# Orthologs of an essential orphan gene vary in their capacities for function and subcellular localization in *Drosophila melanogaster*

Prajal H. Patel[1], Lars A. Eicholt[2], Andreas Lange[2], Kerry L. McDermott[1], Erich Bornberg-Bauer[2,3], Geoffrey D. Findlay[1]

1. Department of Biology, College of the Holy Cross, Worcester, MA, USA 01610

2. Institute for Evolution and Biodiversity, University of Muenster, Huefferstrasse 1, 48149 Muenster, NRW, Germany

3. Department of Protein Evolution, Max-Planck Institute for Biology, Max-Planck-Ring 5, 72076 Tuebingen, BW, Germany

Correspondence to: GDF (gfindlay@holycross.edu)

ORCIDs:

PHP: 0009-0004-9368-4751

LAE: 0000-0002-3985-3698

AL: 0000-0003-3871-1986

EBB: 0000-0002-1826-3576

GDF: 0000-0001-8052-2017

**Abstract**

The rapidly evolving nature of orphan genes raises questions about whether their activities remain conserved or evolve uniquely across species. We explore this issue using *goddard* (*gdrd*), an orphan gene essential for spermatogenesis in *Drosophila melanogaster*. Although the Gdrd protein has maintained structural conservation across the *Drosophila* genus, its overall length and primary sequences, especially at its disordered termini, show substantial divergence. Using gene swaps assays that included *gdrd* orthologs from multiple species across different evolutionary distances, we investigated how lineage-specific evolutionary changes have affected Goddard's ability to function in *D. melanogaster*. Our results indicate that most orthologs possess the ability to interact with axonemes and insect ring centrioles in *D. melanogaster*, suggesting these interactions could have been present in the orthologs' common ancestor. Surprisingly, a highly divergent ortholog from *D. mojavensis* fully rescued fertility in *goddard* null *D. melanogaster*, indicating that the Gdrd protein at the base of the *Drosophila* genus was most likely fully integrated into an essential spermatogenesis pathway. However, several orthologs, including one from a more closely related species, failed to complement fertility in *D. melanogaster*. All of these complementation failures correlated with weaker axonemal binding. Furthermore, we noted that all tested orthologs exhibited divergent subcellular localization patterns. In parallel with these gene swap studies, we computationally analyzed each ortholog's structural stability and flexibility, physicochemical properties within its intrinsically disordered regions, and amino acid sequence divergence. Molecular dynamics simulations indicated structural instability in the most closely related, non-functional ortholog, while shared motifs identified in the disordered termini correlated with functionality, notably present in *D. mojavensis* and *D. melanogaster* but absent in non-rescuing orthologs. These findings suggest that although Goddard's central structure and key interactions likely existed at the base of the *Drosophila* genus, several orthologs have undergone consequential lineage-specific evolutionary changes.

## Introduction

Orphan genes are phylogenetically restricted genes that encode proteins with no detectable homology to existing proteins [1–3]. As a group, orphan genes compose roughly 10-20% of eukaryotic genomes, with 30% being an upper limit [4,5]. Because they lack functionally annotated homologs and characterized structural domains, the functions of orphan genes are difficult to predict and are largely unknown [6]. However, many orphan genes have been implicated in species-specific adaptations for behavior, environmental factors, reproduction, host pathogen interactions, and speciation [4,7,8]. Orphan genes generally exhibit short life cycles that include gene birth, gene maturation, and gene death [1,9–13]. During gene maturation, low selective constraints on orphan genes often result in the fast evolution of their protein-coding sequences [14,15]. These changes are hypothesized to be instrumental for the functional refinement of orphan proteins and their integrations into biochemical pathways [14,16].

Several origination mechanisms account for the presence of orphan genes within a species' genome. An early model suggested a role for gene duplication followed by paralog divergence as a major mechanism [14]. However, recent estimates in yeast, flies and humans suggest that only ~30 percent of orphans can be explained by duplication and divergence [17]. Other mechanisms of orphan gene origination include *de novo* gene birth, rapid divergence of ancestral genes (with and without duplication), transposable element recruitment, horizontal gene transfer, overprinting, and alternative reading frame usage [2,18].

Despite different origination mechanisms, orphan genes often share characteristics such as decreased number of introns and shorter gene lengths compared to non-orphan genes [10,19–21]. These characteristics are consistent with the different gene origin mechanisms mentioned above. Orphan genes also exhibit elevated non-synonymous to synonymous substitution rates ($d_N/d_S$) across species, indicating that these genes' amino acid sequences diverge relatively quickly [10,14,19,22]. Indeed, shorter genes and protein termini (the first and last 50 amino acids in particular) often exhibit higher evolutionary rates in general [23,24]. This amino acid sequence mutability correlates with a protein's intrinsic properties such as solvent exposure or intrinsic disorder [25–27].

Multiple studies have identified intrinsically disordered regions (IDRs) as a common feature of *de novo* emerged orphan genes [16,21,28–34]. While the functions of these disordered regions are currently unknown, IDRs in general confer on proteins crucial functional characteristics such as flexibility or accessible interaction sites [35]. As the properties of IDRs are chiefly physicochemical (e.g. charge, polarity) and thus independent of amino acid position, substitutions, insertions and deletions within IDRs are both frequent and largely inconsequential, allowing conservation of function without conservation of sequence [36–38]. While this lack of constraint on the primary sequence evolution of IDRs can be mistaken for evolution under relaxed selection in orphans, the physicochemical properties of IDRs are nevertheless under evolutionary selection [35]. Indeed, functionally important physicochemical motifs within IDRs are conserved at the sequence level [39] and can be identified using position-independent, alignment-free homology detection tools [40].

As orphan genes evolve, their rapid sequence changes may result in conservation of protein-protein interactions [6,41,42], refinement of gene function, or, at the extreme ends of the spectrum, either evolving new functions or losing all functional capacities. As our current knowledge of both the functions and interacting partners for many orphan genes remains limited, the evolutionary trajectories of orphan genes are difficult to probe. Here, as a case study, we examine *goddard* (*gdrd*), a putatively *de novo* evolved orphan gene that is essential for male fertility in *Drosophila melanogaster* [16,21,43,44]. We combined assessments of ortholog function and localization in *D. melanogaster* with AlphaFold2 (AF2) predictions, molecular dynamics (MD) simulations, physicochemical analyses of IDRs, and sequence homology analyses to test several hypotheses. First, do amino acid sequence changes that generally do not alter protein structure have consequences on a protein's function, or are such substitutions largely evolutionarily inconsequential? Second, was *gdrd* fully functional at the base of the *Drosophila* genus, as might be expected if structure and function are linked, or has the gene undergone functional refinement in the course of its lifespan? Consistent with the protein's structural conservation across the genus and the ability of a highly diverged ortholog to function in *D. melanogaster*, our results suggest that the Gdrd protein present in the hypothetical *Drosophila* ancestor was most likely integrated into a spermatogenesis-related pathway. However, within several descendant lineages of the genus, changes in amino acid sequence have led to structural instabilities and alterations in the physicochemical properties of their terminal IDRs, and these changes correlate with an inability to function in *D. melanogaster*. Along with observations that various orthologs also appear to have divergent subcellular localization patterns when exogenously expressed in *D. melanogaster* testes, these data indicate that Gdrd has maintained, lost, and possibly gained different intrinsic physiochemical and localization properties across the genus.

## Results

This study examined the ability of *gdrd* orthologs from a variety of *Drosophila* species to function in *D. melanogaster* males lacking any endogenous *gdrd* copies. As a readout of gene activity of these evolution-derived alleles, we measured the ability of orthologous proteins to either establish correct subcellular localization patterns in *D. melanogaster* testes or restore fertility to *gdrd* mutant males. We interpret the ability of orthologs to restore either capability as positive evidence of functional conservation. Negative results in these assays, however, can have several non-mutually exclusive explanations, including loss of functional conservation, lineage specific co-evolution with binding partners, and/or differential protein stability or aggregation propensity when expressed exogenously in *D. melanogaster* testes. We addressed the latter issue by examining amino acid sequence divergence among orthologs and predicting their intrinsic protein properties such as folding and stability, using AF2 and MD simulations, respectively.

### Gdrd orthologs from the *Drosophila* subgenus show substantial primary sequence divergence from the *melanogaster* protein

In our investigation, we included four relatively closely related orthologs (inclusive of the *melanogaster* gene) from subgenus *Sophophora* and three divergent orthologs from subgenus *Drosophila* (Figure 1A). These subgenera diverged ~40-43 million years ago (MYA) [45,46]. Using a combination of experimental and computational approaches, we previously showed that these orthologous Gdrd proteins each share a highly similar structural conformation that consists of a central alpha helix flanked by N- and C-terminal IDRs [44]. AF2 predictions confirm our earlier results (Figure 1A). A notable difference between orthologs present in either subgenus is the length of the terminal IDRs. The *D. mojavensis* (*repleta* group) and *D. virilis* (*virilis* group) orthologs encode proteins with extended N- and C-termini (Figure 1A) [43]. By contrast, the orthologous protein present in a Hawaiian *Drosophila* species, *D. grimshawi*, is more similar in length to the *melanogaster* protein and exhibits only a modest C-terminal extension (Figure 1A).

To measure amino acid sequence conservation among divergent orthologs, we performed pairwise BLASTP comparisons between the *melanogaster* protein and orthologs present within the genus. As expected, closely related orthologs within the *Sophophora* subgenus have high levels of amino acid conservation over the full lengths of these proteins, ranging from 46-96% identity (Figure S1). Orthologs from the *Drosophila* subgenus, however, show more limited sequence homology. In the case of *D. mojavensis*, homology is only detected within the protein's central helix and C-terminus. In *D. virilis*, homology is restricted to the central helix (Figure S1). By contrast, the *D. grimshawi* protein shows no significant alignment-based homology to the *D. melanogaster* protein (Figure S1). Pairwise BLASTP analyses between *melanogaster gdrd* N- or C-termini fail to detect significant homology in any of the *Drosophila* subgenus species, emphasizing the strong divergence likely due to increased termini lengths in *D. mojavensis* and *D. virilis* and the overall disorder of the terminal IDRs in these divergent species (Figure S1).

To get around this difficulty in our sequence homology analysis, we first performed a multiple sequence alignment for these orthologous genes using Clustal Omega (ClustalO) [47]. In the case of *Sophophora* subgenus orthologs, percent identity in full length, pairwise protein alignments to *D. melanogaster* Gdrd ranged from 42-96% (Figure 1B). This range decreases to 15-21% identity for *Drosophila* subgenus orthologs, with *D. grimshawi* being the least conserved. When we analyze percent identity by protein region (N-terminus, central helix and C-terminus), we observe similar levels of amino acid sequence conservation across the Gdrd proteins, with a few exceptions. Notably, compared to other regions of the protein, the amino acid sequence of the *D. ananassae* C-terminus appears more divergent, as does the N-terminus of *D. grimshawi* (Figure 1B).

