

RESEARCH ARTICLE

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

Emily L. Rivard¹^{✉aa}, Andrew G. Ludwig¹[✉], Prajal H. Patel¹[✉], Anna Grandchamp², Sarah E. Arnold¹^{aa}, Alina Berger²², Emilie M. Scott¹¹, Brendan J. Kelly¹¹^{bb}, Grace C. Mascha¹¹, Erich Bornberg-Bauer^{2,3}^{2,3}, Geoffrey D. Findlay¹^{1*}

1 College of the Holy Cross, Worcester, Massachusetts, United States of America, **2** University of Münster, Münster, Germany, **3** Max Planck Institute for Developmental Biology, Tübingen, Germany

 These authors contributed equally to this work.

^{aa} Current address: Harvard University, Cambridge, Massachusetts, United States of America

^{bb} Current address: The Ohio State University, Columbus, Ohio, United States of America

* gfindlay@holycross.edu



OPEN ACCESS

Citation: Rivard EL, Ludwig AG, Patel PH, Grandchamp A, Arnold SE, Berger A, et al. (2021) A putative *de novo* evolved gene required for spermatid chromatin condensation in *Drosophila melanogaster*. PLoS Genet 17(9): e1009787. <https://doi.org/10.1371/journal.pgen.1009787>

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, UNITED STATES

Received: June 29, 2021

Accepted: August 19, 2021

Published: September 3, 2021

Copyright: © 2021 Rivard et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its [Supporting Information](#) files.

Funding: This work was supported by NSF (<https://nsf.gov>) CAREER Award #1652013 (to GDF) and an Alexander von Humboldt Foundation (<https://www.humboldt-foundation.de/en/>) Fellowship (to AG). This work relied on resources provided by the Genomics Education Partnership, which is supported by NSF grant #1915544, NIH (<https://nih.gov>) grant #R25GM130517, and hosted by The

Abstract

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

Author summary

Genomes are in flux, as genes are constantly added and lost throughout evolution. New genes were once thought to arise almost exclusively via the modification or duplication of existing genes. Recently, however, interest has grown in alternative modes of new gene origination, such as *de novo* evolution from genetic material that previously did not encode proteins. Many *de novo* genes are expressed in male reproductive tissues, but their significance for fertility is not well understood. We screened likely *de novo* genes

University of Alabama and Washington University in St. Louis. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

expressed in the *Drosophila* testis for reproductive roles and found one gene, *atlas*, essential for male fertility. We leveraged genetic and cell biological experiments to investigate roles for Atlas protein in reproduction and found that it is required during sperm development for proper packaging of DNA in the sperm nucleus. Evolutionary analyses of this gene revealed a complicated history, including loss in some lineages, movement between chromosomes, and fusion with a non-protein-coding gene. Studying both the functions and evolutionary histories of new proteins illustrates how they might evolve critical roles in biological processes despite their relative novelty. Furthermore, the study of *atlas* identifies an essential genetic player in the fly testis, an important model system for understanding how gametes are produced.

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 [1,2]; 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 [3–5]. 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 [6–13]. Multiple models explain how protein-coding *de novo* genes may acquire both an open reading frame (ORF) and regulatory sequences permitting transcription [14–17]. 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 [17–20].

The capacity of protein-coding *de novo* genes to evolve important functions is a topic of interest from evolutionary, physiological and molecular perspectives [21]. 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 [22,23]. *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 [24]. 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 [25,26]. *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 [27]. 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 [28]. 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 [12,29–34]. Proposed causes of this “out of the testis” phenomenon include the high level of promiscuous transcription in testis cells [35,36], the relative simplicity of promoter regions driving expression in the testis [37], and preferential retention of novel genes with male-biased expression [38]. Sexual selection also drives rapid evolution of reproductive proteins [39] and could drive the emergence of new

genes as a mechanism of improving male reproductive ability [33]. 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 [40]. 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 [41,42]. As observed in other biological systems, *Drosophila de novo* genes retained by selection demonstrate enriched expression in the testis [10,19,33,34]. The expression patterns of emerging *de novo* genes in the *Drosophila* testis were recently analyzed at single cell resolution [43], 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 [44] and CRISPR/Cas9-based [45] 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 [46]. *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 [20]. These data suggested that putative *de novo* genes can evolve essential roles in a rapidly evolving reproductive process, spermatogenesis.

Here, we expanded 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 found 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* [46]. This result, and other

CRISPR genome editing indicated that neither gene plays a role in male fertility. In these cases, the RNAi phenotypes might have been due to off-target knockdown.

CRISPR-mediated gene mutation validates *atlas* RNAi results

We validated the observed fertility defect by using CRISPR/Cas9-based genome editing to construct putative loss-of-function alleles for *atlas* (S2 Fig). The principal allele we used for validation and the functional studies described below was a null allele that completely deletes the *atlas* genomic locus. This allele was generated by targeting each end of the locus with a gRNA. We made three additional frameshift alleles by inducing double-stranded breaks at a gRNA target site just downstream of the *atlas* start codon, which induced non-homologous end joining. Males homozygous for the *atlas* deletion allele have the same fertility defect as knockdown males (Fig 2A). Males homozygous for any of three frameshift alleles showed significantly reduced, but non-zero, fertility (Fig 2B). It is possible that residual *atlas* function may be present in these animals, perhaps due to translation initiation at a downstream start codon, which could generate a shorter protein with partial function. Each frameshift allele retains the possibility of encoding an N-terminally truncated, but otherwise in-frame, version of Atlas protein that would lack the first 60 amino acids (out of 172; see S2 Fig). Alternatively, it is possible that the residual fertility of the frameshift alleles is caused by the gene's intact 3' UTR, a topic we discuss in more detail below. Finally, we constructed a genomic rescue construct carrying both the full *atlas* transcribed region and its native regulatory sequences. *atlas* null males that carried a single copy of the rescue construct had fully restored fertility (S3 Fig). Overall, these data demonstrate that *atlas* loss, and not an RNAi or CRISPR off-target, causes nearly complete male sterility.

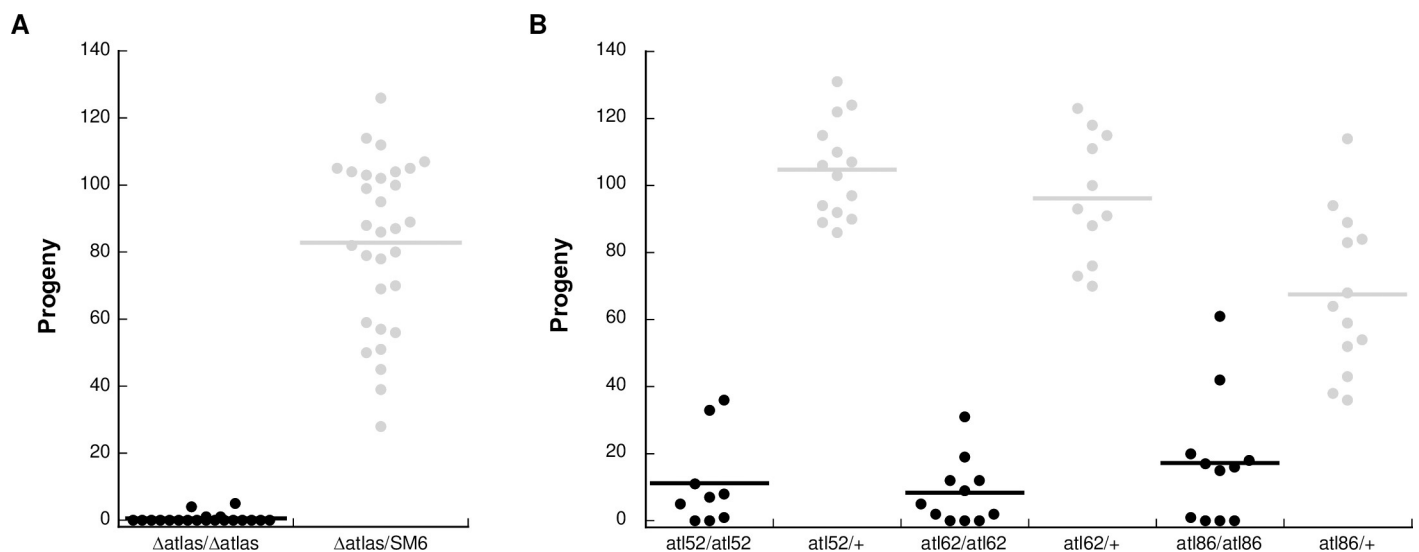


Fig 2. CRISPR-generated deletion and frameshift alleles of *atlas* confirm the gene's requirement for male fertility. A) Single-pair fertility assay for males homozygous for the null ($\Delta atlas$) allele or heterozygous controls ($\Delta atlas/SM6$). Flies homozygous for the deletion had significantly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

<https://doi.org/10.1371/journal.pgen.1009787.g002>

***atlas* is required for proper spermatid nuclear condensation**

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

We next examined two post-meiotic processes: individualization of 64-cell spermatid cysts into mature sperm, and spermatid nuclear condensation. Individualization initiates when an actin-rich individualization complex (IC) associates with the bundle of spermatid nuclei. The IC then proceeds down the sperm tails, expelling cytoplasmic waste and remodeling cell membranes to form 64 individual sperm. We visualized this process in males 0–1 days old, when spermatogenesis occurs at high levels, by staining whole mount testes for actin (Fig 3C and 3D). Although ICs associated with nuclear bundles present at the basal end of the testes in both control and *atlas* null males, we observed significantly fewer nuclear bundle-associated ICs in nulls (Fig 3C and 3D). While control testes typically had several ICs progressing down sperm tails, we saw a significantly reduced proportion of progressed bundles in nulls (Fig 3C and 3D). In some null testes, we also observed individual investment cones dissociated from progressing ICs (Fig 3C).

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

That condensing spermatid nuclei are misshapen in the absence of *atlas* suggests the possibility that Atlas protein is critical for nuclear condensation. This idea is further supported by its predicted biochemical properties. Previously characterized spermatid chromatin binding proteins are small and highly basic [53,56,57], as the excess of positively charged amino acid side chains facilitates ionic interactions with negatively charged DNA. Many such proteins

