
733 *melanogaster* testes, few have major, non-redundant effects on fertility. However, several such
734 genes have acquired critical roles, acting at distinct stages of spermatogenesis and sperm
735 function. We showed previously that the putative *de novo* gene *saturn* is required for maximal
736 sperm production, as well as for the ability of transferred sperm to migrate successfully to sperm
737 storage organs in females (Gubala et al. 2017). Another putative *de novo* gene, *goddard*, is
738 required for sperm production and encodes a cytoplasmic protein that appears to localize to
739 elongating axonemes (Gubala et al. 2017; Lange et al. 2021). Loss of *goddard* impairs the
740 individualization of spermatid bundles (Lange et al. 2021), thus exerting an effect that appears

741 to be upstream of those observed for *saturn* and *atlas*. Here, we report another novel function
742 for a putative *de novo* gene: encoding an essential transition protein that is necessary for proper
743 nuclear condensation in spermiogenesis. Taken together, these results demonstrate that while
744 many *de novo* genes may play subtle roles or share functional redundancy with other genes, *de*
745 *novo* genes can also become essential players in complex cellular processes that mediate
746 successful reproduction.
747

Materials and Methods

RNA Interference Screen

De novo and putative *de novo* genes inferred to be no older than the *Drosophila* genus were identified previously (Heames et al. 2020). We filtered these genes with publicly available RNA-seq data (Brown et al. 2014) to identify those expressed predominantly in the testes (>50% of RPKM sum deriving from the testes from ModENCODE data; Brown et al. 2014), giving a total of 96 genes. To assess each of these candidates for effects on male fertility, we induced knockdown in the male germline by crossing UAS-RNAi flies to Bam-GAL4, UAS-Dicer2 flies (White-Cooper 2012; Gubala et al. 2017). Control flies were generated by crossing the genetic background into which UAS-RNAi was inserted crossed to the same GAL4 line. Flies carrying UAS-RNAi were of two types. Roughly half of the genes had publicly available lines from the Vienna *Drosophila* Resource Center (Dietzl et al. 2007) or the Transgenic RNAi Project (Ni et al. 2011). For the other genes, no publicly available RNAi stock was available, so we constructed TRiP-style stocks in the pValium20 vector as previously described (Findlay et al. 2014). These constructs were integrated into an AttP site in stock BL 25709 (y^1 v^1 P{nos-phiC31\int.NLS}X; P{CaryP}attP40) from the Bloomington *Drosophila* Stock Center (injections by Genetivision; Houston, TX, USA) and crossed into a *y v* background to screen for v^+. We attempted at least two rounds of transgenic production for each gene. In total, we were able to obtain and test RNAi lines for 57 of the 96 identified genes.

We initially screened males knocked down for each candidate gene for major fertility defects by crossing groups of 7 knockdown or control males to 5 virgin Canton S females, letting the adults lay eggs for ~48 hours, and then discarding adults and quantifying the resulting progeny by counting the pupal cases, as previously described (Gubala et al. 2017). To assess the degree of knockdown achieved, 10 whole males of each line were homogenized in TRIzol reagent (Life Technologies, Carlsbad, CA). RNA isolation, DNAse treatment, cDNA synthesis and semi-quantitative RT-PCR with gene-specific primers were performed as previously described; amplification of *RpL32* was used as a positive control (Gubala et al. 2017). Any gene that did not show at least partial knockdown was discarded from further analysis, leaving a total of 42 genes successfully screened. Table S3 lists all lines used and results of tests for effective target gene knockdown.

CRISPR Genome Editing

To validate RNAi results for *atlas*, *CG43072* and *CG33284*, we used CRISPR/Cas9 genome-editing to generate null alleles that could be used for further analysis, as described previously (Lange et al. 2021). Briefly, our general strategy was to design gRNAs in the pU6.3 vector (Drosophila Genome Resource Center (DGRC) #1362) that targeted each end of a locus. These plasmids, along with plasmids encoding gRNAs that targeted the w+ locus, were co-injected by Rainbow Transgenics (Camarillo, CA) into embryos laid by *vasa*-Cas9 females in a w+ background, Bloomington stock #51323 (Ge et al. 2016). G_0 animals were crossed to *w*- flies, and members of G_1 broods with a higher-than-expected fraction of *w*- progeny were individually crossed to an appropriate balancer line and then PCR-screened for the desired deletion of the targeted locus.

We also constructed three frameshift, expected loss-of-function alleles for *atlas* by using CRISPR to induce non-homologous end joining at a single PAM site just downstream of the *atlas* start codon. *Vasa*-Cas9 embryos were co-injected and screened for w- progeny as described above. We then used squish preps to isolate DNA from G1 flies and used a PCR-RFLP assay to detect mutations. PCR products spanning the gRNA-targeted site were digested with *Bfa*I (New England Biolabs (NEB), Ipswich, MA); undigested products in which the