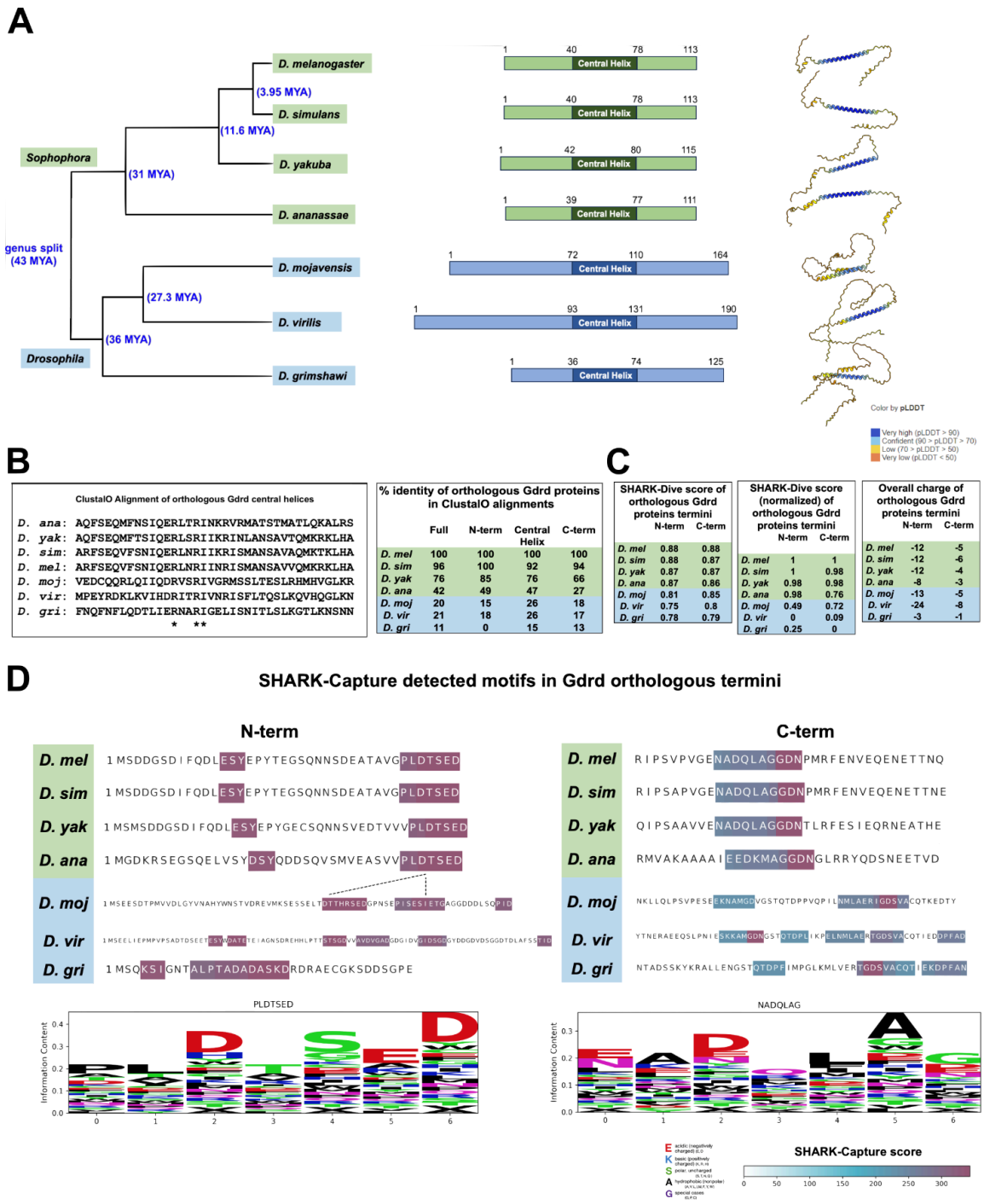



**Figure 1: Gdrd orthologs from the *Drosophila* subgenus exhibit both amino acid sequence divergence and coding sequence lengthening at both the N- and C- termini.**

(A) *Left: gdrd* orthologs are present throughout the *Drosophila* genus within both the *Drosophila* and *Sophophora* subgenera, which diverged around 43 million years ago (MYA). Divergence times (in blue brackets and font) were estimated using TimeTree. *Middle: Sophophora* subgenus *gdrd* genes (green) typically have shorter coding sequences compared to their *Drosophila* subgenus (blue) counterparts. The Gdrd orthologs share a structure and can be divided into regions of disordered N- and C-termini and a

central helix. *Right:* The central helix is conserved in all Gdrd orthologs and predicted with very high confidence by AlphaFold2 (>90 pLDDT).
(B) *Left:* ClustalO multiple sequence alignment of the central helix region. Three conserved amino acid residues common to all seven orthologs are marked with asterisks. *Right:* Analysis of amino acid conservation within ClustalO multiple sequence alignment using percent identity as a metric. The table shows percent identity conservation between *D. melanogaster* Gdrd and orthologs over the proteins' full lengths and by protein region. As expected, distantly related orthologs exhibit decreased amino acid conservation, with the *D. grimshawi* ortholog presenting the greatest sequence divergence. The disordered *D. grimshawi* Gdrd N-terminus shows no detectable primary sequence homology to the *D. melanogaster* protein.
(C) Raw (*left*) and normalized (*middle*) SHARK-Dive scores for orthologous N- and C-terminal regions and each ortholog's overall ionic charge across the lineage (*right*). With regard to SHARK-Dive scores and overall charge, both *D. mojavensis* Gdrd termini are more similar to their respective termini in *Sophophora* than to more closely related subgenus *Drosophila* orthologs. *D. virilis* and *D. grimshawi* termini differ considerably from those of *Sophophora*.
(D) Identified motifs by SHARK-Capture reveal that *D. mojavensis* Gdrd N- and C-termini contain motifs similar to *Sophophora* termini (*top*). Logos of the most common consensus motif identified in each termini (*bottom*).

High levels of disorder lead to rapid sequence evolution due to lack of structural constraints, which can confound alignment-based homology detection. We therefore used k-mer-based, alignment-free tools, SHARK-Dive [40] and SHARK-Capture [48] to analyze the disordered termini. These tools allow identification, in a position-independent manner, of conserved amino acid motifs that share physicochemical similarities. These analyses reveal that the *D. mojavensis* termini are more similar to those of *Sophophora* subgenus orthologs (Figure 1C) and share similar N-terminal motifs (PLDTSED in *Sophophora*; similar DTTHRSED & PISESIETG in *D. mojavensis* Gdrd), which are absent from the *D. grimshawi* and *D. virilis* orthologs (Figure 1D). Furthermore, the net charge of the *D. mojavensis* termini resembles the net charges of *Sophophora* subgenus ortholog termini, whereas the net charges of the *D. grimshawi* and *D. virilis* termini strongly differ (Figure 1C).

Altogether, these data reveal that while Gdrd orthologs across the *Drosophila* genus, including the highly divergent *D. grimshawi* protein, appear structurally conserved, changes in amino acid sequences and physicochemical properties have occurred in Gdrd orthologs within various sublineages of the genus. This pattern could be explained by lineage-specific adaptations through coevolution with other proteins and/or by relaxed genetic constraints.

### Expression of codon-optimized *gdrd* orthologs in *D. melanogaster*

To assess if there are functional consequences to these evolutionary changes, we performed gene swaps to investigate how well *gdrd* orthologs could restore fertility in *gdrd* null *D. melanogaster* males. Such experiments are a powerful way to understand the functional consequences of protein divergence and have previously been used in *Drosophila* to analyze changes in the functional capacities of a variety of DNA-binding proteins [49–51]. Here, we

performed ortholog swaps within a previously reported *gdrd* rescue construct that both restores fertility to *gdrd* null mutants and expresses C-terminally Hemagluttanin (HA) tagged proteins amenable to subcellular localization studies [44]. To ensure robust and comparable expression in *D. melanogaster*, we codon optimized the coding sequences of each *gdrd* ortholog and phiC31-integrated all modified constructs at the same genomic locus. Integrated constructs were then crossed into the *gdrd* null background for functional and expression analysis.

We began our analysis by replacing the wild-type *D. melanogaster gdrd* coding sequence in the original *gdrd* rescue construct with a codon-optimized *D. melanogaster* sequence. To improve on previous subcellular localization analysis performed in whole mount testes, we assessed Gdrd expression and subcellular localization in isolated cysts, which allows for superior developmental staging and imaging. Since we previously observed Gdrd localization to the basal body, transition zone, and axoneme during spermatid elongation [44], we chose to analyze Gdrd expression and localization in relation to Unc, an early marker for centriole to ciliary basal body conversion [52].

Our analysis of codon-optimized *D. melanogaster* Gdrd expression patterns and subcellular localizations in cysts recapitulates our previous observations [44] in whole-mount testes (Figure 2). Gdrd expression begins in mature spermatocytes, in which the protein has a diffuse, mostly cytoplasmic localization (Figure 2A). At this stage, spermatocytes have four Unc positive basal bodies (Figure 2A) that anchor cilia within a plasma membrane cap [52,53]. Besides containing acetylated tubulin and basal body components [54], these structures also contain components of the transition zone/ ciliary gate, including MKS1, Cep290, BD91/ BD92, and Cby [55]. Interestingly, we cannot detect any Gdrd enrichment at these ciliary structures, suggesting that Gdrd does not interact with any of the aforementioned components. Therefore, the ciliary gate and basal bodies of elongating spermatids are either compositionally distinct or are post-translationally modified during spermatid elongation to recruit Gdrd to this structure.

During spermatid elongation, Gdrd localizes both to the cytoplasm (with nuclear and mitochondrial exclusion) and to developing axonemal structures [44]. In early round spermatids characterized by unfurling mitochondrial derivatives, the *melanogaster* protein colocalizes with Unc at the basal body, axoneme, and transition zone (Figure 2B). At later spermatid elongation stages, this pattern becomes refined. Unc remains associated with the basal body and the transition zone but ceases to associate with the axoneme (Figures 2C-2D). In late round and early canoe spermatids, Gdrd appears enriched in a region distal to the Unc-positive basal body and no longer colocalizes with Unc, suggesting that this portion of the basal body or axoneme may constitute a distinct subdomain (Figure 2C). At these stages, Gdrd is both cytoplasmic and enriched at the axoneme. This sharp localization at the axoneme continues until late elongation, when Gdrd expression becomes diffuse (Figure 2C). We also observed Gdrd colocalization with Unc at the transition zones in late elongated cysts (Figure 2D). Interestingly, ciliary gate and transition zone proteins have been reported to disassemble at the onset of sperm individualization [56]. Our immunohistochemical analysis of Gdrd reveals that the protein disappears prior to this disassembly (Figure 2D).