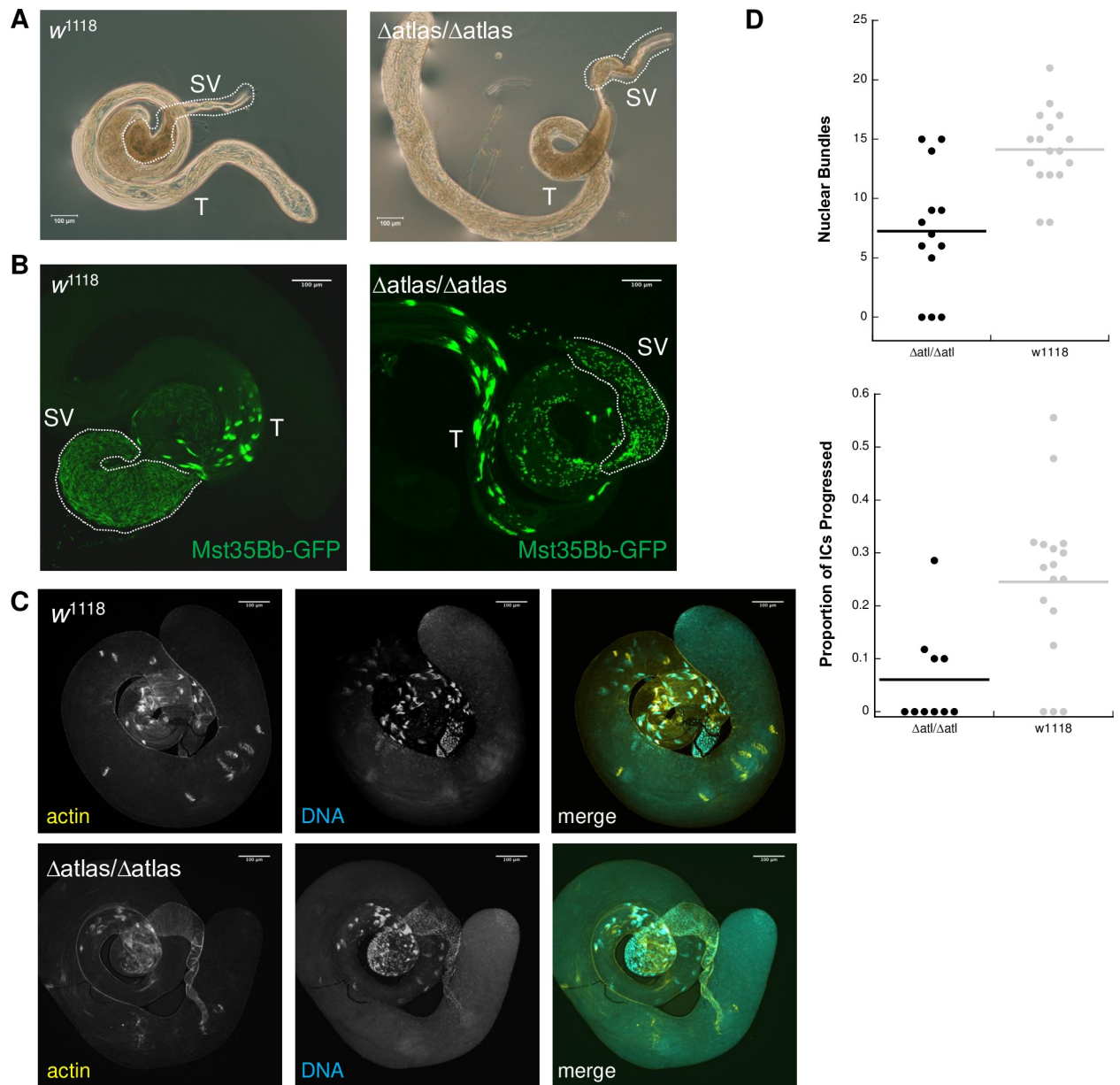


Fig 3. Cytological investigations of the *atlas* mutant fertility defect. A) Phase contrast microscopy of male reproductive tracts dissected from 7-day-old, unmated control (w^{1118}) or *atlas* null males. Control males show the expected accumulation of sperm tails in the seminal vesicle (SV), which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

<https://doi.org/10.1371/journal.pgen.1009787.g003>

(i.e., Tpl94D, Mst35Ba, Mst35Bb, Prt199C and Mst77F) also contain a conserved protein domain, the high-mobility-group box (HMG-box) domain [55,58–64], suggesting that this type of chromatin binding protein could have originated through gene duplication and divergence.

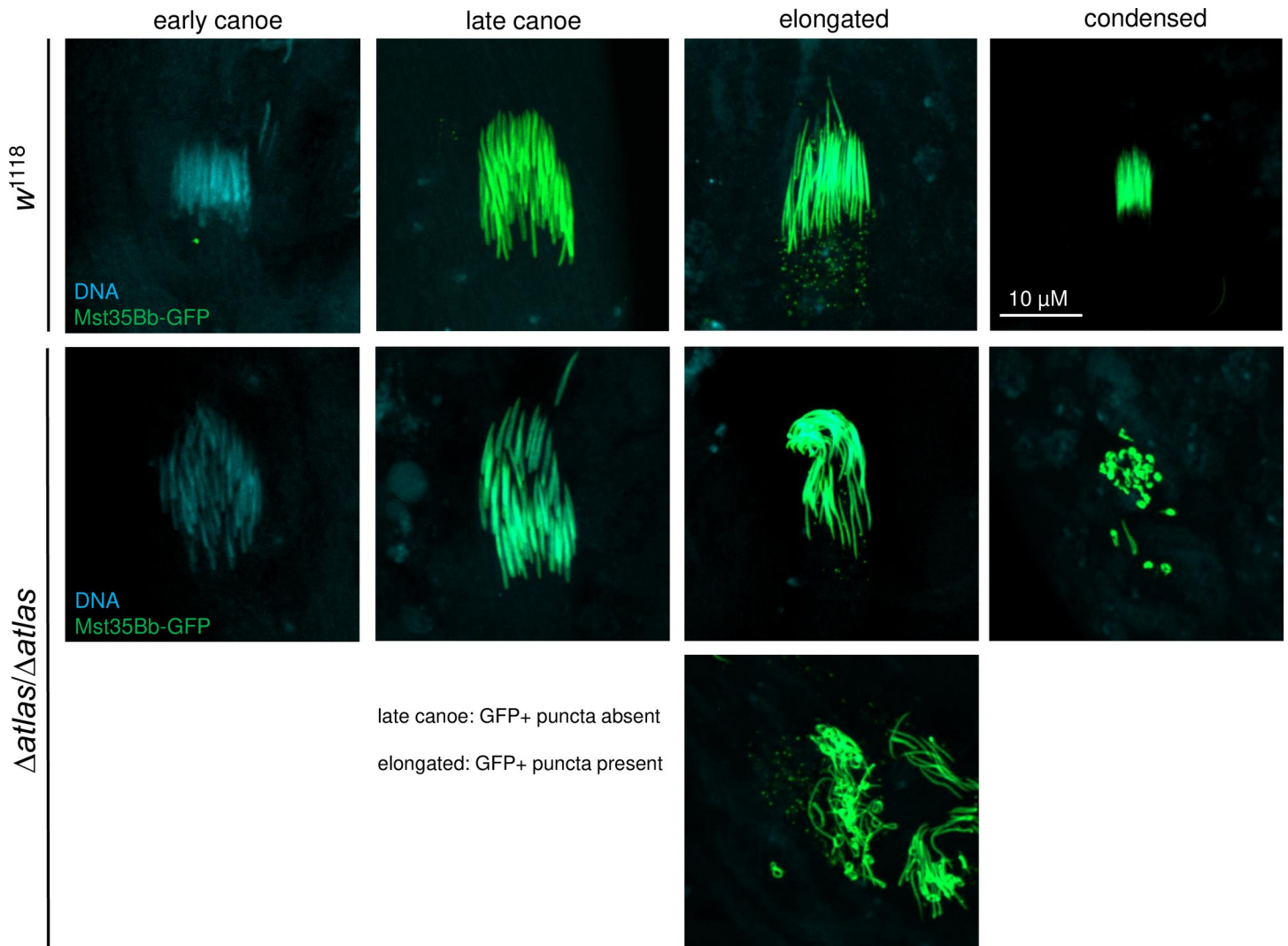


Fig 4. *atlas* null males show aberrant nuclear shaping at and beyond the elongated stage of spermatid nuclear condensation. Early and late canoe stages were distinguished by the absence or presence of Mst35Bb-GFP, respectively. Late canoe and elongated stages were distinguished by the absence or presence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

<https://doi.org/10.1371/journal.pgen.1009787.g004>

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) [65]. To test the hypothesis that Atlas is nuclear localized, and could thus bind DNA, we generated an *atlas*-GFP transgene under UAS control and expressed it ubiquitously using *tubulin*-GAL4 and in the early male germline using *Bam*-GAL4. In both larval salivary glands and early male germline cells, Atlas-GFP appeared to be nuclear localized ([S5 Fig](#)).

While these results were consistent with Atlas protein localizing to the nucleus, they did not allow us to visualize Atlas in the cells in which it is normally expressed. To do so, we used CRISPR/Cas9-induced homology directed repair (<https://flycrispr.org/scarless-gene-editing/>) [66–68] to create an *atlas*-GFP fusion at the endogenous *atlas* locus (see [S6 Fig](#) and [Materials and Methods](#)). We first confirmed the functionality of the knock-in allele by showing that

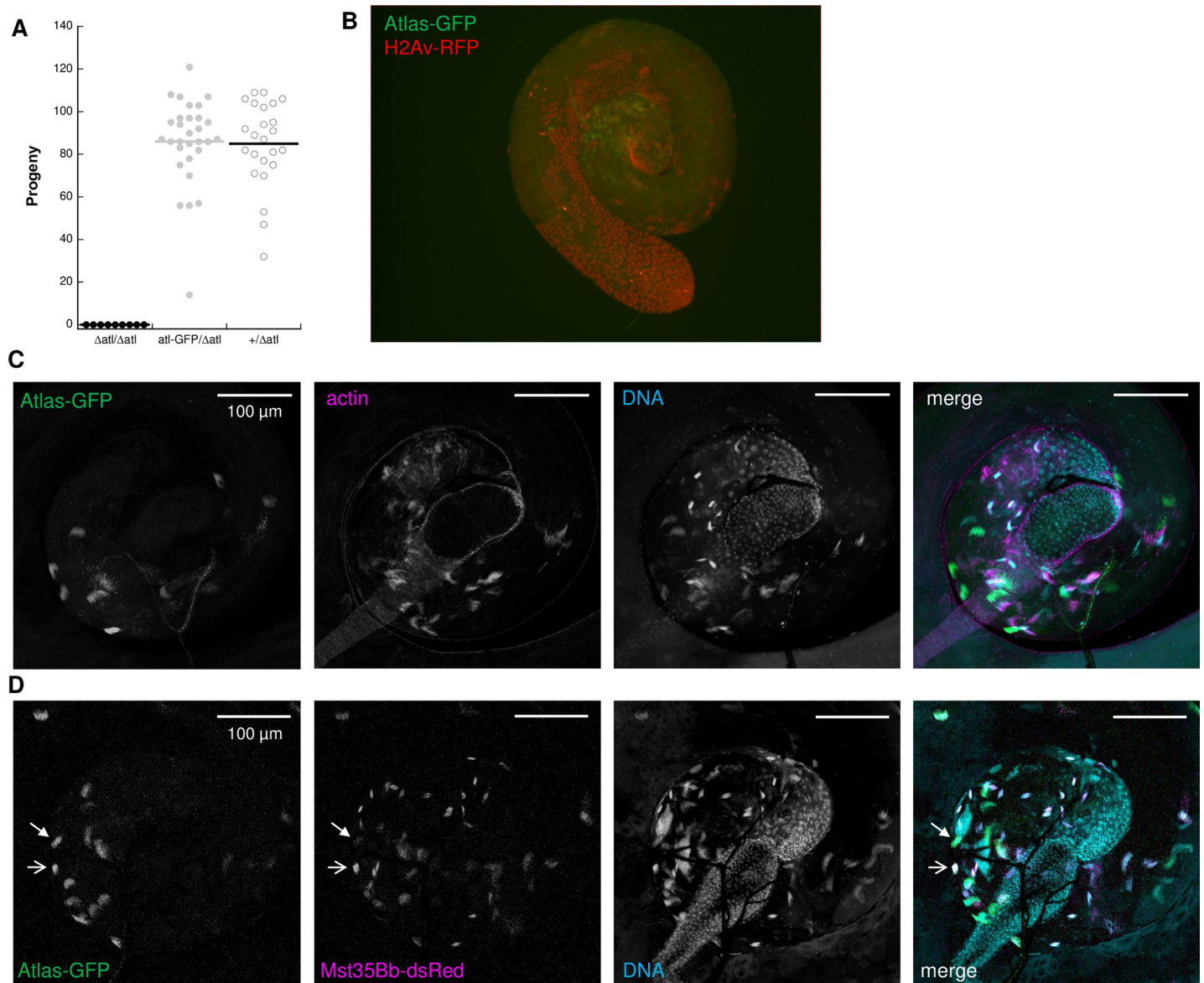


Fig 5. An *atlas*-GFP allele generated at the endogenous *atlas* locus is fully functional for male fertility and encodes a protein that localizes to condensing spermatid nuclei. A) A single copy of the *atlas*-GFP allele is sufficient for full fertility when paired with the Δ *atlas* allele, as compared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

<https://doi.org/10.1371/journal.pgen.1009787.g005>

males with the *atlas* locus genotype *atlas*-GFP/ Δ *atlas* had equivalent fertility to males of genotype *atlas*+/ Δ *atlas* (Fig 5A). We then visualized Atlas-GFP fusion protein in whole-mount testes in conjunction with phalloidin-stained actin (Figs 5C and S7). Atlas-GFP was absent from seminal vesicles, consistent with its absence from the proteome of mature *D. melanogaster* sperm [69,70]. Instead, Atlas-GFP colocalized with condensing nuclear bundles near the basal

end of the testes (Fig 5C). Actin-based ICs were also observed in the basal testes, but generally did not co-localize with Atlas-GFP, suggesting that Atlas-GFP is present in condensing nuclei before IC association (Fig 5C). This result, taken together with the aberrant nuclear condensation in the absence of *atlas* (Fig 4), is consistent with the idea that Atlas is a transition protein. Transition proteins are chromatin components that act transiently during spermatid nuclear condensation. A series of transition proteins first replace histones as the primary DNA binding proteins in the nucleus and then give way to protamines, the proteins that package chromatin in mature sperm [53,55,58].