799 expected *Bfa*I site was lost indicated a mutation, which was balanced and then confirmed by
800 PCR and sequencing of homozygous mutant lines.
801 We used scarless CRISPR editing and homology-directed repair (HDR) to insert the
802 GFP protein-coding sequence in-frame at the end of the *atlas* protein sequence (see Fig. S6;
803 (https://flycrispr.org/scarless-gene-editing/) (Bruckner et al. 2017; Bier et al. 2018; Hill et al.
804 2019). We first generated an *atlas*-GFP DNA construct by cloning the *atlas* protein-coding
805 sequence into pENTR and using LR Clonase II (Thermo Fisher Scientific, Waltham, MA) to
806 recombine the sequence with pTWG (DGRC #1076; T. Murphy), generating a C-terminally
807 tagged *atlas*:GFP construct. We amplified the *atlas* fragment from *vasa*-Cas9 strain #51323
808 genomic DNA. Once *atlas*-GFP was obtained in a plasmid, we amplified it with primers that
809 contained 5' *Esp*3I sites and overhangs designed for Golden Gate Assembly (GGA) and that, in
810 the case of the reverse primer, also added on 42 nucleotides downstream of the *atlas* stop
811 codon to reach a PAM site identified by FlyCRISPR TargetFinder (Gratz et al. 2014) as being
812 optimal for Cas9/gRNA recognition and cleavage. The primer also introduced a mutation in the
813 PAM site so that insertion of the designed piece of DNA into the genome *in vivo* would not be
814 subject to re-cutting. We also used the NEB Q5 Site-Directed Mutagenesis kit to introduce a
815 silent mutation into the *atlas* protein-coding sequence to eliminate an internal *Esp*3I site. The
816 resulting construct was used as the "left" homology arm for homology-directed repair (HDR)
817 editing. We constructed a "right" homology arm by using NEB Q5 PCR to amplify a 982-bp
818 fragment downstream of the PAM site, using primers modified to contain *Esp*3I sites and
819 overhangs compatible with GGA. We performed GGA by combining these left and right arms, a
820 plasmid containing a PiggyBac transposase-excisable 3xP3-dsRed flanked by *Esp*3I sites, and
821 backbone plasmid pXZ13, with *Esp*3I and T4 DNA ligase (NEB). A combination of colony PCR,
822 restriction digestion and sequencing identified properly assembled plasmids suitable for HDR.
823 *Vasa*-Cas9 embryos were co-injected with the assembled plasmid and a pU6.3 plasmid
824 encoding a gRNA targeting the region just downstream of the *atlas* stop codon. G0 flies were
825 crossed to w^{1118} adults, and G1 flies were screened for red fluorescent eyes using the
826 NIGHTSEA system (NIGHTSEA LLC, Lexington, MA). Six balanced lines from two independent
827 G1 broods were established. To remove the dsRed from the *atlas* locus, we crossed these lines
828 to a PiggyBac transposase line (BDSC #8285) and then selected against pBac and dsRed in
829 the following generation. PCR and sequencing confirmed the expected "scarless" insertion of
830 GFP at the *atlas* locus.
831
832 Atlas *Rescue Line*
833
834 We constructed an HA-tagged *atlas* rescue line that contained the *atlas* gene flanked by
835 1345 bp of sequence upstream of the start codon (but excluding the coding sequence of
836 upstream gene CG3124) and 3000 bp of sequence downstream of the stop codon (including the
837 full 3' UTR) as follows. Genomic sequences were PCR amplified using Q5 High fidelity
838 Polymerase (NEB), purified Canton S genomic DNA (Gentra Puregene Tissue Kit, Qiagen,
839 Germantown, MD), and the atlas rescue F1/R1 and atlas rescue F3/R3 primer sets (see Table
840 S4). The 3x-HA tag was likewise amplified from pTWH (DGRC 1100; T. Murphy) using atlas
841 rescue F2/R2 primers. These DNA fragments were subsequently assembled into a XbaI/AscI-
842 linearized w+attB plasmid (Addgene, Watertown, MA, plasmid 30326, deposited by J.
843 Sekelsky). The assembled construct was then phiC31 integrated into the PBac{y^+-attP-
844 3B}VK00037 (Bloomington *Drosophila* Stock Center (BDSC) stock #24872) docking site
845 (Rainbow Transgenics) and crossed into the *atlas* null background to assess rescue.
846
847 *Fertility Assays and Sperm Visualization*
848

To validate the finding of reduced fertility for *atlas* knockdown males in the group fertility assay described above, we performed single-pair fertility assays in which knockdown or mutant males or their controls were mated individually to Canton S virgin females. Based on previous experience analyzing genes that resulted in sterility or near-sterility (Gubala et al. 2017; Lange et al. 2021), we designed assays with *N* = 20-30 flies per male genotype. Matings were observed, and males were discarded after copulation. Females were allowed to lay eggs into the vials for 4 days and then discarded. Pupal cases were counted as a measure of fertility. Crosses to generate and mating assays involving RNAi flies were maintained at 25º to optimize knockdown. Before all assays, flies were reared to sexual maturity (3-7 days) in single-sex groups on cornmeal-molasses food supplemented with dry yeast grains (Gubala et al. 2017).

To assess the ability of *atlas*-GFP to rescue the fertility defect of the Δ*atlas* allele, we crossed atlas-GFP and w1118 flies to Δ*atlas*/SM6. Males with genotypes atlas-GFP/Δ*atlas* and +/Δ*atlas* were compared using the single-pair fertility assay described above.

To observe the production of sperm in knockdown or mutant males, we introduced the Mst35Bb-GFP marker into these males, which labels mature sperm and late-stage spermatid nuclei with GFP (Manier et al. 2010). Samples were prepared, imaged and analyzed as described previously (Gubala et al. 2017); because the large differences in testis shape and sperm production observed in initial phase contrast imaging would be obvious to any experimenter, we were not blind to male genotypes.

Atlas-GFP Ectopic Expression

We used the Gateway cloning system (Thermo) to construct an *atlas*-GFP transgene expressed under UAS control (primers in Table S4). The *atlas* protein-coding sequence in pENTR was recombined with pTWG (*Drosophila* Genomics Resource Center, T. Murphy) as described above. The resulting plasmid was then inserted into w- flies using P-element-mediated transposition (Rainbow Transgenics), w+ G1s were selected, and several independent insertions were balanced. We crossed male UAS-*atlas*:GFP flies to females from two different driver lines: *tubulin*-GAL4 (to drive ubiquitous expression) and *Bam*-GAL4 (to drive expression in the early germline). We dissected larval salivary glands of the *tub*>*atlas*:GFP males, since these cells are exceptionally large and ideal for visualizing subcellular localization. We then dissected the testes of *Bam*>*atlas*:GFP males to evaluate whether the localization pattern observed in the salivary gland was consistent in testis tissue, albeit not the same cells in which endogenous atlas appears to be expressed. Protein localization was visualized by fluorescence confocal microscopy on a Leica SP5 microscope (Leica Microsystems, Wetzlar, Germany) and images were captured with LASAF as described previously (Lange et al. 2021).