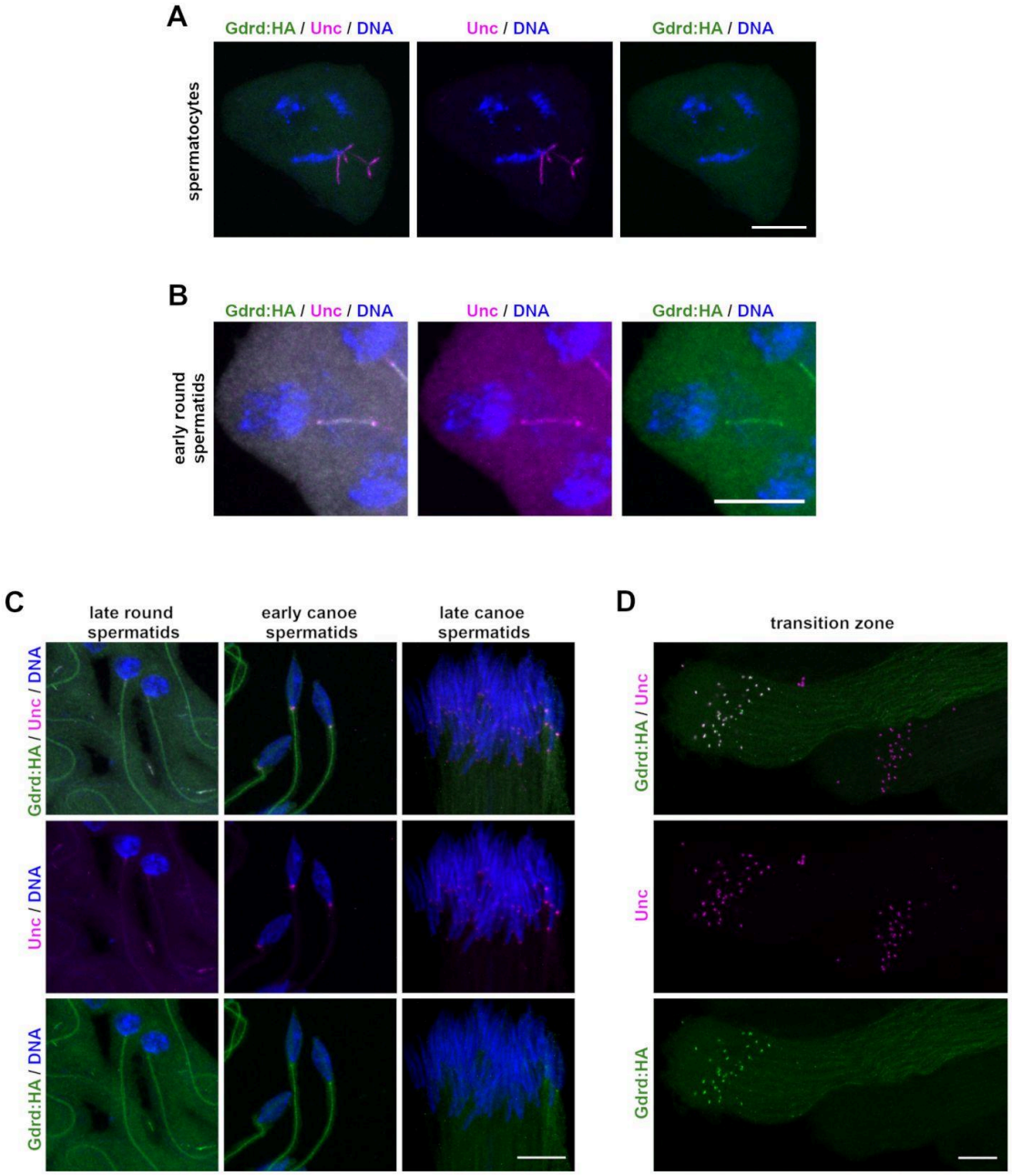



**Figure 2: Gdrd associates with the cytoplasmic flagella during spermatid elongation.**
Genotype in all images: *unc:EGFP rescue, codon-optimized (CO) D. mel gdrd:HA rescue, Δgdrd*. All images are composed of confocal z-stacks.
(A) Codon-optimized *D. melanogaster* (*D. mel*) Gdrd localizes throughout the cytoplasm but shows no enrichment at basal bodies or ciliary gates present in mature spermatocyte flagella.
(B) *D. mel* Gdrd colocalizes with Unc at the basal body, migrating insect ring centriole, and the axoneme in early round spermatids (leaf blade stage).

(C) *D. mel* Gdrd strongly decorates axonemes in late round and early canoe staged spermatid cysts. At these stages, Gdrd is enriched in a region distal to the Unc positive basal body, indicating the presence of a distinct basal body or axonemal subdomain.
(D) At the distal end of elongated spermatid cysts, Gdrd co-localizes with Unc at the transition zone/ ciliary gates (top cyst). Both Gdrd expression and localization disappear prior to the disassembly of transition zones/ ciliary gates (bottom cyst).
Scale bar for all images = 10 μm.

Although the sperm axoneme is principally cytoplasmic in many insect species, the distal tip is sequestered within a ciliary cap [54,57,58]. This compartmentalization is mediated by the insect ring centriole, an electron dense structure that lacks centriole-like organization [58]. An analogous murine sperm structure called the annulus shares protein constituents and functions with the insect ring centriole [54,59,60]. In flies, it has been hypothesized that this compartment is necessary for efficient tubulin polymerization/ axoneme growth [54]. However, while Gdrd decorates the axoneme throughout spermatid elongation, we never observe Gdrd localization beyond the transition zone, indicating either that the compartmentalized, polymerizing end of the axoneme is biochemically distinct from the cytoplasmic axoneme or that the ciliary gate excludes Gdrd from entry (Figures 2B, 2D). Altogether, these colocalization experiments suggest that Gdrd consistently associates with cytoplasmic sperm flagella until a time point just prior to the onset of sperm individualization. This experiment also indicates that the codon-optimized *D. melanogaster* Gdrd fully recapitulates our previous expression and localization results.

**The Gdrd central helix is sufficient for axoneme and transition zone localization**

Our analyses of sequence divergence indicate that the disordered N- and C- termini of Gdrd orthologs are divergent in terms of length (*D. mojavensis* and *D. virilis*), amino acid sequence (*D. ananassae* and *D. grimshawi*), and biophysical properties like charge (*D. virilis* and *D. grimshawi*). To determine the functional importance of the Gdrd N- and C-terminal IDRs, we performed structure-function analysis. We modified the codon-optimized *D. melanogaster* rescue by removing either the N-terminal or the C-terminal coding sequences. The N-terminal truncation (Gdrd ΔN) removes 36 amino acids after the initial methionine, and the C-terminal truncation (Gdrd ΔC) removes the final 35 amino acids of the protein (Figure 3A). MD simulations based on AF2 predictions show that these truncations are stable and form the central alpha helical structure of Gdrd (Figure 3B). After introducing these truncated constructs into flies, we assessed the function of each truncated protein by testing its ability to rescue fertility defects in the *gdrd* null mutant. The full length codon-optimized *gdrd* construct restores fertility to wild-type levels, which is indistinguishable from the rescuing capacity of the non-codon-optimized (wild-type) *D. melanogaster gdrd* rescue construct ($n$ = 25, unequal variance $t$-test, $P$ = 0.08) (Figure 3C). Both truncations, however, appear non-functional, indicating that both the N- and C- terminal protein regions are most likely important for function ($n$ = 25, unequal variance $t$-test, $P < 0.0001$) (Figure 3C).

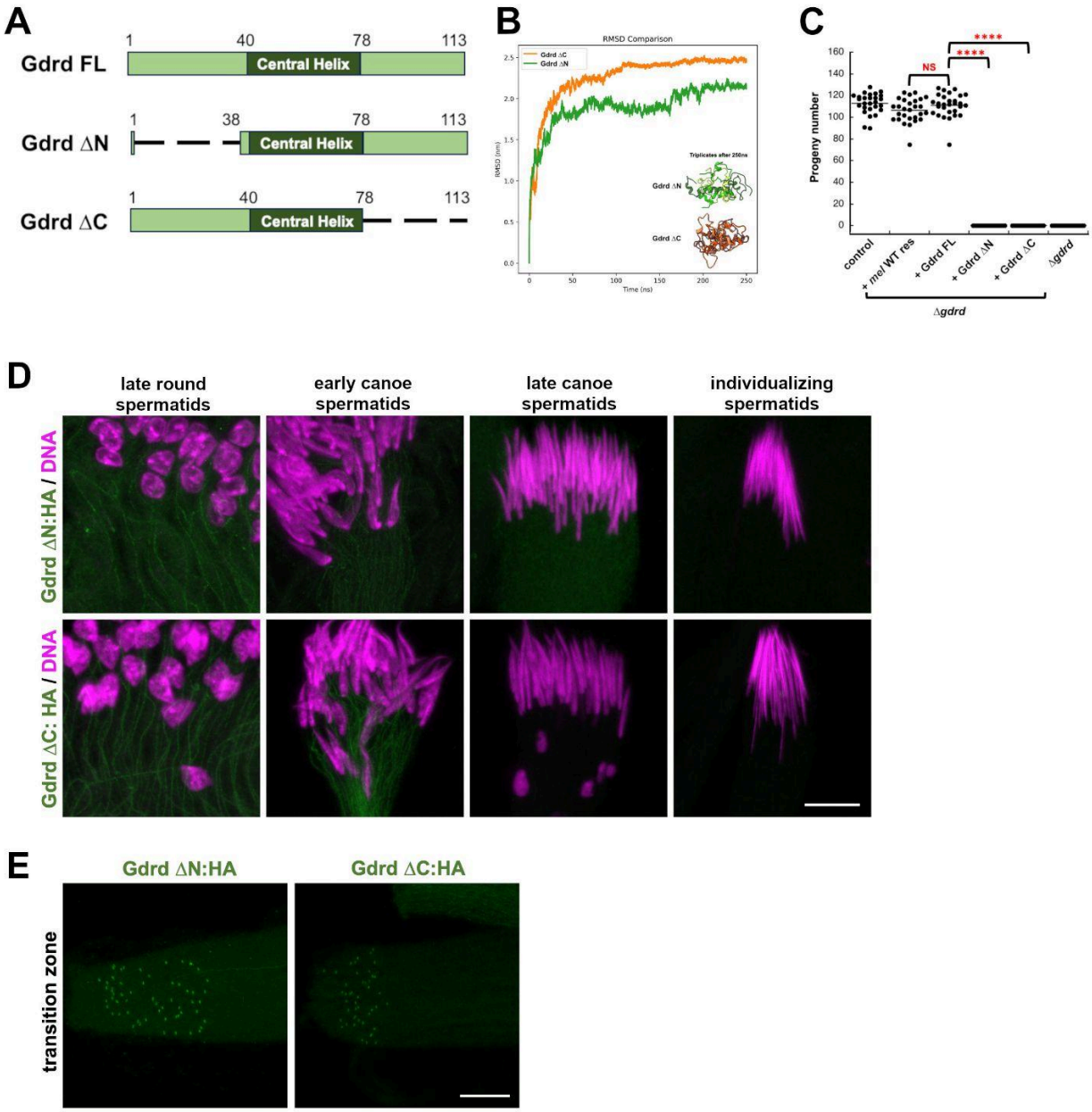



**Figure 3: Gdrd's central helix mediates localization to axonemes and transition zones.**
(A) Diagram of full length codon-optimized *D. mel* Gdrd (Gdrd FL), *D. mel* Gdrd N-terminal truncation (Gdrd ΔN), and *D. mel* Gdrd C-terminal truncation (Gdrd ΔC).
(B) Averaged backbone RMSD over triplicates of MD simulations of Gdrd ΔN and Gdrd ΔC indicate stable folding. Structurally aligned triplicate structures after 250ns show that the central helix remains largely stable. Data for each run are available in supplementary materials (Figure S5).
(C) Fertility assay to test function of full length or truncated forms of *D. mel* Gdrd. Full length codon-optimized Gdrd expression restores fertility to *Δgdrd* mutants to wild-type levels. Neither truncation, however, is functional in this assay. Statistical test: *t*-tests with unequal variance. NS = not significant, **** = $P < 0.0001$.

(D) Both N- and C- terminally truncated Gdrd proteins localize to axonemes in late round and early canoe spermatid cysts.
(E) Both N- and C- terminally truncated Gdrd proteins localize to transition zones.
Scale bar = 10 μm in all images.

We next determined if the failure of N- or C-terminally truncated Gdrd proteins to rescue fertility was due to disruptions to their subcellular localizations. We found that both truncated proteins were difficult to detect, suggesting that they are either poorly expressed or are less soluble due to the IDR truncation, as IDRs can enhance solubility [31,36,61,62]. Indeed, the C-terminally truncated protein was undetectable in late canoe stage cysts (Figure 3D). This diminished protein level could account for the inability of either construct to restore fertility to *gdrd* mutant males. However, even at low protein levels, both truncated proteins localized to axonemes in late round and early canoe staged spermatids (Figure 3D). Furthermore, we can readily detect both the N- and C-terminally truncated proteins at the transition zone (Figure 3E). Altogether, these observations suggest that Gdrd's central helix, which is common to both constructs, most likely mediates the *D. melanogaster* protein's subcellular localizations. However, both terminal IDRs are likely essential for Gdrd's function, in agreement with their structural conservation across the lineage (Figure 1A).