To further elucidate the role of *atlas* in nuclear condensation, we next examined Atlas-GFP localization in the presence of either an early spermatid nuclear marker, histone H2Av-RFP [55,71], or Mst35Bb-dsRed [51], a marker of nuclei from the late canoe stage through final condensation. Atlas-GFP showed no co-localization with H2Av-RFP, suggesting that Atlas functions after histone removal (Fig 5B). In contrast, some GFP-positive bundles co-localized with Mst35Bb-dsRed, but others did not (Figs 5D and S7B). These data suggest that Atlas may be one of the final transition proteins used in nuclear condensation before the chromatin becomes fully condensed with protamines.

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 (Figs 3 and 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 and 6B). 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 S7 Fig). 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 [53] 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 [53]. 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.

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 [46,72]. 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 [73], as assessed by synteny, in all *Drosophila* species examined, including distantly related species such as *D. virilis*

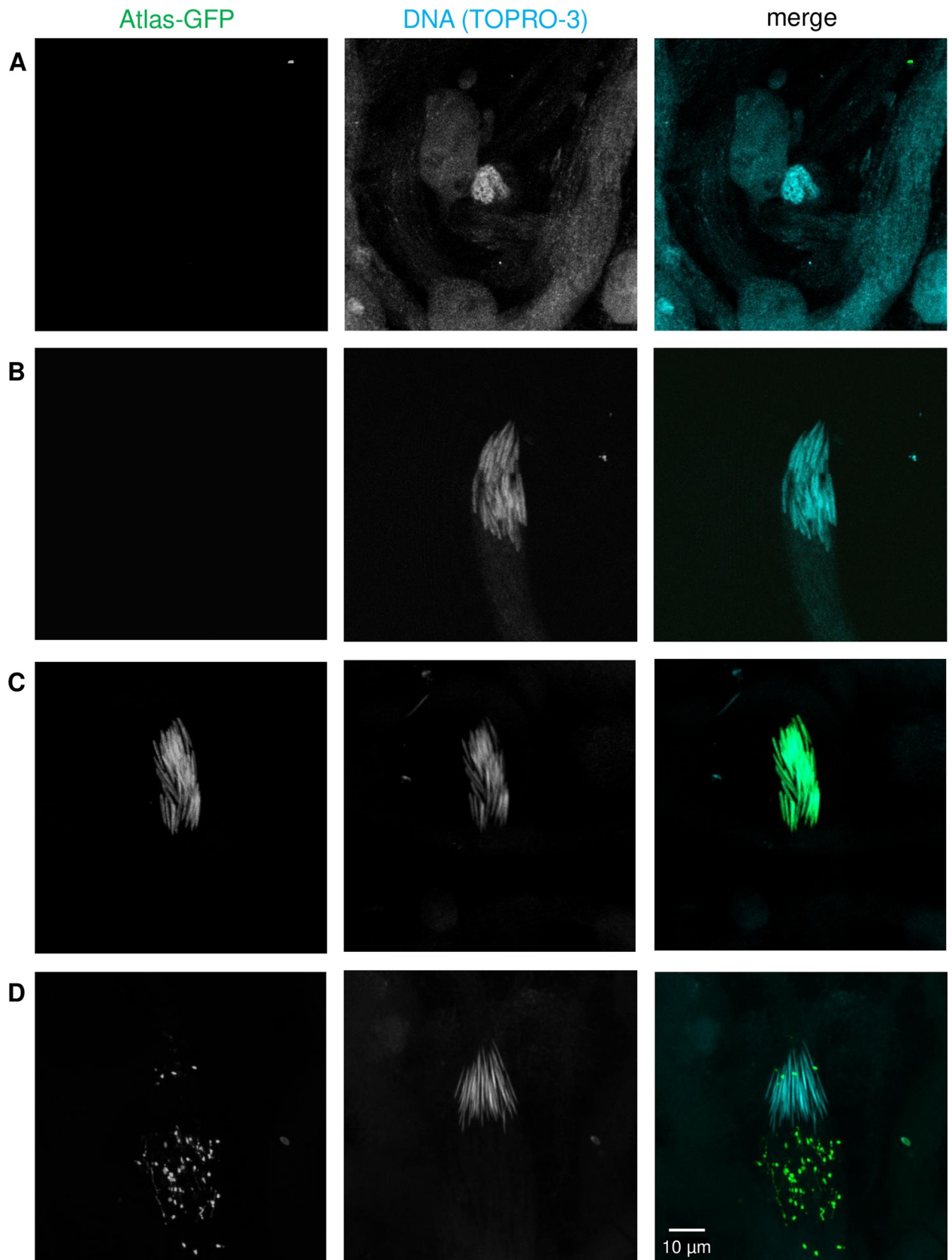


Fig 6. Atlas-GFP is present in late canoe-stage spermatid nuclei and then appears to leave the nucleus in puncta. Staging of condensing spermatid nuclei fixed in paraformaldehyde from *atlas*-GFP males stained with TO-PRO-3 DNA stain. Atlas-GFP is not detectable in (A) round stage or (B) early canoe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

<https://doi.org/10.1371/journal.pgen.1009787.g006>

and *D. grimshawi* (Figs 7 and S8). The protein-coding first exon shows 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 (S2 Table). 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 *D. virilis* in a partially syntenic region on the same Muller A element (S2 Table and S9 Fig). 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.

To confirm the lack of *atlas* protein-coding sequences identifiable by BLASTP or TBLASTN in most non-*melanogaster* group species, we identified the regions syntenic to those containing *atlas* in *D. ananassae* and *D. virilis* in 11 additional *Drosophila* species and used more sensitive methods to search for potential orthologs [72]. Specifically, we: a) relaxed the BLAST cut-offs for detection, since default parameters can cause false-negative results when searching for potential *de novo* genes in divergent species [74]; b) used adult male RNA-seq data to detect transcribed areas within each syntenic region that did not match annotated genes; and, c) predicted the isoelectric point of the potential proteins encoded, under the

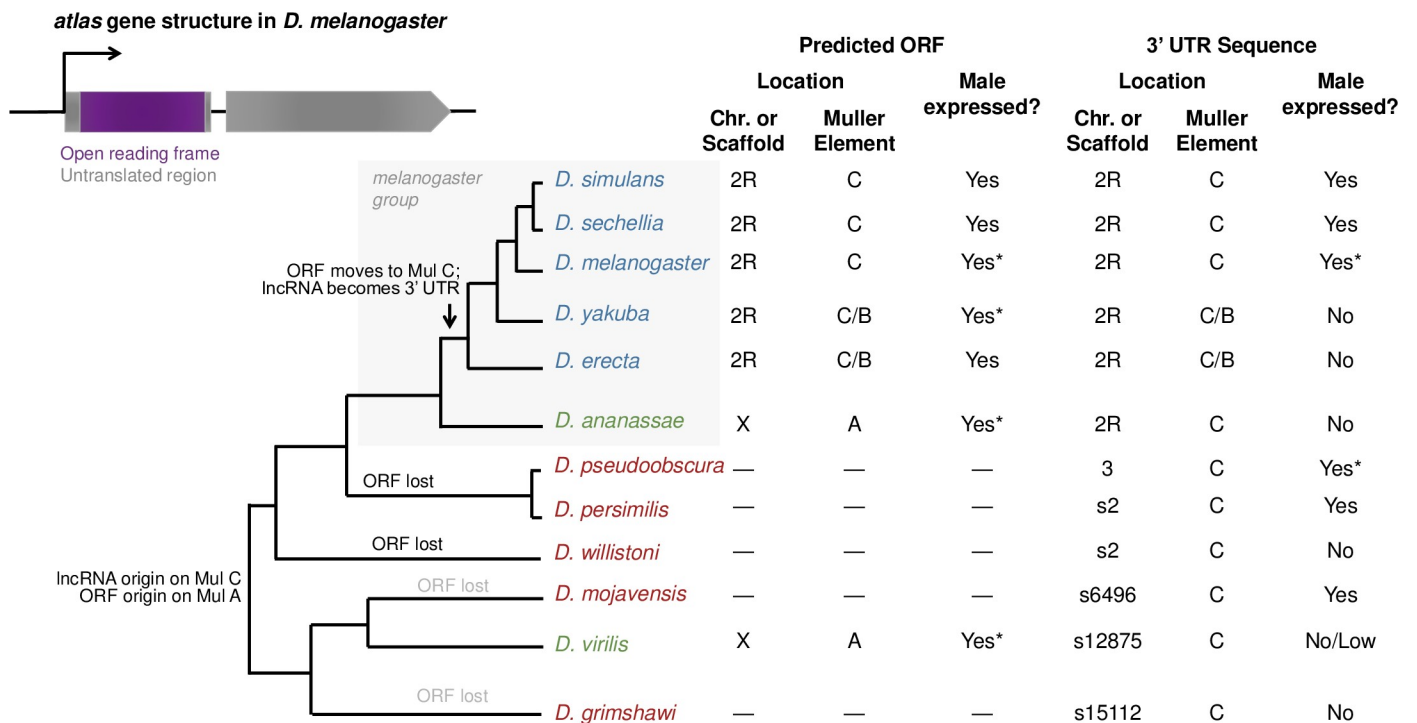


Fig 7. Molecular evolution and gene expression of *atlas* across the *Drosophila* genus. The gene structure of *atlas* in *D. melanogaster* is shown at top left. The predicted protein-coding sequence is contained entirely within exon 1, while exon 2 encodes the presumed 3' UTR. The gene is located on 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 S8 and S10 Figs.

<https://doi.org/10.1371/journal.pgen.1009787.g007>

hypothesis that Atlas orthologs would have conserved, DNA-binding functions. The results are summarized in [S9 Fig](#) and [S2 Table](#). 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 have 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 ([S9 Fig](#)). 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 ([S2 Table](#)). 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 ([S2 Table](#)). 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 (thegep.org) [72] and verified several of these results by RT-PCR ([Figs 7](#) and [S10](#)). In all species in which *atlas* was detected, the protein-coding region is expressed specifically in males regardless of its genomic location ([Fig 7](#)). Interestingly, the non-coding region shows 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 ([S8 Fig](#)). 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 ([S8 Fig](#)), 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 [46], 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* (S2 Table). 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 S2 Table are considered, though an additional loss in the *melanogaster* group is likely due to the absence of a detectable ortholog in *D. kikkawai* and *D. serrata*. As noted above, the pattern of gene loss can also appear due to orthology detection failure [74], 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 S11 Fig, 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 [75] 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 [76]. We also note that a previous HMMER-based analysis to identify HMG-box-containing spermatid chromatin condensation proteins in *Drosophila* and other related insects did not detect *atlas* as a member of the gene family [64], further diminishing the possibility that *atlas* is a divergent paralog of other transition and protamine-like proteins. 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 [77,78]. 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 , $\omega = 0.41$ by PAML model M0), there was no significant evidence for positive selection acting to recurrently diversify a subset of sites within the protein (Table 1).

Table 1. PAML sites tests for positive selection acting on *atlas* in the *melanogaster* group.

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

<https://doi.org/10.1371/journal.pgen.1009787.t001>

Discussion

Across taxa, many *de novo* evolved genes are expressed in the male reproductive system [12,31,33,34]. Identifying those genes that have evolved essential roles in reproduction will provide insight into how newly evolved genes integrate with existing cellular networks [47] 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.

Loss of *atlas* 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 conservation over the last ~15 million years.

Comparison of *atlas* null and frameshift allele phenotypes

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

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

to fertility in a unique way. Future experiments that could distinguish these possibilities are described below.