Imaging Spermatogenesis and Spermatid Nuclear Condensation

We used phase-contrast microscopy to examine the stages of spermatogenesis in whole mount testes (White-Cooper 2004). To assess the processes of nuclear condensation and individualization of 64-cell cysts of spermatids in the post-meiotic stages of spermatogenesis, we used fluorescence and confocal microscopy to visualize actin-based individualization complexes and nuclei. Samples were processed, and actin and nuclear DNA were visualized with TRITC-phalloidin (Molecular Probes, Eugene, OR) and TOPRO-3 iodine (Thermo), respectively, as described previously (Lange et al. 2021). The final stages of nuclear condensation were visualized with the Mst35Bb-GFP marker described above, as well as an equivalent marker, Mst35Bb-dsRed (Manier et al. 2010). Earlier nuclear stages were visualized with histone H2AvD-RFP (BDSC stock #23651), which is present in round spermatid nuclei and the earliest stages of nuclear elongation (Clarkson and Saint 1999; Rathke et al. 2007). Images

899 with H2AvD-RFP were obtained with epifluorescence microscopy, since we lack an appropriate
900 confocal laser for RFP.
901 To examine spermatid nuclei at various stages of condensation, we visualized nuclear
902 bundles using TOPRO-3. Testes of newly eclosed (<1 day old) *atlas* null and control males
903 were dissected in PBS. Testes were then transferred to a droplet of 4% paraformaldehyde on
904 poly-L-lysine treated glass slides and were gently shredded in the post-meiotic region to release
905 sperm bundles. Testes were gently squashed beneath coverslips coated in Sigmacote (Sigma
906 Aldrich, St. Louis, MO). We then froze slides in liquid nitrogen for a few seconds and popped off
907 of the siliconized coverslip with a razor. Slides were incubated in Coplin jars filled with 95%
908 ethanol at -20°C for 30 minutes and then mounted in VECTASHIELD (Vector Laboratories,
909 Burlingame, CA). Nuclear staging was performed by examining the shape of the nuclei. Early
910 and late canoe stages of condensation were distinguished by the absence or presence of
911 Mst35Bb-GFP, respectively. Elongated and condensed stages were distinguished by the
912 presence or absence, respectively, of vesicles of GFP-tagged nuclear proteins (Atlas-GFP or
913 Mst35Bb-GFP) located basal to the nuclei. Examples of stages are given in Fig. 4. Confocal
914 stacks were taken on a Leica SP5 microscope, images were captured by LASAF, and ImageJ
915 was used to flatten stacks into a single, two-dimensional image. All intact nuclear bundles were
916 counted for each dissection.
917 For the experiments measuring nuclear condensation stage (Table S1), a sample size of
918 N = 10 for each genotype was selected based on the magnitude of the *atlas* null phenotype and
919 the consistent differences observed in previous dissections of these genotypes with Mst35Bb-
920 GFP. Likewise, for the IC-nuclear bundle association and IC progression analysis (Fig. 3C-D),
921 we selected sample sizes of N = ~15 per genotype based on pilot experiments showing that
922 aberrant actin phenotypes were highly consistent in null testes and previous experience with
923 such quantification (Lange et al. 2021).
924
925 *Evolutionary and Gene Expression Analysis of* atlas
926
927 We searched for orthologs of the *D. melanogaster* Atlas protein in the original
928 sequenced *Drosophila* species with BLASTP searches in FlyBase (Clark et al. 2007). We also
929 used TBLASTN searches to identify orthologs in species lacking complete protein annotations.
930 We identified syntenic regions for each species by looking for conserved neighboring genes,
931 such as *ord* and *CG3124*. In addition to analyzing the *atlas* coding region, we conducted
932 separate BLASTN searches for the sequence of the *D. melanogaster* 3'UTR across *Drosophila*
933 species since it has a different conservation pattern than the coding sequence.
934 To test for sex-specific expression biases for both the ORF and the 3' UTR sequences,
935 we used adult male- and female-specific RNA-seq data from numerous *Drosophila* species
936 accessed through the Genomics Education Partnership version of the UCSC Genome Browser
937 (http://gander.wustl.edu/) and initially collected by Brown et al. (2014) and Chen et al. (2014).
938 We also confirmed these findings experimentally in several species by performing RT-PCR on
939 cDNA isolated from whole males and whole females, as previously described (Gubala et al.
940 2017).
941 To search for *atlas* orthologs in non-*Drosophila* Dipterans, we obtained from ENSEMBL
942 Metazoa the genomes of *Musca domestica, Glossina morsitans, Lucilia cuprina, Aedes aegypti,*
943 *Anopheles darlingi, Anopheles gambiae, Culex quinquefasciatus* and *Mayetiola destructor*. We
944 performed a synteny search (summarized in Fig. S10) in each species by identifying the nearest
945 neighbors of *atlas* in the *D. ananassae* and *D. virilis* genomes that had an identifiable homolog
946 in each species. In all cases, the homologs of the nearest neighbors on each side of atlas were
947 found on different contigs, suggesting synteny breakdown. We obtained up to 1 Mb of
948 sequence on each side of each identified homolog and queried it with BLASTN, TBLASTN, and
949 Exonerate (Slater and Birney 2005) for regions with significantly similarity to any portion of the

Atlas protein or cDNA sequences. No significant hits, and no hits better than what could be found in other parts of the genome, were found. Finally, we used HMMER to search for orthologs in all annotated proteomes and all metazoan proteomes. We first queried the database with Atlas from either *D. melanogaster* or *D. virilis* and accepted hits that fell below an e-value cutoff of 0.01 and a minimum hit length of 3%. These hits were then included iteratively in subsequent searches until no new significant hits were found.