**The ability of *gdrd* orthologs to function in *D. melanogaster* does not correlate well with phylogenetic distance**

We next replaced the *D. melanogaster gdrd* coding sequence in the rescue construct with codon-optimized *gdrd* coding sequences from the *D. simulans*, *D. yakuba*, *D. ananassae*, *D. mojavensis*, *D. virilis*, or *D. grimshawi* orthologous genes. Since we cannot detect Gdrd-HA on western blots, we analyzed the expression of these constructs in whole testes using immunohistochemistry. We find that all exogenously introduced *gdrd* orthologs are expressible in *D. melanogaster* testes (Figure 4). As all these genes are expressed under the *D. melanogaster* gene's regulatory sequences, including the *melanogaster* gene's 5' and 3' UTRs, the testes expression pattern of all *gdrd* orthologs mimic the *D. melanogaster* gene's pattern, with protein expression peaking in post-meiotic round spermatid cells (Figure 4). To the extent that it can be evaluated using immunofluorescence, however, the abundance of the *D. ananassae*, *D. virilis*, and *D. grimshawi* orthologs appear weaker, suggesting that these proteins are less expressed, unstable, have different or weakened localization patterns, or a combination of these factors (Figure 4). Furthermore, the signal in testes expressing the *D. grimshawi* ortholog is both fainter and more diffuse in older elongating spermatid cysts, indicating that the *D. grimshawi* protein also fails to associate with axonemes (Figure 4). Expression of the *D. grimshawi* ortholog, however, is robust at early spermatid stages, and the protein appears to associate with the plasma membrane. Altogether, these data indicate that orthologs can be expressed in *D. melanogaster* testes.

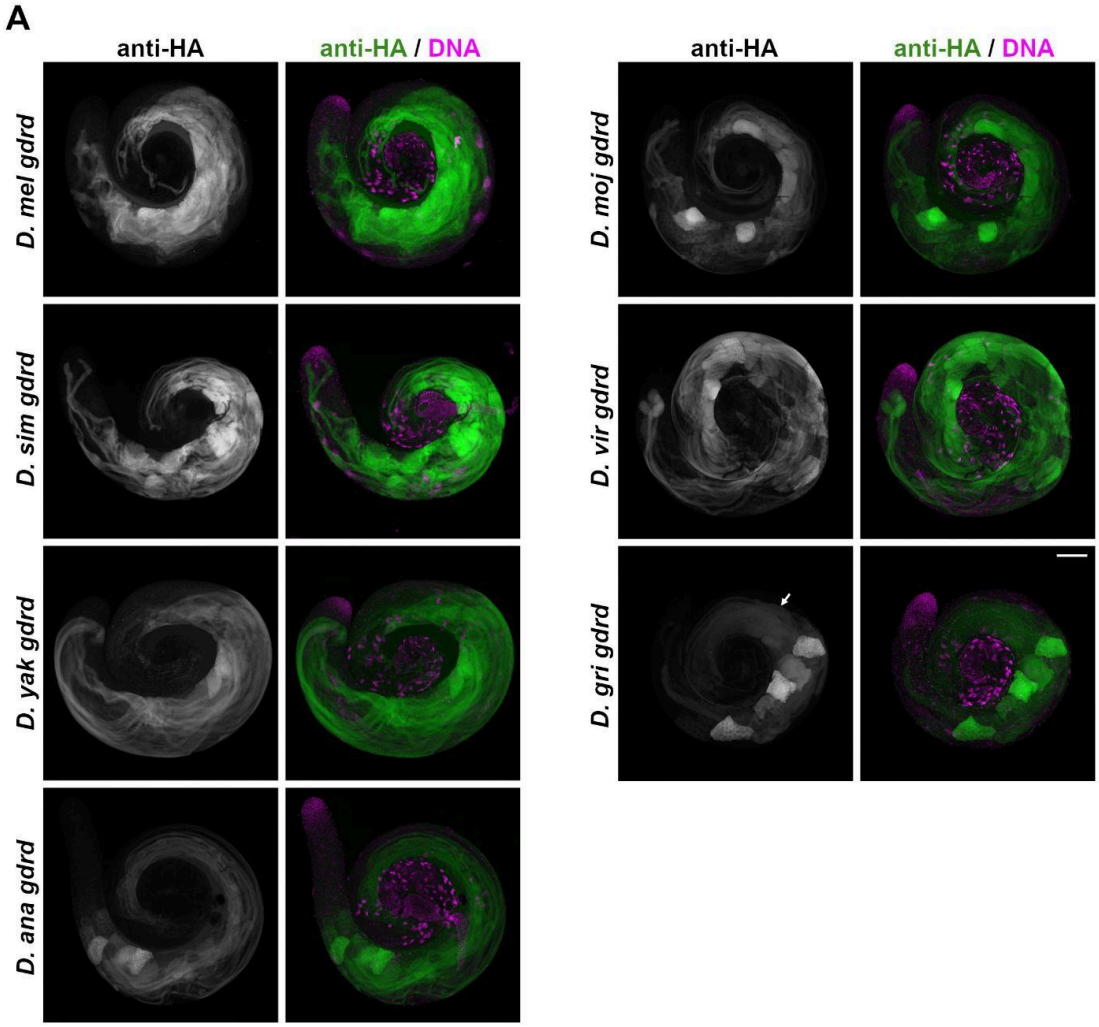



**Figure 4: Exogenous expression of *gdrd* orthologs in *D. melanogaster* testes appears largely robust.**

*D. melanogaster* testes were labeled with anti-HA (gray scale or green) to detect the abundance of orthologous Gdrd proteins. The apical ends of the testes, which contain the youngest cysts, are on the left side of the image, and the basal ends containing the nuclei of older cysts are at the center. The anti-HA signal intensity of *D. melanogaster* (*D. mel*) and *D. simulans* (*D. sim*) orthologs appears robust. This signal appears diminished in more phylogenetically distant orthologs, *D. yakuba* (*D. yak*), *D. ananassae* (*D. ana*), *D. mojavensis* (*D. moj*), *D. virilis* (*D. vir*), and *D. grimshawi* (*D. gri*). The signal appears diffuse in *D. grimshawi* late elongating spermatid cysts (arrow), suggesting that the *D. grimshawi* protein doesn't localize to axonemes (arrow). Nuclei are labeled with DAPI (magenta). Scale bar = 100 µm.

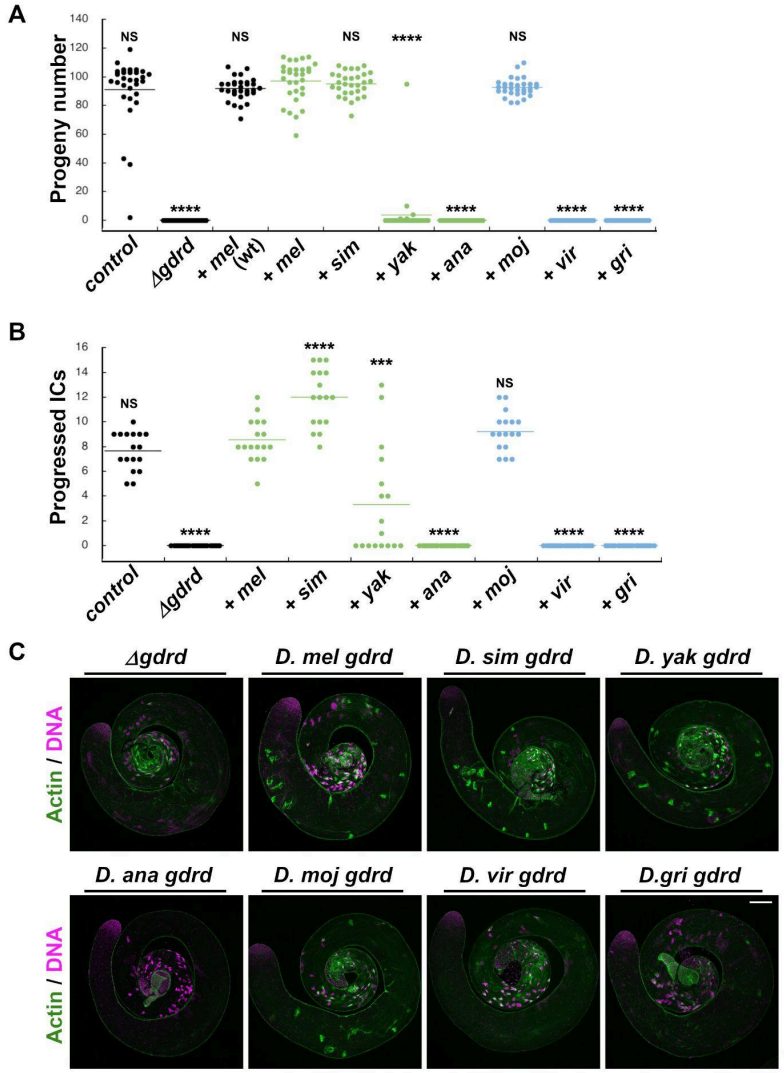



**Figure 5: A *gdrd* ortholog that diverged over 40 million years ago fully restores fertility to *gdrd* null *D. melanogaster* males, while several other orthologs of equal or lesser divergence do not.**

(A) Fertility assay to assess the function of *gdrd* orthologs in *D. melanogaster*. *Δgdrd* mutants expressing codon-optimized *D. melanogaster* (*mel*) Gdrd exhibit similar fertility levels when compared with either control flies ($w^{11118}$) *or Δgdrd* mutants expressing the original, non-codon optimized *D. melanogaster* rescue (*mel wt*) ($P$ = 0.25 and 0.08, respectively). *Δgdrd* mutants expressing codon-optimized *D. melanogaster* (*mel*) also exhibit similar fertility levels when compared with *Δgdrd* mutants expressing the *D. simulans* (*sim*) or *D. mojavensis* (*moj*) orthologs ($P$ = 0.55 and 0.14 respectively). The *D. yakuba* ortholog (*yak*) appears only partially functional (mean progeny number = 3.7; $P < 0.0001$). The expression of *D. ananassae* (*sim*), *D. virilis* (*vir*), or *D. grimshawi* (*gri*) orthologs fails to rescue *Δgdrd* male sterility ($P < 0.0001$ in all cases). Statistical test: *t*-tests with unequal variance. NS = not significant, **** = $P < 0.0001$.

(B) Testes from either non-rescuing (*ana*, *vir, gri*) or poorly rescuing (*yak*) ortholog gene swaps display either no or diminished progressed ICs, suggesting a defect in initiating spermatid individualization. The partial restoration of IC progression in *yak*-expressing testes is consistent with the partial rescue observed in the fertility assay (Panel A). $P < 0.0001$ for *ana*, *vir*, and *gri* orthologs and P = 0.00015 for *yak*.

Expression of the *simulans* ortholog increases the number of progressed ICs ($P < 0.0001$). Statistical test: *t*-tests with unequal variance. NS = not significant, *** = $P < 0.001$, **** = $P < 0.0001$.
(C) In *gdrd* mutants, ICs form but do not translocate along the elongated spermatid cysts. Progressing ICs are readily observable in fertile or partially fertile ortholog gene swaps as determined by phalloidin labeling of actin-rich ICs (green). *D. ananassae* gene swap testes, however, often lack IC formation, indicating that exogenous expression of the *D. ana* ortholog enhances the *gdrd* mutant phenotype. DAPI labels DNA (magenta). Scale bar = 100 μm.

To determine if *gdrd* orthologs can complement fertility defects observed in the *gdrd* mutants, we performed a fertility assay. The codon-optimized *D. melanogaster gdrd* restores fertility to wild-type levels ($n$ = 30, unequal variance *t*-test, $P$ = 0.25) (Figure 5A). Despite sequence and structural conservation of *Sophophora* Gdrd orthologs (Figure 1), however, these proteins vary in their capabilities to restore fertility to *gdrd* mutants. Unsurprisingly, the *simulans gdrd* ortholog, which shares 96% identity with the *D. melanogaster* ortholog, fully restores fertility ($n$ = 30, unequal variance *t*-test, $P$ = 0.55) (Figure 5A). By contrast, the *D. yakuba* ortholog restores only a low level of fertility (4-43% of *D. melanogaster* ortholog's fertility levels, depending on the replicate; average progeny number in *D. yakuba* replicates (±s.e.) = 3.7/45.9 (± 3.2/8.8)), despite a high level (76%) of sequence identity ($n$ = 30, unequal variance *t*-test, $P < 0.0001$) (Figure 5A, Figure S2). The more divergent *D. ananassae* ortholog (42% sequence similarity) consistently fails to restore any fertility to *gdrd* mutants ($n$ = 30, unequal variance *t*-test, $P < 0.0001$) (Figure 5A). These data suggest that even when both sequence and structure are fairly well conserved, several *Sophophora* subgenus *gdrd* orthologs cannot fully compensate for loss of the *D. melanogaster* gene.