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

Atlas is an essential transition protein

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 5). Its lack of overlap with testis-specific histones (Fig 5B), partial overlap with Mst35Bb (Fig 5D), removal from needle-stage nuclei (Fig 6) and absence from mature sperm (Figs 5 and 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 [53,58,59]. 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 [79], but otherwise match *Atlas* in their biochemical properties, transient expression, 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 [53,58,59]. 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 [57,64]. In *D. melanogaster*, protamine-like proteins include Mst35Ba, Mst35Bb, Prtl99C and Mst77F [63,80,81]. Interestingly, while all characterized protamine-like proteins are present in mature sperm, only some are essential for fertility. Knockouts of *Mst35Ba*, *Mst35Bb*, or both show occasional nuclear shaping defects, but male fertility is normal [80,82]. In contrast, mutants of *Prtl99C* or *Mst77F* are sterile. *Prtl99C* and *Mst35Ba/b* bind condensed DNA independently of each other and contribute additively to the shortening of needle-stage nuclei, but *Prtl99C*'s effect is ~3-fold greater [63,64]. This difference is apparently great enough to reduce fertility only in *Prtl99C* mutants. In contrast, *Mst77F* and *Mst35Ba/b* show a genetic interaction, as *Mst35Ba/b* null flies become nearly sterile in an *Mst77F* heterozygous background [81]. Furthermore, while Mst35Bb-GFP is expressed in *Mst77F* nulls, these flies show deformed spermatid nuclei that do not reach a recognizable needle-like stage. Because *atlas* nulls show considerable phenotypic similarity to *Mst77F* nulls, but not *Prtl99C* nulls, we hypothesize that *Atlas* may act in a pathway with *Mst77F*. Our observation of inefficient IC movement down sperm tails in *atlas* null testes is reminiscent of a similar phenotype in *Mst77F* nulls [81], providing further evidence that these proteins may act in a common

pathway. In both cases, ICs can form at misshapen canoe-stage nuclei, but fail at a subsequent step. While the exact relationship between nuclear abnormalities in the late canoe stage and individualization is not entirely understood, it is possible that nuclear shape and the organization of nuclear bundles impact the ability of IC association and IC progression, as is also observed in mutants of another gene, *dPSMG1*, which controls nuclear shape [59].

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

Because *de novo* genes emerge from non-coding sequences, they typically encode proteins that are short and lack complex structure [3,16]. Indeed, expanding the length of the protein-coding region and evolving higher-level protein structures are hypothesized to be among the final stages of new gene evolution [4]. In light of these constraints, what kinds of cellular functions might be available to newly evolved proteins? Vakirlis et al. [83] overexpressed emerging proto-genes in *S. cerevisiae* and found that those encoding proteins with predicted transmembrane (TM) domains were more likely to be adaptive, as assessed by the effect of proto-gene overexpression on growth rate. Such proteins may arise when thymine-rich intergenic regions undergo mutations that allow protein-coding gene birth and expression, since many codons with multiple U nucleotides encode amino acids commonly found in TM domains [83]. Our imaging data, in addition to the prediction tools employed by Vakirlis et al. [83], suggest that Atlas does not contain a TM domain. However, just as the amino acid compositional requirements of a TM domain are not overly complex, neither are those of DNA binding proteins. In essence, these proteins must simply be small, have a high concentration of positively charged residues, and contain a nuclear localization signal, which itself requires a small patch of positively charged residues [65]. Thus, DNA binding proteins may be a relatively easy class of protein to evolve *de novo*.

While many putative *de novo* genes are expressed in the *D. melanogaster* testis [19], *atlas* was the only verified hit from our screen that was essential for male fertility (Fig 1). This result raises two related questions. First, why are so many of these other genes expressed if their knockdown causes no obvious effect? In general, it is common for the knockdown of protein-coding genes expressed in reproductive tissues in *D. melanogaster* to result in no detectable fertility defects [46,84–86]. One set of explanations center on technical issues: a gene could be expressed outside of the cells targeted by the RNAi driver, knockdown level could be insufficient to cause a phenotype, and/or a gene may be functionally redundant such that its knockdown causes no apparent effect. Another hypothesis to explain this pattern is that while the loss of function of such genes may cause small reductions in fertility that would be subject to strong negative selection in nature, the conditions used to assay such knockdown animals in primary screens are rarely tailored to detect differences of this magnitude. A third possibility, not mutually exclusive with the above, is that while some genes may be expendable in non-competitive, non-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 [87–89]. These latter two possibilities are illustrated by some of the other spermatid chromatin binding proteins previously characterized as “non-essential” (e.g., *Mst35Ba/Bb* and *Tpl94D*) [58,80]. While these proteins are not *de novo* evolved, they provide examples of genes whose mutations cause aberrant cellular phenotypes, such as the abnormally shaped spermatid nuclei in *Mst35Ba/Bb*, and that may contribute to fertility defects when placed in a sensitized genetic background [81].

It is also possible that some of the genes we screened are *de novo* genes that have no functional effects and have thus never been selectively advantageous. For example, 8 of the 42

screened genes have annotated orthologs in either *D. melanogaster* alone or in only *D. melanogaster* and *D. simulans*. Such recently “born” *de novo* genes could have been propagated by neutral evolutionary processes but not yet experienced inactivating mutations that would abolish their coding sequences and/or expression. This pattern would be consistent with *de novo* genes’ high rates of both birth and death observed previously in *Drosophila* [38] and yeast [83].

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 that contribute to DNA packaging in mature sperm [57] are found across all sequenced *Drosophila* species [61], and orthologs are also reported in FlyBase from other Dipteran and non-Dipteran insects [90]. However, transition protein Tpl94D is reported to be restricted to species ranging from *D. melanogaster* to *D. pseudoobscura* [60], as are the related proteins tHMG1 and tHMG2 with high-mobility group domains [58,90]. 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 spermiogenesis 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 instance, *D. willistoni* lacks *atlas* but appears to have several additional paralogs of the protamines found only in duplicate in *D. melanogaster*.

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, this movement remains noteworthy. Prior work has found a significant dearth of testis-expressed genes on the X chromosome in *Drosophila* [69,91–93] and other species [94,95]. 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 [96–98]. 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 [99–104]. 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 [102–108], but continued research in this area might inform further interrogation of the forces driving *atlas* off of the X chromosome.

The movement of the *atlas* protein-coding sequence to chromosome 2 also created the two-exon gene observed in *D. melanogaster*, in which the longer second exon appears to be entirely

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

A final issue raised by our results is the exact timing and mechanism of origin for the *atlas* protein-coding sequence. The bioinformatic screen [19] that identified *atlas* and the other genes tested in Fig 1 was designed to identify both “*de novo*” genes, defined as protein-coding regions in *Drosophila* that had recognizable, but non-ORF-maintaining, TBLASTN hits in outgroup species, and “putative *de novo*” genes, which had no TBLASTN hits in outgroup species. (Importantly, the screen also eliminated any protein with an identifiable protein domain, thus reducing the chances of identifying divergent members of gene families.) The vast majority of the genes we tested with RNAi, including *atlas*, fell into the putative *de novo* category. The bioinformatic screen’s criteria were reasonable for a high-throughput analysis, but BLAST-based methods have known limitations for detecting orthologous sequences in diverged species [74,113,114]. The lack of identifiable *atlas* protein-coding genes in several *Drosophila* species (e.g., *D. pseudoobscura* and *D. willistoni*) is unlikely to be due to BLAST homology detection failure, and extensive synteny-based searches confirmed the gene’s absence (S9 Fig). BLAST and synteny-based searches for orthologs in non-*Drosophila* species also did not detect an ortholog, though BLAST searches are not predicted to have adequate sensitivity for a protein of this size and evolutionary rate, at this level of species divergence [74]. Hence, in addition to using synteny to search for orthologs, we used HMMER, which employs hidden Markov models and builds a sequence profile of the target protein using information from multiple orthologs. Since HMMER also did not detect orthologs outside of *Drosophila*, we hypothesize that *atlas* evolved *de novo* at the base of the genus. However, since we remain unable to identify the non-protein-coding sequence from which *atlas* arose, we continue to refer to *atlas* as a putative *de novo* gene [5].

Overall, we find that while many putative *de novo* evolved genes are expressed in the *D. melanogaster* testes, few have major, non-redundant effects on fertility. However, several such genes have evolved critical roles at distinct stages of spermatogenesis and sperm function. We showed previously that the putative *de novo* gene *saturn* is required for maximal sperm production, as well as for the ability of transferred sperm to migrate successfully to sperm storage organs in females [46]. Another putative *de novo* gene, *goddard*, is required for sperm production and encodes a cytoplasmic protein that appears to localize to elongating axonemes [20,46]. Loss of *goddard* impairs the individualization of spermatid bundles [20], thus exerting an effect that appears to be upstream of those observed for *saturn* and *atlas*. Here, we report another novel function for a putative *de novo* gene: encoding an essential transition protein that is necessary for proper nuclear condensation in spermiogenesis. Taken together, these results demonstrate that while many *de novo* genes may play subtle roles or share functional redundancy with other genes, *de novo* genes can also become essential players in complex cellular processes that mediate successful reproduction.

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 [19]. We filtered these genes with publicly available RNA-seq data [115] to identify those expressed predominantly in the testes (>50% of RPKM sum deriving from the testes from ModENCODE data) [115], 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 [46,116]. Control flies were generated by crossing the attP-containing genetic background into which UAS-RNAi was inserted 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 [117] or the Transgenic RNAi Project [50]. For the other genes, no publicly available RNAi stock was available, so we constructed TRiP-style stocks in the pValium20 vector as previously described [85]. These constructs were integrated into an AttP site in stock BL 25709 ($y^1 v^1 P\{\text{nos-}\phi\text{C31}\int\text{int.NLS}\}X; P\{\text{CaryP}\}\text{attP40}$) 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. S3 Table shows all RNAi lines used and lists the short hairpin sequences cloned for the TRiP lines we constructed.

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 [46]. 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 [46]. We evaluated knockdown efficiency by agarose gel electrophoresis of RT-PCR products as one of four levels: “complete” if no product from the knockdown cDNA sample was visible via agarose gel electrophoresis; “near complete” for a very faint knockdown product that was also much less abundant than the control product; “partial” for a more robust knockdown product that was still visibly less intense than control; and “not knocked down” if the product intensity for the knockdown sample equaled or exceeded that of the control. Any gene that did not show at least partial knockdown was discarded from further analysis, leaving a total of 42 genes successfully screened. S3 Table shows the degree of knockdown achieved for each line.

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 [20]. 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 [118]. 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 expected *Bfa*I site was lost indicated a mutation, which was balanced and then confirmed by PCR and sequencing of homozygous mutant lines.

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

Vasa-Cas9 embryos were co-injected with the assembled plasmid and a pU6.3 plasmid encoding a gRNA targeting the region just downstream of the *atlas* stop codon. G0 flies were crossed to *w*¹¹¹⁸ adults, and G1 flies were screened for red fluorescent eyes using the NIGHTSEA system (NIGHTSEA LLC, Lexington, MA). Six balanced lines from two independent G1 broods were established. To remove the dsRed from the *atlas* locus, we crossed these lines to a PiggyBac transposase line (BDSC #8285) and then selected against pBac and dsRed in the following generation. PCR and sequencing confirmed the expected “scarless” insertion of GFP at the *atlas* locus.

***atlas* genomic rescue line**

We constructed an HA-tagged *atlas* rescue line that contained the *atlas* gene flanked by 1345 bp of sequence upstream of the start codon (but excluding the coding sequence of upstream gene *CG3124*) and 3000 bp of sequence downstream of the stop codon (including the full 3' UTR) as follows. Genomic sequences were PCR amplified using Q5 High fidelity Polymerase (NEB), purified Canton S genomic DNA (Gentra Puregene Tissue Kit, Qiagen, Germantown, MD), and the *atlas* rescue F1/R1 and *atlas* rescue F3/R3 primer sets (see S4 Table). The 3x-HA tag was likewise amplified from pTWH (DGRC 1100; T. Murphy) using *atlas* rescue F2/R2 primers. These DNA fragments were subsequently assembled into a *Xba*I/*Asc*I-linearized w

+attB plasmid (Addgene, Watertown, MA, plasmid 30326, deposited by J. Sekelsky). The assembled construct was then phiC31 integrated into the PBac{y⁺-attP-3B}VK00037 (Bloomington *Drosophila* Stock Center (BDSC) stock #24872) docking site (Rainbow Transgenics) and crossed into the *atlas* null background to assess rescue.

Fertility assays and sperm visualization

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 [20,46], 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 25° 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 [46].

To assess the level of fertility conferred by the *atlas*-GFP allele, we crossed *atlas*-GFP and w^{1118} 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 [51]. Samples were prepared, imaged and analyzed as described previously [46].

atlas-GFP ectopic expression

We used the Gateway cloning system (Thermo) to construct an *atlas*-GFP transgene expressed under UAS control (primers in S4 Table). 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 [20].