We analyzed the molecular evolution of the *atlas* protein-coding sequence by obtaining orthologous protein-coding sequences from *melanogaster* group species. (Analysis out of this group was not performed due to high sequence divergence and poor alignment quality.) We used BLASTP to identify these sequences from GenBank and then extracted the coding DNA sequence for each. Sequences were aligned, checked for recombination, used to construct a gene tree, and analyzed with the PAML sites test as described previously (McGeary and Findlay 2020), except that alignment positions that included gaps were masked from the PAML analysis. We initially analyzed a set of 13 species (*melanogaster*, *simulans*, *sechellia*, *yakuba*, *erecta*, *suzukii*, *takahashii*, *biarmipes*, *rhopaloa*, *ficusphila*, *elegans*, *eugracilis* and *ananassae*); we excluded an ortholog detected in *D. bipectinata* due to poor alignment. This initial analysis detected a class of sites with significant evidence of positive selection, but closer inspection of the alignment revealed that the site with the strongest evidence of selection, corresponding to *D. melanogaster* residue 31R, may have been driven by a questionable alignment due to an insertion in that region that was unique to *D. takahashii*. The results reported derive from an analysis with *D. takahashii* excluded, which produced a more reliable alignment and showed no evidence for any sites under positive selection.

973 **Acknowledgements**
974
975 We thank Dr. Alexis Hill for assistance with developing the *atlas*-GFP transgene; Gynesis
976 Vance, Elvis Perez, Ishanpepe Jagusah, Emily Gualdino, and students in the Biology 261 Lab at
977 Holy Cross for assistance with RNAi fertility screens; Dr. Rob Bellin for assistance with confocal
978 microscopy; Dr. Justin McAlister for use of the NIGHTSEA system; the Bloomington Stock
979 Center and the Vienna *Drosophila* Resource Center for fly strains; and Dr. Mariana Wolfner and
980 members of the Findlay and Bornberg-Bauer labs for helpful discussions about the project and
981 the manuscript. This work was supported by NSF CAREER Award #1652013 (to GDF) and a
982 Humboldt Fellowship (to AG). GDF and EBB are also grateful to the Evolution Think Tank at
983 the University of Münster, which funded GDF as a visiting fellow for in-person collaboration.
984

985 **References**
986

987 Abrusán G. 2013. Integration of New Genes into Cellular Networks, and Their Structural
988 Maturation. Genetics 195:1407-1417.
989 Alvi ZA, Chu T-C, Schawaroch V, Klaus AV. 2016. Genomic and expression analysis of
990 transition proteins in *Drosophila*. Spermatogenesis 5:e1178518.
991 Alvi ZA, Chu T-C, Schawaroch V, Klaus AV. 2013. Protamine-like proteins in 12 sequenced
992 species of Drosophila. Protein and peptide letters 20:17-35.
993 Argyridou E, Huylmans AK, Königer A, Parsch J. 2017. X-linkage is not a general inhibitor of
994 tissue-specific gene expression in *Drosophila melanogaster*. Heredity 119:27-34.
995 Argyridou E, Parsch J. 2018. Regulation of the X Chromosome in the Germline and Soma of
996 Drosophila melanogaster Males. Genes 9:242.
997 Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, Jentoft S.
998 2018. *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole
999 genome sequence data. Molecular Biology and Evolution 35:593-606.
1000 Baker RH, Narechania A, Johns PM, Wilkinson GS. 2012. Gene duplication, tissue-specific
1001 gene expression and sexual conflict in stalk-eyed flies (Diopsidae). Philosophical
1002 Transactions of the Royal Society of London. Series B, Biological Sciences 367:2357-
1003 2375.
1004 Begun DJ, Lindfors HA, Kern AD, Jones CD. 2007. Evidence for *de Novo* Evolution of Testis-
1005 Expressed Genes in the *Drosophila yakuba/Drosophila erecta* Clade. Genetics
1006 176:1131-1137.
1007 Bier E, Harrison MM, O'Connor-Giles KM, Wildonger J. 2018. Advances in Engineering the Fly
1008 Genome with the CRISPR-Cas System. Genetics 208:1-18.
1009 Bornberg-Bauer E, Schmitz JF. 2017. Fact or fiction: Updates on how protein-coding genes
1010 might emerge *de novo* from previously non-coding DNA. F1000Research 6:57.
1011 Bouska MJ, Bai H. 2021. Long noncoding RNA regulation of spermatogenesis via the spectrin
1012 cytoskeleton in Drosophila. G3 Genes|Genomes|Genetics.
1013 Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, Booth BW, Wen J, Park S,
1014 Suzuki AM, et al. 2014. Diversity and dynamics of the Drosophila transcriptome. Nature
1015 512:393-399.
1016 Bruckner JJ, Zhan H, Gratz SJ, Rao M, Ukken F, Zilberg G, O'Connor-Giles KM. 2017. Fife
1017 organizes synaptic vesicles and calcium channels for high-probability neurotransmitter
1018 release. Journal of Cell Biology 216:231-246.
1019 Cai J, Zhao R, Jiang H, Wang W. 2008. *De novo* origination of a new protein-coding gene in
1020 *Saccharomyces cerevisiae*. Genetics 179:487-496.
1021 Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, Charloteaux B,
1022 Hidalgo C, Barbette J, Santhanam B, et al. 2012. Proto-genes and *de novo* gene birth.
1023 Nature 487:370-374.
1024 Chamakura KR, Tran JS, O'Leary C, Lisciandro HG, Antillon SF, Garza KD, Tran E, Min L,
1025 Young R. 2020. Rapid de novo evolution of lysis genes in single-stranded RNA phages.
1026 Nature Communications 11.
1027 Chen Z-X, Sturgill D, Qu J, Jiang H, Park S, Boley N, Suzuki AM, Fletcher AR, Plachetzki DC,
1028 FitzGerald PC, et al. 2014. Comparative validation of the *D. melanogaster* modENCODE
1029 transcriptome annotation. Genome Research 24:1209-1223.
1030 Civetta A, Finn S. 2014. Do candidate genes mediating conspecific sperm precedence affect
1031 sperm competitive ability within species? A test case in *Drosophila*. G3 (Bethesda, Md.)
1032 4:1701-1707.
1033 Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, Kaufman TC, Kellis M,
1034 Gelbart W, Iyer VN, et al. 2007. Evolution of genes and genomes on the Drosophila
1035 phylogeny. Nature 450:203-218.