Strikingly, the highly divergent *gdrd* orthologs present within the *Drosophila* subgenus also exhibit variable capacities to restore fertility in *D. melanogaster gdrd* mutants. Much to our surprise, the divergent *D. mojavensis* ortholog (20% amino acid sequence identity) fully restores fertility to wild-type levels (n = 30, unequal variance *t*-test, $P$ = 0.14). This finding has several implications. First, it suggests that the protein is functional in *D. melanogaster* despite considerable amino acid sequence divergence. Second, because the common ancestor of *D. mojavensis* and *D. melanogaster* corresponds to the root of the *Drosophila* genus, the ancestral gene present in this ancestor likely already possessed all of the major functions of extant *D. melanogaster gdrd*. Third, the lengthening of the terminal IDRs in *D. mojavensis* (Figures 1A & 5A) most likely has no major consequence on the protein's central functions. On the other hand, the equally divergent orthologs of *D. virilis* (21% identity), which also contains longer IDRs, and *D. grimshawi* (11% identity) fail to restore fertility (n = 30, unequal variance *t*-tests, $P < 0.0001$) (Figures 1B & 5A). Thus, within the context of these gene swap assays, amino acid sequence divergence has diminished the functional capabilities of some *gdrd* orthologs in *D. melanogaster* while being inconsequential in others.

As *gdrd* loss of function mutations lead to failure of sperm individualization [44], we assessed this process in gene swapped testes by quantifying progressed individualizing complexes (ICs). In the cases of orthologs that show no rescue ability, we observe no progressed ICs, indicating that these orthologs appear fully nonfunctional in this assay (Figure 5B). On the other hand, *gdrd* mutant testes expressing the *D. yakuba* ortholog, which restores low levels of fertility to

*gdrd* mutants, exhibit some, but significantly fewer, translocating ICs than *D. melanogaster*-expressing testes (average number of progressing ICs (± s.e.) in *D. melanogaster* and *D. yakuba* expressing testes, respectively = 8.5 (± 0.4) and 3.3 (± 1.1); $n$ = 15, unequal variance $t$-tests, $P < 0.0001$) (Figure 5B). Interestingly, *gdrd* mutant testes expressing the *simulans* ortholog appear to undergo increased sperm individualization despite restoring fertility to levels observed in either wild-type or codon-optimized *D. melanogaster* expressing flies (Figure 5B). By contrast, *gdrd* mutant testes expressing the *D. ananassae* ortholog exhibit phenotypic defects beyond what we observed in *gdrd* null mutants. While ICs formation is decreased in *gdrd* mutants compared to wild-type [44], the majority of mutant testes expressing the *D. ananassae* ortholog exhibit no IC formation, suggesting that spermatid elongation never reaches completion (Figure 5D). As we observed this phenomenon in two independently generated insertion strains, we believe that loss of IC formation in this genotype is not due to secondary mutations. Expression of the exogenous *D. ananassae* ortholog in *D. melanogaster* flies may interfere with other spermatogenic processes, resulting in the attenuation of spermatid elongation. To address if this observed attenuation occurs in either a dominant or dominant negative manner, we assessed the fertility of flies carrying one copy of the *D. ananassae* ortholog and one copy of the *D. melanogaster* ortholog within a *gdrd* mutant background. We find that these flies have fertility levels similar to controls (Figure S3), suggesting that the *D. ananassae* ortholog acts in a recessive manner and that the attenuation is weak.

**Orthologous Gdrd proteins exhibit divergent subcellular localization patterns**

A potential explanation for diminished function in non-rescuing *gdrd* orthologs might be disrupted protein localization. To examine this possibility, we observed the associations of HA-tagged orthologous Gdrd proteins with either the axoneme or the transition zone. As the *D. grimshawi* protein showed weak expression and no axonemal localization (Figure 4), we excluded this ortholog in all remaining analyses.

As previously described (Figure 2C), the codon-optimized *D. melanogaster* protein decorates the axoneme in late round spermatids, a stage of spermiogenesis characterized by significant axoneme and tail elongation but minimal nuclear shaping. This localization persists well into late canoe stage spermatids and disappears in individualizing cysts that are characterized by needle-shaped nuclei (Figure 6A). Both *D. simulans* Gdrd, which fully rescues fertility defects, and *D. yakuba* Gdrd, which exhibits weaker rescue in fertility assays, show robust localization to the axoneme during spermatid elongation (Figure 6A). By contrast, the *D. ananassae* Gdrd ortholog, which fails to rescue *gdrd* mutant fertility defects, poorly localizes to axonemes. Finally, while both orthologs in the *Drosophila* subgenus localize to axonemes, the localization of the *D. virilis* Gdrd protein appears fainter and more cytoplasmic (Figure 6A). Thus, while all orthologs maintain some ability to localize to axonemes, orthologous proteins that localize more weakly to this structure fail to rescue *gdrd* loss of function mutations.

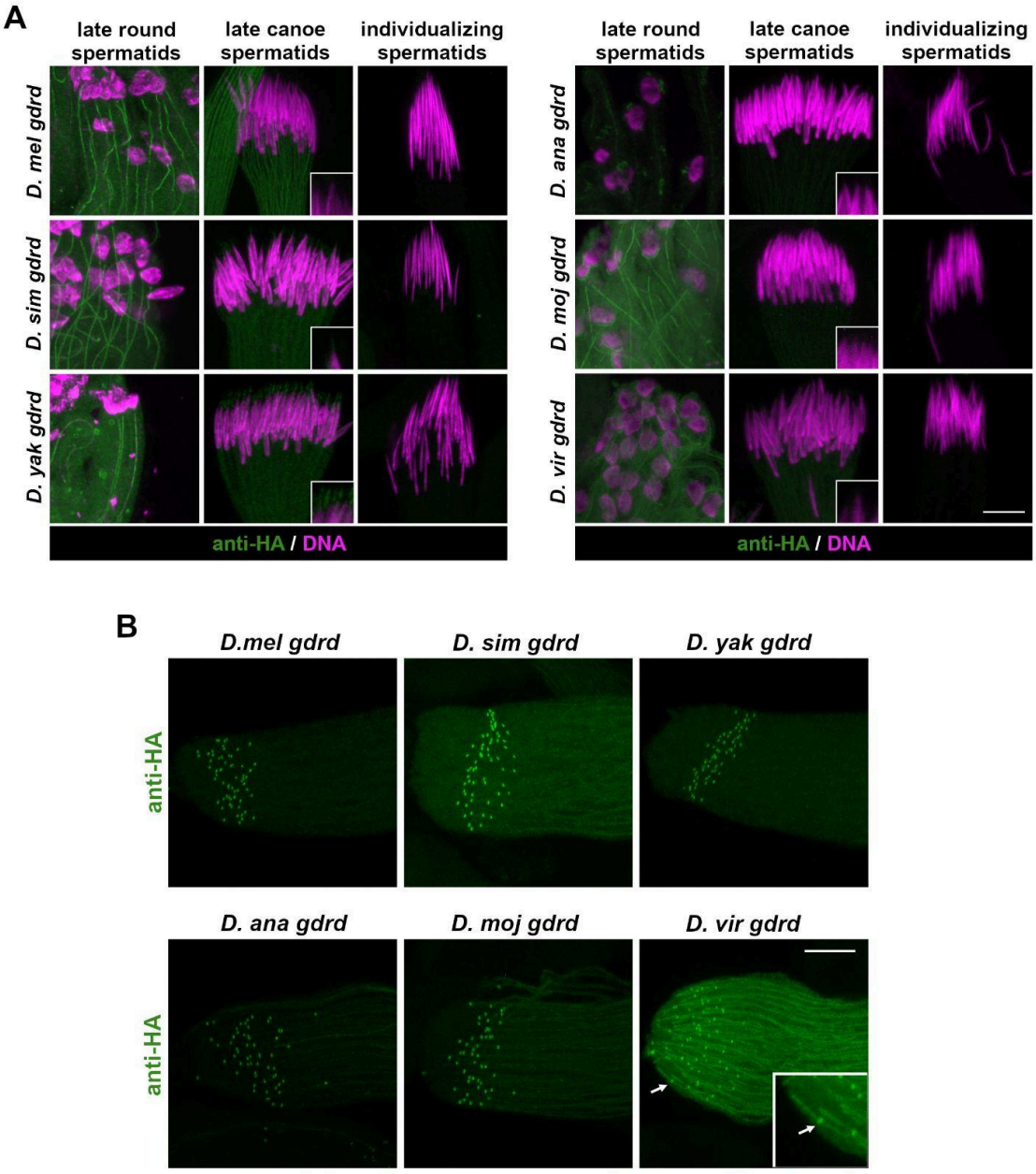



**Figure 6: Gdrd orthologs show variable localization to axonemes and often exhibit divergent subcellular localization patterns in late spermatids.**

In all images, anti-HA (green) and DAPI (magenta) detect HA-tagged Gdrd orthologs and DNA, respectively. (A) Comparisons of ortholog localizations during late spermatid elongation. The basal end of each cyst is at the top of the image. All orthologs are expressed in late round and late canoe stage spermatids. Likewise, expression is lost in individualizing spermatid cysts. *D. mel* Gdrd protein decorates axonemes in late round spermatids and late canoe stage spermatids. This pattern is observed for all rescuing and partially rescuing ortholog gene swaps (*D. sim*, *D. yak*, and *D. moj*). Axonemal localization is reduced or greatly diminished in the *D. vir* and *D. ana* ortholog gene swaps. Orthologs also exhibit divergent localizations at this stage. The *D. sim* and *D. yak* proteins localize to the basal end of nuclei in

late canoe stage cysts (see insets), with the signal in *D. yak* being more robust. The *D. yak* ortholog also exhibits localization at a round organellar structure in late round stage spermatid cysts. Both the *D. moj* and *D. vir* proteins localize to the apical surface of the round nuclei.
(B) All Gdrd orthologs localize to the transition zone. The *D. vir* ortholog breaches the ciliary gate (inset). Scale bar = 10 µm.

Interestingly, every tested ortholog also exhibits some localization pattern that diverges from the *D. melanogaster* protein. In late canoe stage spermatid cysts, the *D. simulans* and *D. yakuba* Gdrd proteins both associate with a structure near the acrosome present at the basal end of spermatid nuclei (Figure 6A). *D. yakuba* Gdrd also exhibits further divergent localization to a round organellar structure visible in late round spermatids (Figure 6A). The *D. ananassae* protein, on the other hand, divergently decorates the surface of large oblong organellar structures that abut either round nuclei or the axoneme (Figure 6A). The two orthologs from the *Drosophila* subgenus both exhibit polarized localization along the apical side of the nuclear membrane reminiscent of the pattern observed for proteins that associate with the nuclear cap. This localization appears weaker in *D. virilis*, as does axoneme localization (Figure 6A). Taken together, these data raise the possibility that divergent Gdrd orthologs might have evolved interactions with additional partners.

As *D. melanogaster* Gdrd also localizes to the transition zone, we examined the ability of orthologous proteins to localize to this structure. As the distal end of spermatid cysts cannot be developmentally staged, we cannot make inferences about localization strength. Interestingly, all tested orthologs maintain localization to the transition zones, including the *D. ananassae* protein, which showed the most diminished localization at axonemes (Figure 6B). Notably, the *D. virilis* protein associates with the axoneme within the ciliary cap (Figure 6B). This suggests that the *D. virilis* Gdrd might have evolved a novel subcellular localization pattern. Alternatively, the presence of *D. virilis* Gdrd within the ciliary cap in *D . melanogaster* might reflect a *gdrd* loss of function phenotype.