Imaging spermatogenesis and spermatid nuclear condensation

We used phase-contrast microscopy to examine the stages of spermatogenesis in whole mount testes [120]. 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 [20]. The final stages of nuclear condensation were visualized with the Mst35Bb-GFP marker

described above, as well as an equivalent marker, Mst35Bb-dsRed [51]. 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 [53,121]. Images with H2AvD-RFP were obtained with epifluorescence microscopy, since we lack an appropriate confocal laser for RFP.

To examine spermatid nuclei at various stages of condensation, we visualized nuclear bundles using TOPRO-3. Testes of newly eclosed (<1 day old) *atlas* null and control males were dissected in PBS. Testes were then transferred to a droplet of 2% paraformaldehyde on poly-L-lysine treated glass slides and were gently shredded in the post-meiotic region to release sperm bundles. Testes were gently squashed beneath coverslips coated in Sigmacote (Sigma Aldrich, St. Louis, MO). We then froze slides in liquid nitrogen for a few seconds and popped off of the siliconized coverslip with a razor. Slides were incubated in Coplin jars filled with 95% ethanol at -20°C for 30 minutes and then mounted in VECTASHIELD (Vector Laboratories, Burlingame, CA). Nuclear staging was performed by examining the shape of the nuclei. Early and late canoe stages of condensation were distinguished by the absence or presence of Mst35Bb-GFP, respectively. Elongated and late canoe stages were distinguished by the presence or absence, respectively, of vesicles of GFP-tagged nuclear proteins (Atlas-GFP or Mst35Bb-GFP) located basal to the nuclei. Examples of stages are given in Fig 4. Confocal stacks were taken on a Leica SP5 microscope, images were captured by LASAF, and ImageJ was used to flatten stacks into a single, two-dimensional image. All intact nuclear bundles were counted for each dissection.

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

Evolutionary and gene expression analysis of *atlas*

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

To test for sex-specific expression biases for both the ORF and the 3' UTR sequences, we used adult male- and female-specific RNA-seq data from numerous *Drosophila* species accessed through the Genomics Education Partnership version of the UCSC Genome Browser (<http://gander.wustl.edu/>) and initially collected by Brown et al. [115] and Chen et al. [123]. We also confirmed these findings experimentally in several species by performing RT-PCR on cDNA isolated from whole males and whole females, as previously described [46].

To search for *atlas* orthologs in non-*Drosophila* Dipterans, we obtained from ENSEMBL Metazoa the genomes of *Musca domestica*, *Glossina morsitans*, *Lucilia cuprina*, *Aedes aegypti*, *Anopheles darlingi*, *Anopheles gambiae*, *Culex quinquefasciatus* and *Mayetiola destructor*. We performed a synteny search (summarized in S11 Fig) in each species by identifying the nearest neighbors of *atlas* in the *D. ananassae* and *D. virilis* genomes that had an identifiable homolog

in each species. In all cases, the homologs of the nearest neighbors on each side of *atlas* were found on different contigs, suggesting synteny breakdown. We obtained up to 1 Mb of sequence on each side of each identified homolog and queried it with BLASTN, TBLASTN, and Exonerate [124] for regions with significant 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 [78], 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*, *ficuspshila*, *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 reported results come from an analysis that excluded *D. takahashii*, which produced a more reliable alignment and showed no evidence for any sites under positive selection.

Supporting information

S1 Fig. Example RT-PCR demonstrating near-complete knockdown of *atlas*. Knockdown was driven by crossing strain VDRC KK-108680 to *Bam*-GAL4, UAS-*Dicer2*. We also attempted to induce knockdown in the same manner with strain VDRC GD-17240, and we produced control flies by crossing VDRC attP strain #60100 to *Bam*-GAL4, UAS-*Dicer2*. cDNA was isolated from whole males of each strain, and a standardized amount of cDNA or control genomic DNA from w^{1118} was assessed for amplification of *atlas* and a housekeeping control gene, *RpL32*. The GD line did not induce detectable knockdown, but the KK line showed near-complete knockdown of *atlas*. Knockdown was assessed in the same way for all other RNAi lines tested; the control cross for TRIP-style RNAi lines was $y v 1509$ crossed to *Bam*-GAL4, UAS-*Dicer2*.
(PDF)

S2 Fig. CRISPR editing methods for constructing *atlas* loss-of-function mutations. A) Graphical representation of CRISPR loss-of-function strategies. Purple boxes represent the *atlas* coding region, while gray represents non-coding regions of the gene. Scissors indicate locations where gRNAs targeted Cas9-mediated double-stranded breaks. The deletion allele was generated using two sgRNAs targeting either side of *atlas* to excise the complete coding region (CDS) and nearly all of the noncoding region. Frameshift alleles were created using one sgRNA targeting a cut at the start of the *atlas* coding region that was repaired with non-homologous end joining (NHEJ). B) Alignment of mutations generated in frameshift alleles. The gene's start codon is indicated with dark green shading. Blue shading indicates bases inserted

by NHEJ, orange shading indicates NHEJ deletions. All three mutations consist of net insertions or deletions that are non-multiples of three, resulting in premature stop codons indicated with underlining. The mutant alleles retain the possibility of encoding a truncated form of Atlas protein if a downstream start codon (light green shading) is used, since this codon is in-frame with the sequence encoding the protein's C terminus. C) Predicted protein sequences encoded by the wild-type and frameshift alleles. Blue shading indicates novel amino acids created by NHEJ indel mutations. Gray shading indicates the potential truncated Atlas protein that could be encoded by the frameshift alleles if translation initiated at the downstream start codon indicated in panel B. Such hypothetical, N-terminally truncated forms of Atlas protein would contain amino acids 61–172 of the wild-type protein. Yellow shading indicates the putative nuclear localization signal.

(PDF)

S3 Fig. Atlas-HA genomic rescue construct restores full fertility in the *atlas* null background.

A single-pair fertility assay was used to assess fertility in males carrying a single copy of the *atlas*-HA genomic rescue construct in the *atlas* null background. These males showed no significant difference in fertility from control (w^{1118}) males (rescue fertility: 92.6 ± 6.6 ; control fertility: 74.8 ± 10.3 ; two-sample t-test assuming unequal variances, $p = 0.16$). Both control and rescue males had significantly higher fertility than *atlas* null males (null fertility: 2.7 ± 2.0 ; two-sample t-tests, both $p < 10^{-5}$).

(PDF)

S4 Fig. Accumulation of sperm in *atlas* null and knockdown testes.

A) Phase contrast imaging of *atlas* null and control males ages 1 day or 4 days. Sperm accumulate in the SV by 4 days in controls, but accumulate in the basal testes of null males on days 1 and 4. The day 7 images are shown in Fig 3 of the main text. B) The same phenotype of sperm accumulation in the basal testis is observed in 7-day-old knockdown males. C) Knockdown males expressing Mst35Bb-GFP show similar patterns of spermatid nuclei to nulls, while control males accumulate many sperm nuclei in the SV (compare to Fig 3 in the main text). Control flies in B-C were generated by crossing VDRC strain #60100 (attP) to Mst35Bb-GFP; +; *Bam*-GAL4, UAS-*Dicer2*. SVs are highlighted for clarity when needed with dotted lines.

(PDF)

S5 Fig. Expression of *atlas*-GFP driven in larval salivary glands and early male germline cells.

A) Dissected larval salivary glands expressing UAS-*atlas*-GFP under the control of *tubulin*-GAL4. B) Apical portion of a testis expressing UAS-*atlas*-GFP under the control of *Bam*-GAL4. In both cases, Atlas-GFP has a predominantly nuclear localization pattern.

(PDF)

S6 Fig. Scarless CRISPR/Cas9 genome editing strategy to produce a GFP knock-in allele at the endogenous *atlas* locus.

Golden Gate assembly was used to construct a plasmid carrying left and right homology arms flanking GFP placed in frame with the end of the *atlas* protein-coding sequence and a dsRed marker under the control of the 3xP3 promoter, which drives expression in the eye. This plasmid was injected into *vasa*-Cas9 flies along with a pU6.3 plasmid containing a gRNA targeting the end of the *atlas* coding sequence. G_0 flies were crossed to w^{1118} , and dsRed positive flies were screened molecularly for the correct *atlas*-GFP insert at the endogenous locus. The dsRed construct was then excised by crossing to a pBac transposase line, which removed the dsRed using flanking TTAA sequences. This protocol was adapted from Hill et al. (2019) and described at <https://flycrispr.org/>.

(PDF)

S7 Fig. Images of whole testes from which the zoomed in basal portions in Fig 5 are taken. A) Whole testis dissection corresponding to Fig 5C. Small GFP-positive puncta are visible near the progressed actin cones, which may represent the removal of Atlas-GFP from condensed nuclei. Also see Fig 6D. B) Whole testis dissection corresponding to Fig 5D.
(PDF)

S8 Fig. The region of the genome corresponding to the non-coding exon of *atlas* in *D. melanogaster* is detectable across *Drosophila*, but expressed in only some species. The non-coding second exon of *atlas* from *D. melanogaster* was compared with BLASTN to 11 other *Drosophila* species. The blue dashed line indicates the region that showed significant sequence identity in these searches. The Adult Male RNA-Seq track shows evidence of male-expressed RNA in the region. Peak heights are not comparable across species because the RNA-Seq was performed at different times. Thus, this analysis gives qualitative information about whether the conserved 3' UTR of *atlas* in *D. melanogaster* is expressed in males in other species. The + or - symbol below each species name indicates whether the top (+) or bottom (-) strand DNA sequence matches the sequence of expressed, 3' UTR mRNA in *D. melanogaster*. Expression of this region was further assessed for some species by RT-PCR; see S9 Fig.
(PDF)

S9 Fig. Synteny analysis across *Drosophila* species of regions of Muller element A that contain *atlas* in *D. ananassae* and *D. virilis*. A) *Atlas* is found downstream of the *Met* gene ortholog in both *D. ananassae* and *D. virilis*. This region was therefore searched in multiple additional species. While the genomes of several species harbored unannotated genes in this general region that showed RNA-seq evidence of male expression, all such predicted genes encoded proteins that were significant BLASTP hits to *D. melanogaster* proteins other than Atlas. B) In *D. ananassae*, *atlas* is found in the middle of an intron of the *Ptp10D* gene. However, in some species, *Ptp10D* is no longer syntenic with *Met*. Therefore, we searched for *atlas* orthologs in and around *Ptp10D* across the same set of species. While orthologs of two other *Drosophila* gene have become inserted into a *Ptp10D* intron in other lineages, no additional *atlas* orthologs were found. In both panels, asterisks indicate unannotated genes supported by RNA-seq evidence that were confirmed with BLASTP to be homologs of genes other than *atlas*.
(PDF)

S10 Fig. RT-PCR of the *atlas* coding sequence (CDS) and non-coding exon (3'UTR) in other *Drosophila* species. cDNA was prepared from whole males or whole females and analyzed with either *atlas* coding sequence primers, primers designed to a portion of the non-coding exon, or housekeeping gene *RpL32* as a control. The protein-coding region is expressed in a male-specific manner in *D. ananassae* and *D. virilis*, consistent with available RNA-seq data. The non-coding region shows robust male-specific expression in *D. pseudoobscura*, but was not detectable in *D. mojavensis* (one primer pair attempted), *D. ananassae* (two primer pairs attempted) or *D. willistoni* (two primer pairs attempted). One of two primer pairs attempted gave faint, non-sex-specific amplification in *D. virilis*.
(PDF)

S11 Fig. Synteny-based method for searching other Dipteran genomes for potential *atlas* orthologs.
(PDF)

S1 Table. Stages of nuclear condensation observed in spermiogenesis in wild-type and *atlas* null testes. Each line shows the distribution of staged nuclear bundles dissected from

one individual pair of testes. Examples of nuclear stages and the curled nuclear phenotype observed in *atlas* null males are shown in Fig 4.

(PDF)

S2 Table. Predicted biochemical properties and levels of conservation for Atlas orthologs identified across *Drosophila* species.

(PDF)

S3 Table. Sources of RNAi lines, degrees of knockdown observed, and hairpins cloned to generate TRiP-style RNAi lines.