1036 Clarkson M, Saint R. 1999. A His2AvDGFP fusion gene complements a lethal His2AvD mutant
1037 allele and provides an in vivo marker for Drosophila chromosome behavior. DNA and
1038 Cell Biology 18:457-462.
1039 Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. 2015. Young genes out of the male: an
1040 insight from evolutionary age analysis of the pollen transcriptome. Molecular Plant
1041 8:935-945.
1042 Demarco RS, Eikenes TH, Haglund K, Jones DL. 2014. Investigating spermatogenesis in
1043 *Drosophila melanogaster*. Methods 68:218-227.
1044 Dietzl G, Chen D, Schnorrer F, Su K-C, Barinova Y, Fellner M, Gasser B, Kinsey K, Oppel S,
1045 Scheiblauer S, et al. 2007. A genome-wide transgenic RNAi library for conditional gene
1046 inactivation in *Drosophila*. Nature 448:151-156.
1047 Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. 2006. Genomic and functional
1048 evolution of the *Drosophila melanogaster* sperm proteome. Nature Genetics 38:1440-
1049 1445.
1050 Dorus S, Freeman ZN, Parker ER, Heath BD, Karr TL. 2008. Recent Origins of Sperm Genes in
1051 Drosophila. Molecular Biology and Evolution 25:2157-2166.
1052 Doyen CM, Chalkley GE, Voets O, Bezstarosti K, Demmers JA, Moshkin YM, Verrijzer CP.
1053 2015. A Testis-Specific Chaperone and the Chromatin Remodeler ISWI Mediate
1054 Repackaging of the Paternal Genome. Cell Reports 13:1310-1318.
1055 Emerson JJ, Kaessmann H, Betrán E, Long M. 2004. Extensive gene traffic on the mammalian
1056 X chromosome. Science 303:537-540.
1057 Eren-Ghiani Z, Rathke C, Theofel I, Renkawitz-Pohl R. 2015. Prtl99C Acts Together with
1058 Protamines and Safeguards Male Fertility in *Drosophila*. Cell Reports 13:2327-2335.
1059 Fabian L, Brill JA. 2012. Drosophila spermiogenesis. Spermatogenesis 2:197-212.
1060 Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. 2014. Evolutionary rate
1061 covariation identifies new members of a protein network required for *Drosophila*
1062 *melanogaster* female post-mating responses. PLoS Genetics 10:e1004108.
1063 Gao G, Vibranovski MD, Zhang L, Li Z, Liu M, Zhang YE, Li X, Zhang W, Fan Q, Vankuren NW,
1064 et al. 2014. A long-term demasculinization of X-linked intergenic noncoding RNAs in
1065 *Drosophila melanogaster*. Genome Research 24:629-638.
1066 Gärtner SMK, Hundertmark T, Nolte H, Theofel I, Eren-Ghiani Z, Tetzner C, Duchow TB,
1067 Rathke C, Krüger M, Renkawitz-Pohl R. 2019. Stage-specific testes proteomics of
1068 *Drosophila melanogaster* identifies essential proteins for male fertility. European Journal
1069 of Cell Biology 98:103-115.
1070 Gärtner SMK, Rothenbusch S, Buxa MK, Theofel I, Renkawitz R, Rathke C, Renkawitz-Pohl R.
1071 2015. The HMG-box-containing proteins tHMG-1 and tHMG-2 interact during the
1072 histone-to-protamine transition in *Drosophila* spermatogenesis. European Journal of Cell
1073 Biology 94:46-59.
1074 Ge DT, Tipping C, Brodsky MH, Zamore PD. 2016. Rapid screening for CRISPR-directed
1075 editing of the *Drosophila* genome using *white* coconversion. G3 (Bethesda, Md.) 6:3197-
1076 3206.
1077 Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, O'Connor-Giles
1078 KM. 2014. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed
1079 repair in Drosophila. Genetics 196:961-971.
1080 Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, Findlay GD.
1081 2017. The *goddard* and *saturn* genes are essential for *Drosophila* male fertility and may
1082 have arisen *de novo*. Molecular Biology and Evolution 34:1066-1082.
1083 Guerzoni D, McLysaght A. 2016. *De novo* genes arise at a slow but steady rate along the
1084 primate lineage and have been subject to incomplete lineage sorting. Genome biology
1085 and evolution 8:1222-1232.

Hales KG, Korey CA, Larracuente AM, Roberts DM. 2015. Genetics on the fly: a primer on the *Drosophila* model system. Genetics 201:815-842.

Heames B, Schmitz J, Bornberg-Bauer E. 2020. A continuum of evolving de novo genes drives protein-coding novelty in *Drosophila*. Journal of Molecular Evolution 88:382-398.

Hill A, Jain P, Ben-Shahar Y. 2019. The *Drosophila* ERG channel seizure plays a role in the neuronal homeostatic stress response. PLoS Genetics 15:e1008288.