These divergent localization patterns, however, are not restricted to elongating spermatid stages, when the Gdrd protein most likely functions, but are often mirrored in mature spermatocytes and early round spermatids. In mature spermatocytes, the *D. melanogaster* Gdrd protein has a diffuse cytoplasmic distribution, as do the *D. simulans* and *D. yakuba* orthologs (Figure 7A). Notably, *D. yakuba* Gdrd expression appears consistently weaker at this stage of spermatogenesis. By contrast, *D. ananassae* Gdrd localizes to round organellar structures present throughout the spermatocyte cytoplasm, while the *D. mojavensis* and *D. virilis* proteins decorate the spermatocyte's nuclear perimeter (Figure 7A).

We also observe similar divergent localizations in post-meiotic early round spermatid staged cysts. During these stages, the *D. melanogaster* Gdrd protein has a dynamic expression and localization pattern. In developing sperm cells that have just exited meiosis, Gdrd expression is initially weak, but increases during the onion and leaf blade stages, when the round mitochondrial derivative called the nebenkern forms or unfurls, respectively (Figure 7B). The *D. melanogaster* Gdrd protein at all three stages is chiefly cytoplasmic, showing exclusion at both nuclei and all mitochondrial derivatives (Figure 7B). At the leaf blade stage, when axonemes are

forming, Gdrd decorates the basal body, the growing axoneme, and the transition zone/ ciliary gate (Figure 2B, Figure 7B).

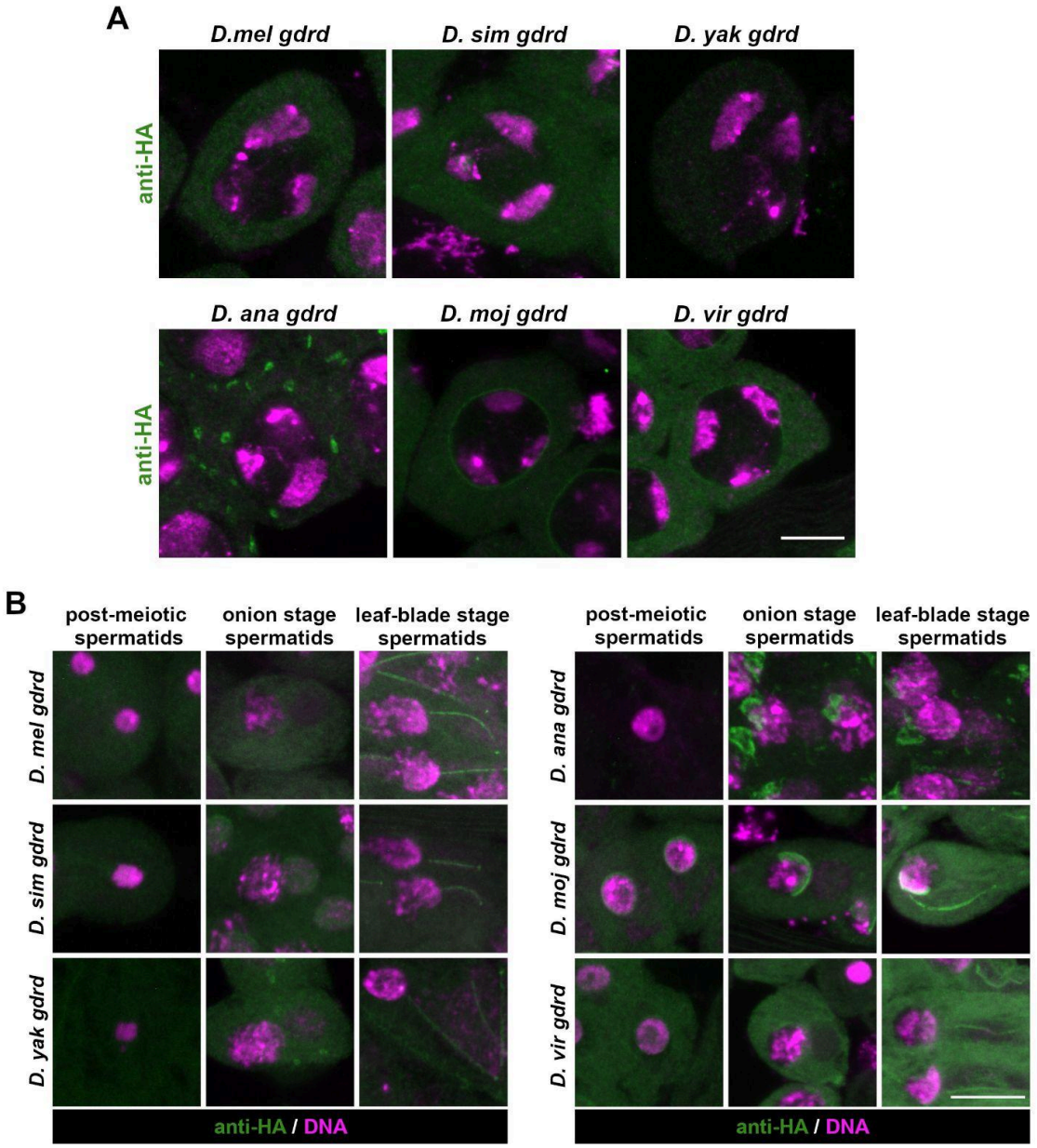



**Figure 7: Gdrd orthologs exhibit divergent subcellular localization patterns in spermatocytes and early round spermatids.**

In all images, anti-HA (green) and DAPI (magenta) detect HA-tagged Gdrd orthologs and DNA, respectively. (A) Spermatocytes expressing Gdrd orthologs. *D. mel*, *D. sim*, and *D. yak* Gdrd proteins appear chiefly cytoplasmic, with substantial nuclear exclusion. The *D. ana* protein decorates round vesicular organelle structures, while the *D. moj* and *D. vir* proteins are enriched at the nuclear perimeter.
(B) Early round spermatids expressing Gdrd orthologs. Gdrd ortholog expression and subcellular localization was followed at three stages: round spermatid just after meiosis (post-meiotic spermatids), onion stage spermatids characterized by nebenkern formation, and leaf-blade stage spermatids

characterized by mitochondrial derivative unfurling and axoneme formation. DAPI labels the nuclei and the mitochondrial DNA present in the nebenkern/ mitochondrial derivatives. *D. mel* Gdrd is cytoplasmic until the leaf blade stage when it localizes to the basal body, axoneme, and transition zone. *D. sim* and *D. yak* Gdrd have a similar dynamic except that the proteins transiently localize to the nebenkern and round organellar structures, respectively, during the onion stage. The *D. ana* protein localizes to a structure reminiscent of the acroblast in both the onion and leaf blade stages. Notably, the *D. ana* protein shows weak localization to axonemes at the leaf blade stage. At all three stages, *D. moj* and *D. vir* proteins have identical dynamics. The proteins decorate the nuclear perimeter in post-meiotic spermatids. This nuclear envelope distribution becomes restricted to the apical side of the nucleus during onion and leaf blade stages. In general, all *D. vir* ortholog localizations appear weaker when compared to the *D. moj* protein. Scale bar = 10 µm.

The *simulans* Gdrd protein localizes to the nebenkern at the onion stage, but is not associated with the mitochondrial derivative once axoneme elongation begins at the leaf blade stage (Figure 7B). The *D. yakuba* protein shows a similar dynamic. At the onion stage, *D. yakuba* Gdrd is present in round organellar structures similar to the ones observed at later developmental stages (Figure 6A), but at the leaf blade stage, the protein localizes exclusively to the axoneme (Figure 7B). Potentially, this suggests that when Gdrd orthologs are unable to bind to/ interact with partners present in the developing axonemal structure, they instead either mislocalize or interact with other partners. An alternate possibility is that the interacting partners present at either the nebenkern or round organellar structures disappear in the leaf blade stage. By contrast, the *D. ananassae* protein has weak expression in post-meiotic spermatids (Figure 7B). The protein, however, localizes to the developing acrosome in both onion stage and leaf-blade stage spermatids suggesting that vesicular localizations observed in spermatocytes may correspond with Golgi bodies (Figure 7B). Interestingly, we observed very little localization to the axoneme at the leaf blade stage, which potentially suggests that the *D. ananassae* protein preferentially interacts with acrosome (Figure 7B). Finally, the *D. mojavensis* and *D. virilis* Gdrd proteins have similar expression and localization dynamics at these stages. Just after meiosis, both proteins localize to the periphery of the nucleus (Figure 7B). At the onion stage, however, this uniform nuclear membrane distribution shifts so that the proteins become asymmetrically localized to the apical nuclear surface (Figure 7B). This nuclear cap localization is maintained during the leaf blade stage (Figure 7B). Interestingly, the *D. virilis* protein remains heavily cytoplasmic at all stages, confirming that the protein has a weakened affinity for the *D. melanogaster* axoneme.

Altogether, these data suggest that despite structural conservation, the amino acid sequence divergence observed in Gdrd orthologs has major effects on the functions and localizations of these proteins, at least to the extent that they can be evaluated in *D. melanogaster* cells.

### *D. ananassae* and *D. virilis* Gdrd orthologs appear unstable in MD simulations

We next investigated whether the ability of an ortholog to function in *D. melanogaster* correlated with its predicted structural stability. To do so, we performed structure predictions using AF2 and subsequently assessed their stability through molecular dynamics (MD) simulations conducted in triplicate over 250 ns. Previous studies have shown that structural predictions for

orphan proteins can yield implausible results [33,63,64], and MD simulations can test the stability of the predicted structures [6,16,44].

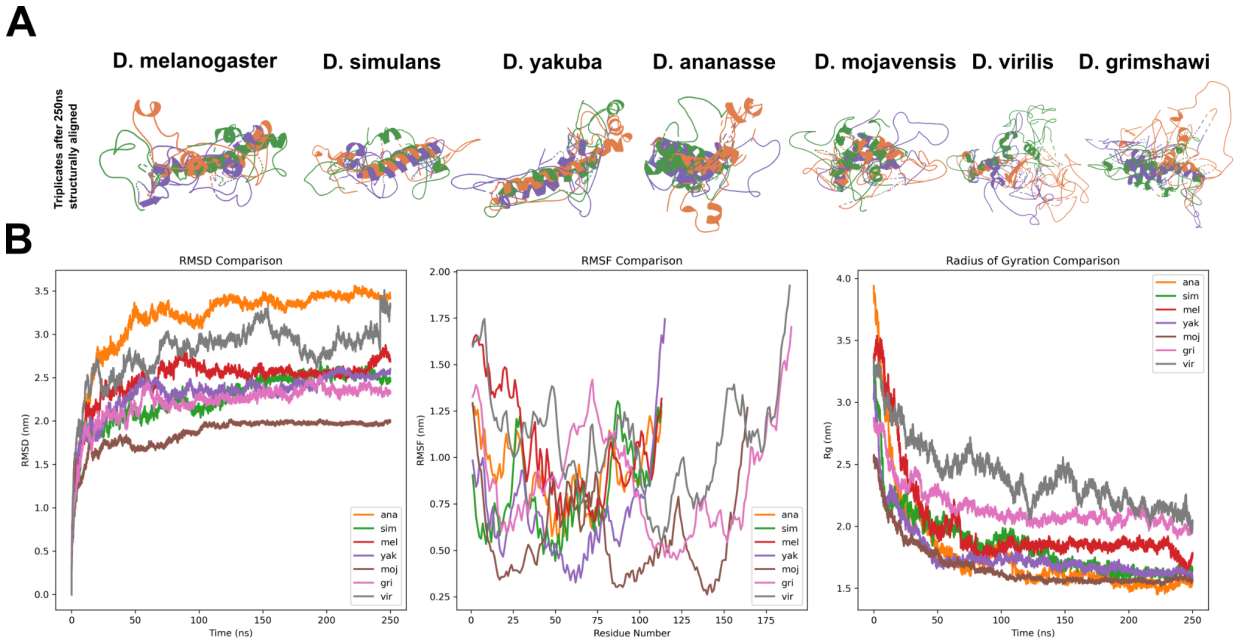



**Figure 8: MD simulations of Gdrd orthologs over 250ns.** (A) Structural alignments of triplicate MD simulations after 250 ns highlight the dynamic structures of Gdrd orthologs while the central helix remains stable. (B) Averaged backbone RMSD over time (left), $\alpha$-C RMSF per residue (center) and radius of gyration over time (right) of all Gdrd orthologs. *D. ananassae* shows by far the highest RMSD indicating relative instability. $\alpha$-C RMSF is high at the termini for all orthologs, in accordance with disordered termini, while central residues fluctuate less and remain a structured helix overall. Trend of radii of gyration shows that all orthologs become more compact over time, likely forming molten globule-like structures. Single runs available in supplementary materials (Figure S4).