(XLSX)

S4 Table. Primers used for constructing *atlas*-HA genomic rescue construct and *atlas*-GFP donor template construct.

(PDF)

Acknowledgments

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

Author Contributions

Conceptualization: Emily L. Rivard, Prajal H. Patel, Erich Bornberg-Bauer, Geoffrey D. Findlay.

Data curation: Prajal H. Patel, Anna Grandchamp, Emilie M. Scott, Geoffrey D. Findlay.

Formal analysis: Emily L. Rivard, Andrew G. Ludwig, Prajal H. Patel, Anna Grandchamp, Sarah E. Arnold, Alina Berger, Emilie M. Scott, Brendan J. Kelly, Grace C. Mascha, Geoffrey D. Findlay.

Funding acquisition: Anna Grandchamp, Erich Bornberg-Bauer, Geoffrey D. Findlay.

Investigation: Emily L. Rivard, Andrew G. Ludwig, Prajal H. Patel, Anna Grandchamp, Sarah E. Arnold, Alina Berger, Emilie M. Scott, Brendan J. Kelly, Grace C. Mascha, Geoffrey D. Findlay.

Methodology: Emily L. Rivard, Prajal H. Patel.

Project administration: Prajal H. Patel, Erich Bornberg-Bauer, Geoffrey D. Findlay.

Resources: Prajal H. Patel, Erich Bornberg-Bauer, Geoffrey D. Findlay.

Software: Anna Grandchamp, Alina Berger.

Supervision: Prajal H. Patel, Anna Grandchamp, Erich Bornberg-Bauer, Geoffrey D. Findlay.

Validation: Emily L. Rivard, Andrew G. Ludwig, Prajal H. Patel, Anna Grandchamp, Sarah E. Arnold.

Visualization: Emily L. Rivard, Andrew G. Ludwig, Prajal H. Patel, Geoffrey D. Findlay.

Writing – original draft: Emily L. Rivard, Prajal H. Patel, Geoffrey D. Findlay.

Writing – review & editing: Emily L. Rivard, Andrew G. Ludwig, Prajal H. Patel, Anna Grandchamp, Sarah E. Arnold, Emilie M. Scott, Brendan J. Kelly, Grace C. Mascha, Erich Bornberg-Bauer, Geoffrey D. Findlay.

References

- Zhang J. Evolution by gene duplication: an update. *Trends Ecol Evol.* 2003; 18(6):292–8. [https://doi.org/10.1016/S0169-5347\(03\)00033-8](https://doi.org/10.1016/S0169-5347(03)00033-8).
- Lipinski KJ, Farslow JC, Fitzpatrick KA, Lynch M, Katju V, Bergthorsson U. High spontaneous rate of gene duplication in *Caenorhabditis elegans*. *Curr Biol.* 2011; 21(4):306–10. <https://doi.org/10.1016/j.cub.2011.01.026> PMID: 21295484
- Van Oss SB, Carvunis A-R. *De novo* gene birth. *PLoS Genet.* 2019; 15(5):e1008160. <https://doi.org/10.1371/journal.pgen.1008160> PMID: 31120894
- Bornberg-Bauer E, Schmitz JF. Fact or fiction: Updates on how protein-coding genes might emerge *de novo* from previously non-coding DNA. *F1000Res.* 2017; 6:57. <https://doi.org/10.12688/f1000research.10079.1> PMID: 28163910
- McLysaght A, Hurst LD. Open questions in the study of *de novo* genes: what, how and why. *Nat Rev Genet.* 2016; 17:567. <https://doi.org/10.1038/nrg.2016.78> PMID: 27452112
- Li Z-W, Chen X, Wu Q, Hagmann J, Han T-S, Zou Y-P, et al. On the origin of *de novo* genes in *Arabidopsis thaliana* populations. *Genome Biol Evol.* 2016; 8(7):2190–202. <https://doi.org/10.1093/gbe/evw164> PMID: 27401176
- Guerzoni D, McLysaght A. *De novo* genes arise at a slow but steady rate along the primate lineage and have been subject to incomplete lineage sorting. *Genome Biol Evol.* 2016; 8(4):1222–32. <https://doi.org/10.1093/gbe/evw074> PMID: 27056411
- Lu T-C, Leu J-Y, Lin W-C. A comprehensive analysis of transcript-supported *de novo* genes in *Saccharomyces sensu stricto* yeasts. *Mol Biol Evol.* 2017; 34(11):2823–38. <https://doi.org/10.1093/molbev/msx210> PMID: 28981695
- Zhang L, Ren Y, Yang T, Li G, Chen J, Gschwend AR, et al. Rapid evolution of protein diversity by *de novo* origination in *Oryza*. *Nat Ecol Evol.* 2019; 3(4):679–90. <https://doi.org/10.1038/s41559-019-0822-5> PMID: 30858588
- Zhao L, Saelao P, Jones CD, Begun DJ. Origin and spread of *de novo* genes in *Drosophila melanogaster* populations. *Science.* 2014; 343(6172):769–72. <https://doi.org/10.1126/science.1248286> PMID: 24457212
- Puntambekar S, Newhouse R, Navas JSM, Chauhan R, Vernaz G, Willis T, et al. Evolutionary divergence of novel open reading frames in cichlids speciation. *Sci Rep.* 2020; 10(1). <https://doi.org/10.1038/s41598-020-78555-0> PMID: 33299045
- Ruiz-Orera J, Hernandez-Rodriguez J, Chiva C, Sabidó E, Kondova I, Bontrop R, et al. Origins of *de novo* genes in human and chimpanzee. *PLoS Genet.* 2016; 11(12):e1005721.
- Chamakura KR, Tran JS, O'Leary C, Lisciandro HG, Antillon SF, Garza KD, et al. Rapid *de novo* evolution of lysis genes in single-stranded RNA phages. *Nat Commun.* 2020; 11(1). <https://doi.org/10.1038/s41467-020-19860-0> PMID: 33243984
- Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N, et al. Proto-genes and *de novo* gene birth. *Nature.* 2012; 487:370–4. <https://doi.org/10.1038/nature11184> PMID: 22722833
- Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as predicted by the preadaptation hypothesis of *de novo* gene birth. *Nat Ecol Evol.* 2017; 1(6):0146. <https://doi.org/10.1038/s41559-017-0146> PMID: 28642936
- Schlötterer C. Genes from scratch—the evolutionary fate of *de novo* genes. *Trends Genet.* 2015; 31(4):215–9. <https://doi.org/10.1016/j.tig.2015.02.007> PMID: 25773713
- Schmitz JF, Ullrich KK, Bornberg-Bauer E. Incipient *de novo* genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol.* 2018; 2(10):1626–32. <https://doi.org/10.1038/s41559-018-0639-7> PMID: 30201962
- Vakirlis N, Hebert AS, Opulente DA, Achaz G, Hittinger CT, Fischer G, et al. A molecular portrait of *de novo* genes in yeasts. *Mol Biol Evol.* 2018; 35(3):631–45. <https://doi.org/10.1093/molbev/msx315> PMID: 29220506

19. Heames B, Schmitz J, Bornberg-Bauer E. A continuum of evolving *de novo* genes drives protein-coding novelty in *Drosophila*. *J Mol Evol*. 2020; 88(4):382–98. <https://doi.org/10.1007/s00239-020-09939-z> PMID: 32253450
20. Lange A, Patel PH, Heames B, Damry AM, Saenger T, Jackson CJ, et al. Structural and functional characterization of a putative *de novo* gene in *Drosophila*. *Nat Commun*. 2021; 12:1667. <https://doi.org/10.1038/s41467-021-21667-6> PMID: 33712569
21. Keeling DM, Garza P, Nartey CM, Carvunis A-R. The meanings of 'function' in biology and the problematic case of *de novo* gene emergence. *eLife*. 2019; 8:e47014. <https://doi.org/10.7554/eLife.47014> PMID: 31674305
22. Cai J, Zhao R, Jiang H, Wang W. *De novo* origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics*. 2008; 179(1):487–96. <https://doi.org/10.1534/genetics.107.084491> PMID: 18493065
23. Li D, Yan Z, Lu L, Jiang H, Wang W. Pleiotropy of the *de novo*-originated gene *MDF1*. *Sci Rep*. 2014; 4(1):7280. <https://doi.org/10.1038/srep07280> PMID: 25452167
24. Xiao W, Liu H, Li Y, Li X, Xu C, Long M, et al. A rice gene of *de novo* origin negatively regulates pathogen-induced defense response. *PLoS One*. 2009; 4(2):e4603. <https://doi.org/10.1371/journal.pone.0004603> PMID: 19240804
25. Baalsrud HT, Tørresen OK, Solbakken MH, Salzburger W, Hanel R, Jakobsen KS, et al. *De novo* gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol Biol Evol*. 2018; 35(3):593–606. <https://doi.org/10.1093/molbev/msx311> PMID: 29216381
26. Zhuang X, Yang C, Murphy KR, Cheng CHC. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc Natl Acad Sci U S A*. 2019; 116(10):4400–5. <https://doi.org/10.1073/pnas.1817138116> PMID: 30765531
27. Xie C, Bekpen C, Künzel S, Keshavarz M, Krebs-Wheaton R, Skrabar N, et al. A *de novo* evolved gene in the house mouse regulates female pregnancy cycles. *eLife*. 2019;8. <https://doi.org/10.7554/eLife.44392> PMID: 31436535
28. Suenaga Y, Islam SMR, Alagu J, Kaneko Y, Kato M, Tanaka Y, et al. NCYM, a *cis*-antisense gene of MYCN, encodes a *de novo* evolved protein that inhibits GSK3 β resulting in the stabilization of MYCN in human neuroblastomas. *PLoS Genet*. 2014; 10(1):e1003996. <https://doi.org/10.1371/journal.pgen.1003996> PMID: 24391509
29. Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. Emergence of young human genes after a burst of retroposition in primates. *PLoS Biol*. 2005; 3(11):e357. <https://doi.org/10.1371/journal.pbio.0030357> PMID: 16201836
30. Baker RH, Narechania A, Johns PM, Wilkinson GS. Gene duplication, tissue-specific gene expression and sexual conflict in stalk-eyed flies (Diopsidae). *Philos Trans R Soc Lond B Biol Sci*. 2012; 367(1600):2357–75. <https://doi.org/10.1098/rstb.2011.0287> PMID: 22777023
31. Cui X, Lv Y, Chen M, Nikoloski Z, Twell D, Zhang D. Young genes out of the male: an insight from evolutionary age analysis of the pollen transcriptome. *Mol Plant*. 2015; 8(6):935–45. <https://doi.org/10.1016/j.molp.2014.12.008>; 20. PMID: 25670339
32. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. *Genome Res*. 2010; 20(10):1313–26. <https://doi.org/10.1101/gr.101386.109> PMID: 20651121
33. Levine MT, Jones CD, Kern AD, Lindfors HA, Begun DJ. Novel genes derived from noncoding DNA in *Drosophila melanogaster* are frequently X-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*. 2006; 103(26):9935–9. <https://doi.org/10.1073/pnas.0509809103> PMID: 16777968
34. Begun DJ, Lindfors HA, Kern AD, Jones CD. Evidence for *de Novo* Evolution of Testis-Expressed Genes in the *Drosophila yakuba/Drosophila erecta* Clade. *Genetics*. 2007; 176(2):1131–7. <https://doi.org/10.1534/genetics.106.069245> PMID: 17435230
35. Necsulea A, Kaessmann H. Evolutionary dynamics of coding and non-coding transcriptomes. *Nat Rev Genet*. 2014; 15(11):734–48. <https://doi.org/10.1038/nrg3802> PMID: 25297727
36. Soumillon M, Necsulea A, Weier M, Brawand D, Zhang X, Gu H, et al. Cellular source and mechanisms of high transcriptome complexity in the mammalian testis. *Cell Rep*. 2013; 3(6):2179–90. <https://doi.org/10.1016/j.celrep.2013.05.031> PMID: 23791531
37. Sorourian M, Kunte MM, Domingues S, Gallach M, Özdil F, Río J, et al. Relocation facilitates the acquisition of short *cis*-regulatory regions that drive the expression of retrogenes during spermatogenesis in *Drosophila*. *Mol Biol Evol*. 2014; 31(8):2170–80. <https://doi.org/10.1093/molbev/msu168> PMID: 24855141
38. Palmieri N, Kosiol C, Schlötterer C. The life cycle of *Drosophila* orphan genes. *eLife*. 2014; 3:e01311. <https://doi.org/10.7554/eLife.01311> PMID: 24554240