Hundertmark T, Gärtner SMK, Rathke C, Renkawitz-Pohl R. 2018. Nejire/dCBP-mediated histone H3 acetylation during spermatogenesis is essential for male fertility in *Drosophila melanogaster*. PLOS ONE 13:e0203622.

Jayaramaiah Raja S, Renkawitz-Pohl R. 2005. Replacement by *Drosophila melanogaster* protamines and Mst77F of histones during chromatin condensation in late spermatids and role of Sesame in the removal of these proteins from the male pronucleus. Molecular and Cellular Biology 25:6165-6177.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. Genome Research 20:1313-1326.

Kanippayoor RL, Alpern JHM, Moehring AJ. 2013. Protamines and spermatogenesis in *Drosophila* and *Homo sapiens*. Spermatogenesis 3:e24376.

Keeling DM, Garza P, Nartey CM, Carvunis A-R. 2019. The meanings of 'function' in biology and the problematic case of de novo gene emergence. eLife 8:e47014.

Kemkemer C, Catalán A, Parsch J. 2014. ‘Escaping’ the X chromosome leads to increased gene expression in the male germline of *Drosophila melanogaster*. Heredity 112:149-155.

Kimura S, Loppin B. 2016. The Drosophila chromosomal protein Mst77F is processed to generate an essential component of mature sperm chromatin. Open Biology 6:160207.

Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan L, Pang N, Aradhya R, Siepel A, Steinhauer J, Lai EC. 2017. New genes often acquire male-specific functions but rarely become essential in *Drosophila*. Genes & development 31:1841-1846.

Landeen EL, Muirhead CA, Wright L, Meiklejohn CD, Presgraves DC. 2016. Sex Chromosome-wide Transcriptional Suppression and Compensatory *Cis*-Regulatory Evolution Mediate Gene Expression in the *Drosophila* Male Germline. PLOS Biology 14:e1002499.

Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH. 2007. Classical nuclear localization signals: definition, function, and interaction with Importin α. Journal of Biological Chemistry 282:5101-5105.

Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, Findlay GD, Bornberg-Bauer E. 2021. Structural and functional characterization of a putative de novo gene in *Drosophila*. Nature Communications 12:1667.

Larkin A, Marygold SJ, Antonazzo G, Attrill H, Gilberto, Garapati PV, Joshua, Millburn G, Strelets VB, Tabone CJ, et al. 2021. FlyBase: updates to the *Drosophila melanogaster* knowledge base. Nucleic acids research 49:D899-D907.

Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. 2006. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. Proceedings of the National Academy of Sciences 103:9935-9939.

Li D, Yan Z, Lu L, Jiang H, Wang W. 2014. Pleiotropy of the *de novo*-originated gene *MDF1*. Scientific reports 4:7280.

Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, Ge S, Guo Y-L. 2016. On the origin of *de novo* genes in *Arabidopsis thaliana* populations. Genome biology and evolution 8:2190-2202.

Lim C, Tarayrah L, Chen X. 2012. Transcriptional regulation during Drosophila spermatogenesis. Spermatogenesis 2:158-166.

1135 Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. 2011. High
1136 spontaneous rate of gene duplication in Caenorhabditis elegans. Current biology : CB
1137 21:306-310.
1138 Lu T-C, Leu J-Y, Lin W-C. 2017. A comprehensive analysis of transcript-supported de novo
1139 genes in *Saccharomyces sensu stricto* yeasts. Molecular Biology and Evolution 34:2823-
1140 2838.
1141 Mahadevaraju S, Fear JM, Akeju M, Galletta BJ, Pinheiro MMLS, Avelino CC, Cabral-De-Mello
1142 DC, Conlon K, Dell'Orso S, Demere Z, et al. 2021. Dynamic sex chromosome
1143 expression in Drosophila male germ cells. Nature Communications 12.
1144 Manier MK, Belote JM, Berben KS, Novikov D, Stuart WT, Pitnick S. 2010. Resolving
1145 mechanisms of competitive fertilization success in *Drosophila melanogaster*. Science
1146 328:354-357.
1147 Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of
1148 young human genes after a burst of retroposition in primates. PLOS Biology 3:e357.
1149 McGeary MK, Findlay GD. 2020. Molecular evolution of the sex peptide network in *Drosophila*.
1150 Journal of Evolutionary Biology 33:629-641.
1151 McLysaght A, Hurst LD. 2016. Open questions in the study of *de novo* genes: what, how and
1152 why. Nature Reviews Genetics 17:567.
1153 Meiklejohn CD, Landeen EL, Cook JM, Kingan SB, Presgraves DC. 2011. Sex Chromosome-
1154 Specific Regulation in the *Drosophila* Male Germline But Little Evidence for
1155 Chromosomal Dosage Compensation or Meiotic Inactivation. PLOS Biology 9:e1001126.
1156 Mikhaylova LM, Nurminsky DI. 2011. Lack of global meiotic sex chromosome inactivation, and
1157 paucity of tissue-specific gene expression on the Drosophila X chromosome. BMC
1158 Biology 9:29.
1159 Moyers BA, Zhang J. 2015. Phylostratigraphic bias creates spurious patterns of genome
1160 evolution. Molecular Biology and Evolution 32:258-267.
1161 Moyers BA, Zhang J. 2018. Toward reducing phylostratigraphic errors and biases. Genome
1162 biology and evolution 10:2037-2048.
1163 Necsulea A, Kaessmann H. 2014. Evolutionary dynamics of coding and non-coding
1164 transcriptomes. Nature Reviews Genetics 15:734-748.
1165 Ni J-Q, Zhou R, Czech B, Liu L-P, Holderbaum L, Yang-Zhou D, Shim H-S, Tao R, Handler D,
1166 Karpowicz P, et al. 2011. A genome-scale shRNA resource for transgenic RNAi in
1167 *Drosophila*. Nature Methods 8:405-407.
1168 Palmieri N, Kosiol C, Schlötterer C. 2014. The life cycle of *Drosophila* orphan genes. eLife
1169 3:e01311.
1170 Parisi M, Nuttall R, Edwards P, Minor J, Naiman D, Lü J, Doctolero M, Vainer M, Chan C,
1171 Malley J, et al. 2004. A survey of ovary-, testis-, and soma-biased gene expression in
1172 *Drosophila melanogaster* adults. Genome biology 5:R40.
1173 Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B. 2003.
1174 Paucity of genes on the *Drosophila* X chromosome showing male-biased expression.
1175 Science 299:697-700.
1176 Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web server: 2018
1177 update. Nucleic acids research 46:W200-W204.
1178 Puntambekar S, Newhouse R, Navas JSM, Chauhan R, Vernaz G, Willis T, Wayland MT,
1179 Umrania Y, Miska EA, Prabakaran S. 2020. Evolutionary divergence of novel open
1180 reading frames in cichlids speciation. Scientific reports 10.
1181 Rathke C, Baarends WM, Awe S, Renkawitz-Pohl R. 2014. Chromatin dynamics during
1182 spermiogenesis. In. Chromatin and epigenetic regulation of animal development. p. 155-
1183 168.
1184 Rathke C, Baarends WM, Jayaramaiah-Raja S, Bartkuhn M, Renkawitz R, Renkawitz-Pohl R.
1185 2007. Transition from a nucleosome-based to a protamine-based chromatin