Our modeling and MD simulations revealed that all orthologous Gdrd proteins exhibit significant structural flexibility. Notably, flexible terminal regions were consistently observed, characterized by disorder interspersed with transient formation of short helices. In contrast, the central helical region remained largely stable across all simulations (Figure 8). We further quantified this intrinsic flexibility through Root Mean Square Deviation (RMSD) analyses, which indicated relatively high deviation values reflective of pronounced dynamic behavior [65] (Figure 8B, left). Among the studied orthologs, *D. ananassae* showed the highest RMSD values, suggesting it to be either the most structurally flexible or least stable of the group. On the other hand, *D. mojavensis* appears to be the most stable structure.

Residue-specific flexibility examined using Root Mean Square Fluctuation (RMSF) analysis (Figure 8B, center) confirmed that the stable central helix exhibits substantially fewer fluctuations than the disordered termini, which demonstrate extensive flexibility as anticipated. Furthermore, the Radius of Gyration (Rg) analysis provides additional insights into the overall compactness and structural stability of each ortholog (Figure 8B, right). Consistent with RMSD and RMSF observations, orthologs display decreasing radii of gyration over simulation time, indicative of initial structural relaxation followed by stable yet flexible conformations, indicative of

molten globule-like structures. Together, these findings substantiate the high structural plasticity of Gdrd proteins, underscoring both the conserved central helix and the dynamic terminal IDRs, which may reflect distinct functional adaptations among orthologs.

## Discussion

The use of orthologous gene swaps allows us to test protein function and conservation in the context of physiologically similar cells, with the caveat that lineage-specific co-evolution of both the focal protein and its interacting partners can confound the interpretation of results. To partially circumvent this issue, we coupled these functional analyses with *in silico* structural modeling and sequence analysis. Our goal was to determine whether intrinsic protein qualities, such as protein stability or flexibility, have changed in various evolutionary lineages and whether these changes might account for functional differences among orthologs. We assume here that both the folded and the disordered regions of a protein's structure would evolve under selective constraints if they are required to maintain an orphan protein's activities, such as interactions with membranes, droplets, ligands or other proteins [6,41,42,66]. As a case study, we analyzed the functionally characterized and structurally conserved orphan gene *gdrd* for the purposes of answering several questions about the gene's evolution. First, are there functional consequences to amino acid sequence divergences observed in orphan genes, while structure remains conserved? *gdrd* is an ideal candidate for addressing this question as the gene shares many of the characteristics of orphan genes (particularly those of likely *de novo* origin), including a short coding sequence, lack of introns, high levels of protein disorder, and sequence divergence in distant species. Second, can distant orthologs functionally substitute for the *D. melanogaster* protein? Third, can we correlate alterations in protein sequence and/or structure with the procurement of essential functions?

One of our motivations for performing this analysis was to investigate whether structural or sequence changes that occurred specifically in *Sophophora* subgenus Gdrd orthologs were important in developing the gene's essential function in *D. melanogaster* spermatogenesis. Protein modeling of Gdrd orthologs across the *Drosophila* genus revealed that they all share a central helix with terminal IDRs (Figures 1 & 8). Orthologs present in two *Drosophila* subgenus members, *D. mojavensis* and *D. virilis*, have elongated N-termini (Figure 1), suggesting that the shortening of this IDR could potentially be important for function in *D. melanogaster*. Contrary to this hypothesis, our results demonstrate that the Gdrd ortholog from *D. mojavensis* (of the *Drosophila* subgenus) is fully functional in *D. melanogaster*. Thus, the evolution of a shortened N-terminal IDR was likely not critical for the protein's function. Furthermore, this result indicates that *gdrd* was most likely functional at the base of the *Sophophora* and *Drosophila* subgenera and that at least some of the gene's functional and protein-protein interaction capabilities are both conserved and emerged long ago (~40-43 MYA, Figure 1).

This finding is not wholly surprising. Several studies have shown that *de novo* evolved orphan genes maintain their characteristics over time, rather than evolving into more structured proteins [6,16,42,44,67,68]. Indeed, our previous [44] and current analyses of Gdrd ortholog structures

(Figures 1 & 8) show that the extant orthologs share a high level of structural conservation of the central helix and terminal IDRs, suggestive of evolution under selective constraints. Thus, the earliest forms of the protein likely shared these conserved structural features.

Using fertility and spermatid individualization as metrics for *gdrd* function, we also observed no strong relationship between phylogenetic distance from *D. melanogaster* and an ortholog's ability to complement *gdrd* mutant fertility (Figure 5). Closely related orthologs present in *D. simulans* and *D. yakuba* are clearly functional when expressed in *D. melanogaster* testes, but so is the extremely divergent *D. mojavensis* ortholog. By contrast, we also identified an ortholog within subgenus *Sophophora* (*D. ananassae*) and two distant orthologs (*D. grimshawi* and *D. virilis*) that fail to restore either fertility or spermatid individualization.

Interestingly, even among orthologs that appear non-functional in our fertility assays, all expressible Gdrd orthologs appear to maintain some association with the axoneme and the insect ring centriole in late elongating spermatids (Figures 4 & 6), indicative of conserved protein-protein interactions across the genus. Of the two subcellular localizations, the ability of orthologous Gdrd to interact with the axoneme shows greater variability, raising the possibility that the protein-protein interactions that mediate insect ring centriole localization may be either stronger, more stable, and/or more conserved. This observation is also suggestive of the possibility that Gdrd might have evolved first as an insect ring centriole constituent. On the other hand, the ability of an ortholog to associate with axonemes often correlates with its ability to restore fertility to *D. melanogaster gdrd* mutants (Figures 5 & 6). While axonemal localization is robust for the *D. simulans*, *D. yakuba*, and *D. mojavensis* orthologs, the non-rescuing *D. virilis* and *D. ananassae* orthologs exhibit weakened associations with axonemes. These observations suggest that, in *D. melanogaster* (and potentially other species), the Gdrd protein's ability to localize to the axoneme may mediate its essential function in fertility.

This conservation of protein-protein interactions at the insect ring centriole and axonemes is consistent with the overall conservation of protein structure and organization amongst all tested orthologs across the *Drosophila* genus. Our structure-function experiments indicate that the central helix mediates the protein's localization to both of these structures (Figure 3). Thus, this region of the protein most likely governs at least one protein-protein interaction. Indeed, multiple sequence alignment of all six tested orthologs reveals the presence of three conserved amino acid residues within the central helix (Figure 1B). However, as our structure-function assays also indicate the importance of the protein termini for function, the protein's IDRs most likely provide additional interaction points with other proteins.

The requirement of terminal IDRs might also explain the ability of *D. mojavensis gdrd* to restore fertility. Although *D. mojavensis* Gdrd has a longer helical region and is evolutionarily more divergent compared to *D. ananassae*, the latter cannot restore *D. melanogaster* fertility. The IDRs of *D. mojavensis* Gdrd bear more resemblance to the IDRs of *Sophophora* species than the IDRs of the *Drosophila* subgenus, according to k-mer based analysis (Figure 1). The *Sophophora* and *D. mojavensis* IDRs contain similar motifs and charges, which likely enable interactions analogous to those seen in *D. melanogaster* Gdrd (Figure 1). Interestingly, *D.*

*mojavensis* Gdrd also exhibits greater stability in MD simulations than *D. ananassae* Gdrd, which might partially explain why only *mojavensis* can rescue fertility (Figure 8). Conversely, the reduced stability of *D. ananassae* Gdrd could reflect greater flexibility, potentially allowing interactions with alternative molecular partners, while also leading to the inability of *D. ananassae gdrd* to restore fertility in *D. melanogaster*.

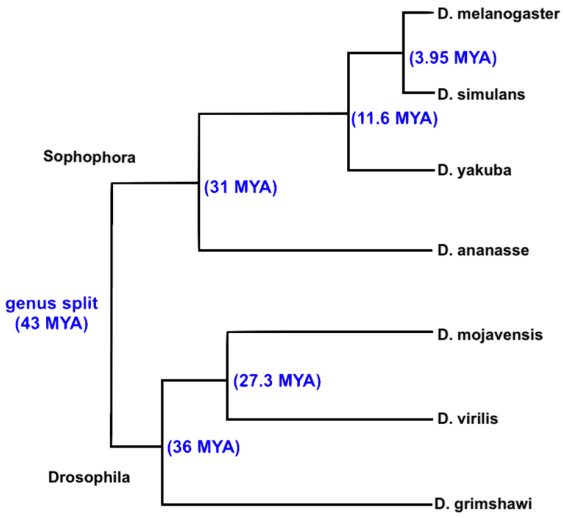
Besides localization to axonemes and insect ring centrioles, non-*melanogaster* Gdrd orthologs often exhibit additional subcellular localizations when exogenously expressed in *D. melanogaster* testes (Figures 6 & 7). These altered localization patterns could result from either evolutionary changes or innovations in their respective *Drosophila* species or protein mislocalizations caused by diminished interactions with the axonemal/ transition zone binding partner. Unsurprisingly, related orthologs often have similarly divergent localizations, presumably due to similarities in their amino acid sequences. The related *D. simulans* and *D. yakuba* proteins both localize to the basal end of elongated spermatid nuclei in late canoe spermatid cysts, while the related *D. mojavensis* and *D. virilis* orthologs associate with the nuclear cap in both early and late round spermatids. Furthermore, we find that these divergent localizations fall into two patterns. In the first pattern, orthologs appear to associate with other structures when axonemal structures have not yet formed. For example, *D. yakuba* Gdrd decorates a round organellar structure in onion stage spermatids. This pattern, however, is lost in leaf blade staged spermatids, when axoneme assembly begins. We observe a similar pattern for the *D. simulans* ortholog, which transiently associates with the nebenkern during the onion stage. Likewise, divergent localization of the *D. simulans* and *D. yakuba* orthologs to the basal end of nuclei only occurs in late canoe stage orthologs, when Gdrd's axonemal localization becomes diminished. These observations suggest that these localization patterns are most likely the result of weak, promiscuous interactions between the orthologous Gdrd proteins and these structures, suggestive of mislocalizations. However, these promiscuous binding/ mislocalization patterns may have consequences. In the case of the *D. yakuba* ortholog gene swap, localization to round organellar structures is restored in older elongating cysts. Therefore, the diminished ability of *D. yakuba* ortholog to restore fertility to *gdrd* mutants could potentially reflect subcellular binding competition for Gdrd between the axoneme and the organelle. On the other hand, divergent localization patterns observed in the *D. mojavensis, D. virilis*, and *D. ananassae* gene swaps may reflect the evolution of distinct binding capacities of these proteins. In the case of *D. mojavensis* and *D. virilis* orthologs, localization to the nuclear cap and the axoneme are similarly robust when compared to localization to the axoneme. By contrast, the *D. ananassae* protein localizes weakly to the axoneme but robustly to acroblast structures, suggesting that protein may have potentially evolved different preferred interaction partners.

Beyond displaying a preferential, divergent interaction with the acroblast, exogenous expression of *D. ananassae* Gdrd also causes attenuated spermatid elongation that enhances the *gdrd* null loss of function phenotype. As we do not observe any dominant effects on fertility, the observed attenuation is most likely weak, which might be expected for orthologous genes evolving under selective constraints that abrogate toxicity [69–71]. At present, the source of this attenuation is unknown. One possibility is that the protein's divergent localization to the acroblast mildly interferes with cellular functions at that organelle. As discussed above, this divergent

localization and interaction could be facilitated by the more flexible structure of *D. ananassae* Gdrd.