39. Wilburn DB, Swanson WJ. From molecules to mating: Rapid evolution and biochemical studies of reproductive proteins. *J Proteomics*. 2016; 135:12–25. <https://doi.org/10.1016/j.jprot.2015.06.007> PMID: 26074353
40. VanKuren NW, Long M. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. *Nat Ecol Evol*. 2018; 2(4):705–12. <https://doi.org/10.1038/s41559-018-0471-0> PMID: 29459709
41. Hales KG, Korey CA, Larracuente AM, Roberts DM. Genetics on the fly: a primer on the *Drosophila* model system. *Genetics*. 2015; 201(3):815–42. <https://doi.org/10.1534/genetics.115.183392> PMID: 26564900
42. Demarco RS, Eikenes TH, Haglund K, Jones DL. Investigating spermatogenesis in *Drosophila melanogaster*. *Methods*. 2014; 68(1):218–27. <https://doi.org/10.1016/j.ymeth.2014.04.020> PMID: 24798812
43. Witt E, Benjamin S, Svetec N, Zhao L. Testis single-cell RNA-seq reveals the dynamics of *de novo* gene transcription and germline mutational bias in *Drosophila*. *eLife*. 2019; 8. <https://doi.org/10.7554/eLife.47138> PMID: 31418408
44. Reinhardt JA, Wanjiru BM, Brant AT, Saelao P, Begun DJ, Jones CD. *De novo* ORFs in *Drosophila* are important to organismal fitness and evolved rapidly from previously non-coding sequences. *PLoS Genet*. 2013; 9(10):e1003860. <https://doi.org/10.1371/journal.pgen.1003860> PMID: 24146629
45. Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan L, Pang N, et al. New genes often acquire male-specific functions but rarely become essential in *Drosophila*. *Genes Dev*. 2017; 31(18):1841–6. <https://doi.org/10.1101/gad.303131.117> PMID: 29051389
46. Gubala AM, Schmitz JF, Kearns MJ, Vinh TT, Bornberg-Bauer E, Wolfner MF, et al. The *goddard* and *saturn* genes are essential for *Drosophila* male fertility and may have arisen *de novo*. *Mol Biol Evol*. 2017; 34(5):1066–82. <https://doi.org/10.1093/molbev/msx057> PMID: 28104747
47. Abrusán G. Integration of new genes into cellular networks, and their structural maturation. *Genetics*. 2013; 195(4):1407–17. <https://doi.org/10.1534/genetics.113.152256> PMID: 24056411
48. Zhang W, Landback P, Gschwend AR, Shen B, Long M. New genes drive the evolution of gene interaction networks in the human and mouse genomes. *Genome Biol*. 2015; 16(1). <https://doi.org/10.1186/s13059-015-0726-x> PMID: 27391693
49. Wu DD, Irwin DM, Zhang YP. *De novo* origin of human protein-coding genes. *PLoS Genet*. 2011; 7(11). <https://doi.org/10.1371/journal.pgen.1002379> PMID: 22102831
50. Ni J-Q, Zhou R, Czech B, Liu L-P, Holderbaum L, Yang-Zhou D, et al. A genome-scale shRNA resource for transgenic RNAi in *Drosophila*. *Nat Methods*. 2011; 8(5):405–7. <https://doi.org/10.1038/nmeth.1592> PMID: 21460824
51. Manier MK, Belote JM, Berben KS, Novikov D, Stuart WT, Pitnick S. Resolving mechanisms of competitive fertilization success in *Drosophila melanogaster*. *Science*. 2010; 328(5976):354–7. <https://doi.org/10.1126/science.1187096> PMID: 20299550
52. Steinhauer J. Separating from the pack: molecular mechanisms of *Drosophila* spermatid individualization. *Spermatogenesis*. 2015; 5(2):e1041345. <https://doi.org/10.1080/21565562.2015.1041345> PMID: 26413413
53. Rathke C, Baarends WM, Jayaramaiah-Raja S, Bartkuhn M, Renkawitz R, Renkawitz-Pohl R. Transition from a nucleosome-based to a protamine-based chromatin configuration during spermiogenesis in *Drosophila*. *J Cell Sci*. 2007; 120(9):1689–700. <https://doi.org/10.1242/jcs.004663> PMID: 17452629
54. Fabian L, Brill JA. *Drosophila* spermiogenesis. *Spermatogenesis*. 2012; 2(3):197–212. <https://doi.org/10.4161/spmg.21798> PMID: 23087837
55. Rathke C, Baarends WM, Awe S, Renkawitz-Pohl R. Chromatin dynamics during spermiogenesis. *Biochim Biophys Acta*. 2014; 1839(3):155–68. <https://doi.org/10.1016/j.bbagr.2013.08.004> PMID: 24091090
56. Kanippayoor RL, Alpern JHM, Moehring AJ. Protamines and spermatogenesis in *Drosophila* and *Homo sapiens*. *Spermatogenesis*. 2013; 3(2):e24376. <https://doi.org/10.4161/spmg.24376> PMID: 23885304
57. Jayaramaiah Raja S, Renkawitz-Pohl R. 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. *Mol Cell Biol*. 2005; 25(14):6165–77. <https://doi.org/10.1128/MCB.25.14.6165-6177.2005> PMID: 15988027
58. Gärtner SMK, Rothenbusch S, Buxa MK, Theofel I, Renkawitz R, Rathke C, et al. The HMG-box-containing proteins tHMG-1 and tHMG-2 interact during the histone-to-protamine transition in *Drosophila* spermatogenesis. *Eur J Cell Biol*. 2015; 94(1):46–59. <https://doi.org/10.1016/j.ejcb.2014.10.005> PMID: 25464903

59. Gärtner SMK, Hundertmark T, Nolte H, Theofel I, Eren-Ghiani Z, Tetzner C, et al. Stage-specific testes proteomics of *Drosophila melanogaster* identifies essential proteins for male fertility. *Eur J Cell Biol*. 2019; 98(2):103–15. <https://doi.org/10.1016/j.ejcb.2019.01.001> PMID: 30679029
60. Alvi ZA, Chu T-C, Schawaroch V, Klaus AV. Genomic and expression analysis of transition proteins in *Drosophila*. *Spermatogenesis*. 2016; 5(3):e1178518. <https://doi.org/10.1080/21565562.2016.1178518> PMID: 27512614
61. Alvi ZA, Chu T-C, Schawaroch V, Klaus AV. Protamine-like proteins in 12 sequenced species of *Drosophila*. *Protein Pept Lett*. 2013; 20(1):17–35. <https://doi.org/10.2174/092986613804096847> PMID: 22789106
62. Dorus S, Freeman ZN, Parker ER, Heath BD, Karr TL. Recent Origins of Sperm Genes in *Drosophila*. *Mol Biol Evol*. 2008; 25(10):2157–66. <https://doi.org/10.1093/molbev/msn162> PMID: 18653731
63. Eren-Ghiani Z, Rathke C, Theofel I, Renkawitz-Pohl R. Prt199C Acts Together with Protamines and Safeguards Male Fertility in *Drosophila*. *Cell Rep*. 2015; 13(11):2327–35. <https://doi.org/10.1016/j.celrep.2015.11.023> PMID: 26673329
64. Doyen CM, Chalkley GE, Voets O, Bezstarosti K, Demmers JA, Moshkin YM, et al. A Testis-Specific Chaperone and the Chromatin Remodeler ISWI Mediate Repackaging of the Paternal Genome. *Cell Rep*. 2015; 13(7):1310–8. <https://doi.org/10.1016/j.celrep.2015.10.010> PMID: 26549447
65. Lange A, Mills RE, Lange CJ, Stewart M, Devine SE, Corbett AH. Classical nuclear localization signals: definition, function, and interaction with Importin α . *J Biol Chem*. 2007; 282(8):5101–5. <https://doi.org/10.1074/jbc.R600026200> PMID: 17170104
66. Hill A, Jain P, Ben-Shahar Y. The *Drosophila* ERG channel seizure plays a role in the neuronal homeostatic stress response. *PLoS Genet*. 2019; 15(8):e1008288. <https://doi.org/10.1371/journal.pgen.1008288> PMID: 31393878
67. Bier E, Harrison MM, O'Connor-Giles KM, Wildonger J. Advances in Engineering the Fly Genome with the CRISPR-Cas System. *Genetics*. 2018; 208(1):1–18. <https://doi.org/10.1534/genetics.117.1113> PMID: 29301946
68. Bruckner JJ, Zhan H, Gratz SJ, Rao M, Ukken F, Zilberg G, et al. Fife organizes synaptic vesicles and calcium channels for high-probability neurotransmitter release. *J Cell Biol*. 2017; 216(1):231–46. <https://doi.org/10.1083/jcb.201601098> PMID: 27998991
69. Dorus S, Busby SA, Gerike U, Shabanowitz J, Hunt DF, Karr TL. Genomic and functional evolution of the *Drosophila melanogaster* sperm proteome. *Nat Genet*. 2006; 38(12):1440–5. <https://doi.org/10.1038/ng1915> PMID: 17099714
70. Wasbrough ER, Dorus S, Hester S, Howard-Murkin J, Lilley K, Wilkin E, et al. The *Drosophila melanogaster* sperm proteome-II (DmSP-II). *J Proteomics*. 2010; 73(11):2171–85. <https://doi.org/10.1016/j.jprot.2010.09.002> PMID: 20833280
71. Schuh M, Lehner CF, Heidmann S. Incorporation of *Drosophila* CID/CENP-A and CENP-C into Centromeres during Early Embryonic Anaphase. *Curr Biol*. 2007; 17(3):237–43. <https://doi.org/10.1016/j.cub.2006.11.051> PMID: 17222555
72. Rele CP, Sandlin KM, Leung W, Reed LK. Manual annotation of genes within *Drosophila* Species: the Genomics Education Partnership protocol. *bioRxiv* [Preprint]. 2021 [posted 2020 Dec 11; revised 2020 Dec 12; revised 2020 Dec 14; revised 2020 Dec 15; revised 2021 Jan 27; revised 2021 May 4; revised 2021 May 21; revised 2021 June 5; cited 2021 Aug 10]. Available from: <https://www.biorxiv.org/content/10.1101/2020.12.10.420521v8.article-info> <https://doi.org/10.1101/2020.12.10.420521>
73. Schaeffer SW, Bhutkar A, McAllister BF, Matsuda M, Matzkin LM, O'Grady PM, et al. Polytene chromosomal maps of 11 *Drosophila* species: the order of genomic scaffolds inferred from genetic and physical maps. *Genetics*. 2008; 179(3):1601–55. <https://doi.org/10.1534/genetics.107.086074> PMID: 18622037
74. Weisman CM, Murray AW, Eddy SR. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol*. 2020; 18(11):e3000862. <https://doi.org/10.1371/journal.pbio.3000862> PMID: 33137085
75. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. HMMER web server: 2018 update. *Nucleic Acids Res*. 2018; 46(W1):W200–W4. <https://doi.org/10.1093/nar/gky448> PMID: 29905871
76. Zdobnov EM, Tegenfeldt F, Kuznetsov D, Waterhouse RM, Simão FA, Ioannidis P, et al. OrthoDB v9.1: cataloging evolutionary and functional annotations for animal, fungal, plant, archaeal, bacterial and viral orthologs. *Nucleic Acids Res*. 2017; 45(D1):D744–D9. <https://doi.org/10.1093/nar/gkw1119> PMID: 27899580
77. Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 2007; 24(8):1586–91. <https://doi.org/10.1093/molbev/msm088> PMID: 17483113