1186 configuration during spermiogenesis in *Drosophila*. Journal of cell science 120:1689-
1187 1700.
1188 Rathke C, Barckmann B, Burkhard S, Jayaramaiah-Raja S, Roote J, Renkawitz-Pohl R. 2010.
1189 Distinct functions of Mst77F and protamines in nuclear shaping and chromatin
1190 condensation during *Drosophila* spermiogenesis. European Journal of Cell Biology
1191 89:326-338.
1192 Ravi Ram K, Wolfner MF. 2007. Sustained post-mating response in *Drosophila melanogaster*
1193 requires multiple seminal fluid proteins. PLoS Genetics 3:e238.
1194 Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. 2013. *De novo* ORFs in
1195 Drosophila are important to organismal fitness and evolved rapidly from previously non-
1196 coding sequences. PLoS Genetics 9:e1003860.
1197 Reinke V, San Gil I, Ward S, Kazmer K. 2004. Genome-wide germline-enriched and sex-biased
1198 expression profiles in *Caenorhabditis elegans*. Development 131:311-323.
1199 Rele CP, Sandlin KM, Leung W, Reed LK. 2020. Manual Annotation of Genes within *Drosophila*
1200 Species: the Genomics Education Partnership protocol. bioRxiv 12.10.420521.
1201 Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, Marqués-
1202 Bonet T, Albà MM. 2016. Origins of *de novo* genes in human and chimpanzee. PLoS
1203 Genetics 11:e1005721.
1204 Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, Rohde C,
1205 Valente VLS, Aguadé M, Anderson WW, et al. 2008. Polytene Chromosomal Maps of 11
1206 Drosophila Species: The Order of Genomic Scaffolds Inferred From Genetic and
1207 Physical Maps. Genetics 179:1601-1655.
1208 Schlötterer C. 2015. Genes from scratch – the evolutionary fate of *de novo* genes. Trends in
1209 Genetics 31:215-219.
1210 Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient *de novo* genes can evolve from
1211 frozen accidents that escaped rapid transcript turnover. Nature Ecology & Evolution
1212 2:1626-1632.
1213 Schnakenberg SL, Matias WR, Siegal ML. 2011. Sperm-storage defects and live birth in
1214 *Drosophila* females lacking spermathecal secretory cells. PLOS Biology 9:e1001192.
1215 Schuh M, Lehner CF, Heidmann S. 2007. Incorporation of *Drosophila* CID/CENP-A and CENP-
1216 C into Centromeres during Early Embryonic Anaphase. Current Biology 17:237-243.
1217 Slater G, Birney E. 2005. Automated generation of heuristics for biological sequence
1218 comparison. BMC Bioinformatics 6:31.
1219 Sorourian M, Kunte MM, Domingues S, Gallach M, Özdil F, Río J, Betrán E. 2014. Relocation
1220 facilitates the acquisition of short *cis*-regulatory regions that drive the expression of
1221 retrogenes during spermatogenesis in *Drosophila*. Molecular Biology and Evolution
1222 31:2170-2180.
1223 Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, Barthès P, Kokkinaki M, Nef S,
1224 Gnirke A, et al. 2013. Cellular source and mechanisms of high transcriptome complexity
1225 in the mammalian testis. Cell Reports 3:2179-2190.
1226 Steinhauer J. 2015. Separating from the pack: molecular mechanisms of Drosophila spermatid
1227 individualization. Spermatogenesis 5:e1041345.
1228 Suenaga Y, Islam SMR, Alagu J, Kaneko Y, Kato M, Tanaka Y, Kawana H, Hossain S,
1229 Matsumoto D, Yamamoto M, et al. 2014. NCYM, a *cis*-antisense gene of MYCN,
1230 encodes a *de novo* evolved protein that inhibits GSK3β resulting in the stabilization of
1231 MYCN in human neuroblastomas. PLoS Genetics 10:e1003996.
1232 Tirmarche S, Kimura S, Sapey-Triomphe L, Sullivan W, Landmann F, Loppin B. 2014.
1233 Drosophila Protamine-Like Mst35Ba and Mst35Bb Are Required for Proper Sperm
1234 Nuclear Morphology but Are Dispensable for Male Fertility. G3
1235 Genes|Genomes|Genetics 4:2241-2245.