78. McGeary MK, Findlay GD. Molecular evolution of the sex peptide network in *Drosophila*. *J Evol Biol*. 2020; 33(5):629–41. <https://doi.org/10.1111/jeb.13597> PMID: 31991034
79. Hundertmark T, Gärtner SMK, Rathke C, Renkawitz-Pohl R. Nejire/dCBP-mediated histone H3 acetylation during spermatogenesis is essential for male fertility in *Drosophila melanogaster*. *PLoS One*. 2018; 13(9):e0203622. <https://doi.org/10.1371/journal.pone.0203622> PMID: 30192860
80. Rathke C, Barckmann B, Burkhard S, Jayaramaiah-Raja S, Roote J, Renkawitz-Pohl R. Distinct functions of Mst77F and protamines in nuclear shaping and chromatin condensation during *Drosophila* spermiogenesis. *Eur J Cell Biol*. 2010; 89(4):326–38. <https://doi.org/10.1016/j.ejcb.2009.09.001> PMID: 20138392
81. Kimura S, Loppin B. The *Drosophila* chromosomal protein Mst77F is processed to generate an essential component of mature sperm chromatin. *Open Biol*. 2016; 6(11):160207. <https://doi.org/10.1098/rsob.160207> PMID: 27810970
82. Tirmarche S, Kimura S, Sapey-Triomphe L, Sullivan W, Landmann F, Loppin B. *Drosophila* protamine-like Mst35Ba and Mst35Bb are required for proper sperm nuclear morphology but are dispensable for male fertility. *G3 (Bethesda)*. 2014; 4(11):2241–5. <https://doi.org/10.1534/g3.114.012724> PMID: 25236732
83. Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A, et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat Commun*. 2020; 11(1). <https://doi.org/10.1038/s41467-020-14500-z> PMID: 32034123
84. Ravi Ram K, Wolfner MF. Sustained post-mating response in *Drosophila melanogaster* requires multiple seminal fluid proteins. *PLoS Genet*. 2007; 3(12):e238. <https://doi.org/10.1371/journal.pgen.0030238> PMID: 18085830
85. Findlay GD, Sitnik JL, Wang W, Aquadro CF, Clark NL, Wolfner MF. Evolutionary rate covariation identifies new members of a protein network required for *Drosophila melanogaster* female post-mating responses. *PLoS Genet*. 2014; 10(1):e1004108. <https://doi.org/10.1371/journal.pgen.1004108> PMID: 24453993
86. Schnakenberg SL, Matias WR, Siegal ML. Sperm-storage defects and live birth in *Drosophila* females lacking spermathecal secretory cells. *PLoS Biol*. 2011; 9(11):e1001192. <https://doi.org/10.1371/journal.pbio.1001192> PMID: 22087073
87. Yeh SD, Do T, Chan C, Cordova A, Carranza F, Yamamoto EA, et al. Functional evidence that a recently evolved *Drosophila* sperm-specific gene boosts sperm competition. *Proc Natl Acad Sci U S A*. 2012; 109(6):2043–8. <https://doi.org/10.1073/pnas.1121327109> PMID: 22308475
88. Civetta A, Finn S. Do candidate genes mediating conspecific sperm precedence affect sperm competitive ability within species? A test case in *Drosophila*. *G3 (Bethesda)*. 2014; 4(9):1701–7. <https://doi.org/10.1534/g3.114.012476> PMID: 25031180
89. Wong A, Albright SN, Giebel JD, Ram KR, Ji S, Fiumera AC, et al. A role for Acp29AB, a predicted seminal fluid lectin, in female sperm storage in *Drosophila melanogaster*. *Genetics*. 2008; 180(2):921–31. <https://doi.org/10.1534/genetics.108.092106> PMID: 18757944
90. Larkin A, Marygold SJ, Antonazzo G, Attrill H, Gilberto, Garapati PV, et al. FlyBase: updates to the *Drosophila melanogaster* knowledge base. *Nucleic Acids Res*. 2021; 49(D1):D899–D907. <https://doi.org/10.1093/nar/gkaa1026> PMID: 33219682
91. Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, et al. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. *Science*. 2003; 299(5607):697–700. <https://doi.org/10.1126/science.1079190> PMID: 12511656
92. Parisi M, Nuttall R, Edwards P, Minor J, Naiman D, Lü J, et al. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. *Genome Biol*. 2004; 5(6):R40. <https://doi.org/10.1186/gb-2004-5-6-r40> PMID: 15186491
93. Vrbancovski MD, Zhang Y, Long M. General gene movement off the X chromosome in the *Drosophila* genus. *Genome Res*. 2009; 19(5):897–903. <https://doi.org/10.1101/gr.088609.108> PMID: 19251740
94. Emerson JJ, Kaessmann H, Betrán E, Long M. Extensive gene traffic on the mammalian X chromosome. *Science*. 2004; 303(5657):537–40. <https://doi.org/10.1126/science.1090042> PMID: 14739461
95. Reinke V, San Gil I, Ward S, Kazmer K. Genome-wide germline-enriched and sex-biased expression profiles in *Caenorhabditis elegans*. *Development*. 2004; 131(2):311–23. <https://doi.org/10.1242/dev.00914> PMID: 14668411
96. Argyridou E, Huylmans AK, Königer A, Parsch J. X-linkage is not a general inhibitor of tissue-specific gene expression in *Drosophila melanogaster*. *Heredity (Edinb)*. 2017; 119(1):27–34. <https://doi.org/10.1038/hdy.2017.12> PMID: 28295031

97. Kemkemer C, Catalán A, Parsch J. 'Escaping' the X chromosome leads to increased gene expression in the male germline of *Drosophila melanogaster*. *Heredity* (Edinb). 2014; 112(2):149–55. <https://doi.org/10.1038/hdy.2013.86> PMID: 24022496
98. Argyridou E, Parsch J. Regulation of the X chromosome in the germline and soma of *Drosophila melanogaster* males. *Genes* (Basel). 2018; 9(5):242. <https://doi.org/10.3390/genes9050242> PMID: 29734690
99. Vibranovski MD, Lopes HF, Karr TL, Long M. Stage-specific expression profiling of *Drosophila* spermatogenesis suggests that meiotic sex chromosome inactivation drives genomic relocation of testis-expressed genes. *PLoS Genet*. 2009; 5(11):e1000731. <https://doi.org/10.1371/journal.pgen.1000731> PMID: 19936020
100. Zhang YE, Vibranovski MD, Krinsky BH, Long M. Age-dependent chromosomal distribution of male-biased genes in *Drosophila*. *Genome Res*. 2010; 20(11):1526–33. <https://doi.org/10.1101/gr.107334.110> PMID: 20798392
101. Zhang YE, Vibranovski MD, Landback P, Marais GAB, Long M. Chromosomal redistribution of male-biased genes in mammalian evolution with two bursts of gene gain on the X chromosome. *PLOS Biol*. 2010; 8(10):e1000494. <https://doi.org/10.1371/journal.pbio.1000494> PMID: 20957185
102. Gao G, Vibranovski MD, Zhang L, Li Z, Liu M, Zhang YE, et al. A long-term demasculinization of X-linked intergenic noncoding RNAs in *Drosophila melanogaster*. *Genome Res*. 2014; 24(4):629–38. <https://doi.org/10.1101/gr.165837.113> PMID: 24407956
103. Vibranovski MD, Zhang YE, Kemkemer C, Lopes HF, Karr TL, Long M. Re-analysis of the larval testis data on meiotic sex chromosome inactivation revealed evidence for tissue-specific gene expression related to the drosophila X chromosome. *BMC Biol*. 2012; 10(1):49. <https://doi.org/10.1186/1741-7007-10-49> PMID: 22691264
104. Mahadevaraju S, Fear JM, Akeju M, Galletta BJ, Pinheiro MMLS, Avelino CC, et al. Dynamic sex chromosome expression in *Drosophila* male germ cells. *Nat Commun*. 2021; 12(1). <https://doi.org/10.1038/s41467-021-20897-y> PMID: 33563972
105. Mikhaylova LM, Nurminsky DI. Lack of global meiotic sex chromosome inactivation, and paucity of tissue-specific gene expression on the *Drosophila* X chromosome. *BMC Biol*. 2011; 9(1):29. <https://doi.org/10.1186/1741-7007-9-29> PMID: 21542906
106. Landeen EL, Muirhead CA, Wright L, Meiklejohn CD, Presgraves DC. Sex chromosome-wide transcriptional suppression and compensatory *cis*-regulatory evolution mediate gene expression in the *Drosophila* male germline. *PLoS Biol*. 2016; 14(7):e1002499. <https://doi.org/10.1371/journal.pbio.1002499> PMID: 27404402
107. Meiklejohn CD, Landeen EL, Cook JM, Kingan SB, Presgraves DC. Sex chromosome-specific regulation in the *Drosophila* male germline but little evidence for chromosomal dosage compensation or meiotic inactivation. *PLoS Biol*. 2011; 9(8):e1001126. <https://doi.org/10.1371/journal.pbio.1001126> PMID: 21857805
108. Vibranovski MD. Meiotic Sex Chromosome Inactivation in *Drosophila*. *J Genomics*. 2014; 2:104–17. <https://doi.org/10.7150/jgen.8178> PMID: 25057326
109. Wen K, Yang L, Xiong T, Di C, Ma D, Wu M, et al. Critical roles of long noncoding RNAs in *Drosophila* Spermatogenesis. *Genome Res*. 2016; 26(9):1233–44. <https://doi.org/10.1101/gr.199547.115> PMID: 27516619
110. Bouska MJ, Bai H. Long noncoding RNA regulation of spermatogenesis via the spectrin cytoskeleton in *Drosophila*. *G3* (Bethesda). 2021. <https://doi.org/10.1093/g3journal/jkab080> PMID: 33720346
111. Lim C, Tarayrah L, Chen X. Transcriptional regulation during *Drosophila* spermatogenesis. *Spermatogenesis*. 2012; 2(3):158–66. <https://doi.org/10.4161/spmg.21775> PMID: 23087835
112. White-Cooper H. Molecular mechanisms of gene regulation during *Drosophila* spermatogenesis. *Reproduction*. 2010; 139(1):11–21. <https://doi.org/10.1530/REP-09-0083> PMID: 19755484
113. Moyers BA, Zhang J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol Biol Evol*. 2015; 32(1):258–67. <https://doi.org/10.1093/molbev/msu286> PMID: 25312911
114. Moyers BA, Zhang J. Toward reducing phylostratigraphic errors and biases. *Genome Biol Evol*. 2018; 10(8):2037–48. <https://doi.org/10.1093/gbe/evy161> PMID: 30060201
115. Brown JB, Boley N, Eisman R, May GE, Stoiber MH, Duff MO, et al. Diversity and dynamics of the *Drosophila* transcriptome. *Nature*. 2014; 512(7515):393–9. <https://doi.org/10.1038/nature12962> PMID: 24670639
116. White-Cooper H. Tissue, cell type and stage-specific ectopic gene expression and RNAi induction in the *Drosophila* testis. *Spermatogenesis*. 2012; 2(1):11–22. <https://doi.org/10.4161/spmg.19088> PMID: 22553486

117. Dietzl G, Chen D, Schnorrer F, Su K-C, Barinova Y, Fellner M, et al. A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature*. 2007; 448(7150):151–6. <https://doi.org/10.1038/nature05954> PMID: 17625558
118. Ge DT, Tipping C, Brodsky MH, Zamore PD. Rapid screening for CRISPR-directed editing of the *Drosophila* genome using *white* coconversion. *G3 (Bethesda)*. 2016; 6(10):3197–206. <https://doi.org/10.1534/g3.116.032557> PMID: 27543296
119. Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in *Drosophila*. *Genetics*. 2014; 196(4):961–71. <https://doi.org/10.1534/genetics.113.160713> PMID: 24478335
120. White-Cooper H. Spermatogenesis: analysis of meiosis and morphogenesis. In: Henderson D, editor. *Drosophila Cytogenetics Protocols*. *Methods Mol Biol*. 247. Totowa, NJ: Humana Press; 2004. p. 45–75.
121. Clarkson M, Saint R. A His2AvDGFP fusion gene complements a lethal His2AvD mutant allele and provides an in vivo marker for *Drosophila* chromosome behavior. *DNA Cell Biol*. 1999; 18(6):457–62. <https://doi.org/10.1089/104454999315178> PMID: 10390154
122. Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA, et al. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature*. 2007; 450(7167):203–18. <https://doi.org/10.1038/nature06341> PMID: 17994087
123. Chen Z-X, Sturgill D, Qu J, Jiang H, Park S, Boley N, et al. Comparative validation of the *D. melanogaster* modENCODE transcriptome annotation. *Genome Res*. 2014; 24(7):1209–23. <https://doi.org/10.1101/gr.159384.113> PMID: 24985915
124. Slater G, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*. 2005; 6(1):31. <https://doi.org/10.1186/1471-2105-6-31> PMID: 15713233