1236 Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, Medetgul-Ernar K,
1237 Bowman RW, Hines CP, Iannotta J, et al. 2020. De novo emergence of adaptive
1238 membrane proteins from thymine-rich genomic sequences. Nature Communications 11.
1239 Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, Coon JJ, Lafontaine I.
1240 2018. A molecular portrait of de novo genes in yeasts. Molecular Biology and Evolution
1241 35:631-645.
1242 Van Oss SB, Carvunis A-R. 2019. *De novo* gene birth. PLoS Genetics 15:e1008160.
1243 VanKuren NW, Long M. 2018. Gene duplicates resolving sexual conflict rapidly evolved
1244 essential gametogenesis functions. Nature Ecology & Evolution 2:705-712.
1245 Vibranovski MD. 2014. Meiotic Sex Chromosome Inactivation in *Drosophila*. Journal of
1246 Genomics 2:104-117.
1247 Vibranovski MD, Lopes HF, Karr TL, Long M. 2009. Stage-specific expression profiling of
1248 *Drosophila* spermatogenesis suggests that meiotic sex chromosome inactivation drives
1249 genomic relocation of testis-expressed genes. PLoS Genetics 5:e1000731.
1250 Vibranovski MD, Zhang Y, Long M. 2009. General gene movement off the X chromosome in the
1251 *Drosophila* genus. Genome Research 19:897-903.
1252 Vibranovski MD, Zhang YE, Kemkemer C, Lopes HF, Karr TL, Long M. 2012. Re-analysis of the
1253 larval testis data on meiotic sex chromosome inactivation revealed evidence for tissue-
1254 specific gene expression related to the drosophila X chromosome. BMC Biology 10:49.
1255 Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, Polpitiya A, Petritis K,
1256 Karr TL. 2010. The *Drosophila melanogaster* sperm proteome-II (DmSP-II). Journal of
1257 Proteomics 73:2171-2185.
1258 Weisman CM, Murray AW, Eddy SR. 2020. Many, but not all, lineage-specific genes can be
1259 explained by homology detection failure. PLOS Biology 18:e3000862.
1260 Wen K, Yang L, Xiong T, Di C, Ma D, Wu M, Xue Z, Zhang X, Long L, Zhang W, et al. 2016.
1261 Critical roles of long noncoding RNAs in *Drosophila* Spermatogenesis. Genome
1262 Research 26:1233-1244.
1263 White-Cooper H. 2010. Molecular mechanisms of gene regulation during Drosophila
1264 spermatogenesis. REPRODUCTION 139:11-21.
1265 White-Cooper H. 2004. Spermatogenesis: analysis of meiosis and morphogenesis. In:
1266 Henderson D, editor. Drosophila Cytogenetics Protocols. Totowa, NJ: Humana Press. p.
1267 45-75.
1268 White-Cooper H. 2012. Tissue, cell type and stage-specific ectopic gene expression and RNAi
1269 induction in the Drosophila testis. Spermatogenesis 2:11-22.
1270 Wilburn DB, Swanson WJ. 2016. From molecules to mating: Rapid evolution and biochemical
1271 studies of reproductive proteins. Journal of Proteomics 135:12-25.
1272 Wilson BA, Foy SG, Neme R, Masel J. 2017. Young genes are highly disordered as predicted
1273 by the preadaptation hypothesis of *de novo* gene birth. Nature Ecology & Evolution
1274 1:0146.
1275 Witt E, Benjamin S, Svetec N, Zhao L. 2019. Testis single-cell RNA-seq reveals the dynamics of
1276 de novo gene transcription and germline mutational bias in *Drosophila*. eLife 8.
1277 Wong A, Albright SN, Giebel JD, Ram KR, Ji S, Fiumera AC, Wolfner MF. 2008. A role for
1278 Acp29AB, a predicted seminal fluid lectin, in female sperm storage in *Drosophila*
1279 *melanogaster*. Genetics 180:921-931.
1280 Wu DD, Irwin DM, Zhang YP. 2011. De novo origin of human protein-coding genes. PLoS
1281 Genetics 7.
1282 Xiao W, Liu H, Li Y, Li X, Xu C, Long M, Wang S. 2009. A rice gene of *de novo* origin negatively
1283 regulates pathogen-induced defense response. PLOS ONE 4:e4603.
1284 Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, Ullrich KK, Tautz D.
1285 2019. A de novo evolved gene in the house mouse regulates female pregnancy cycles.
1286 eLife 8.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24:1586-1591.

Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, Abbassi M, Gandasetiawan KA, Librado P, Damia E, et al. 2012. Functional evidence that a recently evolved Drosophila sperm-specific gene boosts sperm competition. Proceedings of the National Academy of Sciences 109:2043-2048.

Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, Seppey M, Loetscher A, Kriventseva EV. 2017. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. Nucleic acids research 45:D744-D749.

Zhang J. 2003. Evolution by gene duplication: an update. Trends in Ecology & Evolution 18:292-298.

Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, Yu Y, Hou G, Zi J, Zhou R, et al. 2019. Rapid evolution of protein diversity by de novo origination in *Oryza*. Nature Ecology & Evolution 3:679-690.

Zhang W, Landback P, Gschwend AR, Shen B, Long M. 2015. New genes drive the evolution of gene interaction networks in the human and mouse genomes. Genome biology 16.

Zhang YE, Vibranovski MD, Krinsky BH, Long M. 2010. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. Genome Research 20:1526-1533.

Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. 2010. Chromosomal Redistribution of Male-Biased Genes in Mammalian Evolution with Two Bursts of Gene Gain on the X Chromosome. PLOS Biology 8:e1000494.

Zhao L, Saelao P, Jones CD, Begun DJ. 2014. Origin and spread of *de novo* genes in *Drosophila melanogaster* populations. Science 343:769-772.

Zhuang X, Yang C, Murphy KR, Cheng CHC. 2019. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. Proceedings of the National Academy of Sciences 116:4400-4405.