While previous work has focused on the conservation of protein structure in *de novo* evolved orphan genes following their originations [16], this study provides evidence that function can be just as old. Our data, however, also indicate that this ancestral protein has most likely diverged in ways that altered the intrinsic properties of orthologs within a subset of descendant lineages. MD simulations of two non-rescuing orthologs, *D. ananassae* and *D. virilis*, reveal that these proteins appear less structurally stable than other tested orthologs (Figure 8). This might account for the diminished axoneme localization for both orthologs. In the case of the *D. virilis* ortholog, the protein also appears more cytoplasmic, suggesting that the protein's intrinsic structural instability may diminish its binding capacity with its interacting partner. As structural instability can also be indicative of poor folding, reduced protein stability, and aggregation propensity, our MD simulations of the *D. ananassae* ortholog can potentially also explain the ortholog's diminished expression and attenuation in our gene swap assays. Our k-mer based analyses of IDRs also indicate that motifs that were most likely present in the *Sophophora* and *Drosophila* subgenera common ancestor have been lost in various lineages (Figure 1). Collectively, our data suggest that Gdrd was likely fully functional in the common ancestor, but that its intrinsic biophysical properties have been maintained in some descendant lineages and modified or lost in others (Figure 9).

Indeed, this pattern may be applicable to most orphan genes that both undergo rapid sequence changes and contain IDRs. The gain and loss of *de novo* emerged orphan genes have been described as a continuum [11,21,72]. Similarly, their acquisition of function or activities likely follows a similar continuum, driven by the binding promiscuity or polyvalency of the IDRs they often contain. This promiscuity, enabled by structural flexibility, facilitates co-evolution with diverse interaction partners over different evolutionary lineages and cellular environments, allowing for rapid integration into cellular networks [38,73]. This may also explain why Gdrd proteins from the *Drosophila* subgroup, except for *D. mojavensis*, which shares a similar motif and charge (Figure 1), fail to restore fertility despite preserving the overall structure. These orthologs could have co-evolved with different partners, leading to divergent IDR patterns and structural stability (Figure 9).




| Structural stability compared to *D. melanogaster* | Terminal IDR motif conservation | Fertility | Axoneme localization | Divergent localization |
|---|---|---|---|---|
| NA | conserved | full rescue | strong | None |
| similar | conserved | full rescue | strong | mitochondria |
| similar | conserved | partial rescue | strong | unknown vesicular organelle |
| weak | conserved | sterile | weak | acroblast/ golgi |
| similar | conserved | full rescue | strong | nuclear membrane |
| weak | not conserved | sterile | weak | nuclear membrane |
| similar | not conserved | sterile | undetectable | plasma membrane |

**Figure 9: Summary of structural, k-mer, and gene swap analyses.** Orthologs that lack function in gene swap assays typically exhibit weaker structural stability in MD simulations, have lost conserved terminal IDR motifs, or both. Most orthologs also exhibited divergent localization patterns. Divergence times (in blue brackets and font) were estimated using TimeTree. NA = Not Applicable.

Our gene swap experiments suggest that amino acid sequence divergence in orphan genes can be consequential, as exogenously expressed orthologs often exhibit either diminished or lost function as well as altered subcellular localizations. However, this approach also has limitations. First, we analyzed the orthologs within a functional rescue construct, which often produces lower expression levels when compared to the endogenous gene locus. Although the construct containing the *melanogaster* ortholog expressed at high enough levels to restore fertility at levels equivalent to flies expressing two endogenous copies (Figures 3 and 5), and most of the other orthologs appeared to be expressed at levels that were similar to or higher than the rescuing *mojavensis* ortholog, future strategies might include CRISPR-Cas9 genome editing to insert orthologous genes directly into the endogenous genomic locus [50,51,74]. This strategy could potentially discern whether orthologs such as *D. virilis* and *D. ananassae* are completely incapable of restoring fertility in *D. melanogaster*, or whether they can muster some degree of rescue when expressed more highly. On the other hand, higher exogenous expression of the *D. ananassae gdrd* could lead to greater attenuation of spermatid elongation, which could hinder our ability to perform the gene swap in the first place. A second limitation to our methodology is that gene swaps constitute exogenous expression. While testes may be physiologically similar between closely related species, they also differ in important ways [75–77]. Future work to confirm these findings would require raising antibodies to orthologous Gdrd proteins to confirm divergent subcellular localizations within testes of divergent species (a process that has, to date, been challenging for the *D. melanogaster* ortholog) and/or generating knockouts and HA-tagged knock-ins in non-*melanogaster* species [78–81].

In conclusion, this study indicates that Gdrd likely evolved in the common ancestor of the *Sophophora* and *Drosophila* subgenera the core interactions and/or capabilities that render it essential for spermatogenesis in present day *D. melanogaster*. Based on the strong conservation of transition zone localization and the observed correlations between function and axonemal localization, this work highlights the importance of Gdrd localization to these structures. However, our gene swap analyses also raise the possibility that Gdrd function has not remained static in all lineages. The orthologs from two species within the *Drosophila* subgenus (*D. mojavensis* and *D. virilis*) have distinct and expanded localization patterns when examined in *D. melanogaster*, suggesting that Gdrd might have an expanded subcellular distribution in these lineages, while a fairly conserved ortholog within the Sophophora subgenus (*D. ananassae*) might have evolved new preferred interactions. Future experiments will address if these altered subcellular localizations reflect expanded or novel gene functions.

## Materials and Methods

### Fly stocks and husbandry

For a complete list of fly strains used in this study, please see the Reagents Table. For all experiments, flies were raised on standard cornmeal molasses media [82] at 25°C.

### Sequence alignment and truncated *D. melanogaster* Gdrd design

Based on our earlier study, which included concordant experimental results from circular dichroism spectroscopy and 2D NMR spectroscopy [44], we subdivided *D. melanogaster* Gdrd protein into three regions: a 40 amino acid (aa) disordered N-terminus, a 38 aa central helix, and a 35 aa disordered C-terminus. Homology analyses between the *D. melanogaster* protein and *Drosophila* genus orthologs were performed using the pairwise BLASTP algorithm using either the full length or the aforementioned regional *D. melanogaster* amino acid sequences as a query. As distant orthologs within the *Drosophila* subgenus were highly divergent at the N- and C- termini of the protein, amino acid sequence conservation (percent identity) was alternatively determined within the context of a Clustal Omega (ClustalO) multiple sequence alignment [47] and analyzed using MEGA 11 [83]. Truncated versions of *D. melanogaster* Gdrd were designed based on AlphaFold2 [84] predictions, experimental structure data [44], and alignment of the central helix (Figure 1B). The N-terminal truncation (Gdrd ΔN) removes 36 amino acids following the initial methionine. The C-terminal truncation (Gdrd ΔC) removes the final 35 amino acids of the protein, immediately downstream of the central helix. We used the DeNoFo toolkit [85] to annotate the methods used to identify Gdrd in earlier studies [43,44].

### Alignment-free analysis of terminal IDRs

Based on the sequence alignment of the central helices (Figure 1B), sequences of the disordered termini were extracted from each ortholog. Homology of each ortholog's N- and C-termini to those of *D. melanogaster* Gdrd was calculated using the alignment-free, k-mer based tool SHARK-Dive [40]. Similarity was normalized with D. *melanogaster* Gdrd self-comparison set as 1. Shared motifs were identified using SHARK-Capture in default mode

[48]. Overall charge of tails was calculated using a custom python script (D, E = -1, K, R, H = +1, other amino acids = 0).

**Transgenic flies**

All transgenic constructs used in this analysis constitute modifications of the previously reported *gdrd:HA* rescue [44]. For ortholog gene swaps, the *D. melanogaster*, *D. simulans*, *D. yakuba*, *D. ananassae*, *D. mojavensis*, *D. virilis*, and *D. grimshawi gdrd* coding sequences were first codon-optimized using GENEius software (Eurofins MWG Operon). Forty nucleotide upstream and downstream homology arm sequences (corresponding to the 5'UTR and HA-tag sequences in the original rescue, respectively) were then attached to each codon-optimized gene sequence prior to synthesis (Eurofins MWG Operon). The upstream regulatory regions containing the *D. melanogaster gdrd* 5'UTR and the downstream regulatory regions containing the Hemagluttanin (HA) tag were PCR amplified using Q5 High Fidelity Polymerase (NEB). These PCR fragments, along with the synthesized codon-optimized gene, were cloned into a XbaI/AscI-linearized w+attB plasmid (Sekelsky, Addgene plasmid 30326) using Gibson Assembly (NEB). Gdrd ΔN and Gdrd ΔC constructs are modifications of the codon-optimized *mel gdrd* rescue construct. Sequences that omit these protein regions were amplified using Q5 High Fidelity Polymerase (NEB) and extended primers designed with 20 bp homology regions. These DNA fragments were then cloned into w+attB plasmid using Gibson assembly. To control for chromatin position effects on expression, each rescue construct was then phiC31 integrated into the *PBac{y$^+$-attP-9A}VK00020* (BL24867) docking site (Rainbow Transgenics). Upon establishing transgenic lines in the *D. melanogaster gdrd* null background, we used PCR and sequencing to re-confirm ortholog insertions and the absence of the endogenous gene. To maximize gene expression from transgenes, all tested males carry two copies of each rescue insertion, unless otherwise noted. See the Reagents Table for primer and synthesized DNA sequences.

**Fertility assay**

To assess the functional abilities of *gdrd* orthologs and truncations in *D. melanogaster*, we performed single male fertility assays by mating individual males with two *Canton-S* virgin females. All males were collected as virgins and then aged in uncrowded conditions for four to six days prior to mating. The age range of virgin females fell between four to eight days. Likewise, virgin females used in the assay were isolated in uncrowded condition and reared on a high yeast diet prior to mating to increase fecundity. Mating and egg laying proceeded for 48 hours post-crossing before parents were discarded. Progeny number was determined by counting the number of pupal cases on the side of each vial 10 days (~216 hours) after setting the cross. Final sample sizes for each genotype were $n$ = 25-30 males, depending on the experiment. Progeny count data were analyzed in Microsoft Excel using *t*-tests with unequal variances.

**Fluorescent labeling and confocal microscopy of *Drosophila* whole testes and testes squashes**

Antibody labeling of whole testis and testis squashes were performed as previously described in Lange *et al*. [44] and Sitaram *et al*. [86], respectively. In brief, processed tissues were fixed following dissection in 4% paraformaldehyde/ 1X phosphate buffered saline (PBS) for 20

minutes. After washing the fixative away, tissues were blocked for 1-2 hours in 1X PBS containing 3% bovine serum albumin, 5% normal goat serum, and 0.1% Triton-X. The samples were then incubated overnight in rabbit anti:HA (C29F4, Cell Signaling Technologies) diluted at 1:100 in blocking solution. The following day, the antibody was washed off before applying goat anti-rabbit Alexa Fluor 488 conjugated secondary (A-11008, ThermoFisher Scientific) diluted at 1:200 in blocking solution. The secondary was then washed away prior to mounting samples in Vectashield Plus Antifade Mounting Medium Plus DAPI (Vector labs). To analyze individualization complexes during spermatogenesis, fixed whole testes were incubated overnight with TRITC conjugated phalloidin (R415, Life Technologies/ Molecular Probes) diluted at 1:200 in PBS containing 0.1% Triton-X. Microscopy was performed on an SP8 X confocal microscope (Leica Microsystems) using HC PL APO CS2 20x/0.75 ILL and HC PL APO CS2 63x/1.40 oil objectives. DAPI, Alexa Fluor 488, and TRITC were excited using 405, 488, and 546 wavelengths, respectively. Post-acquisition processing was performed using ImageJ Fiji (version 1.0) [87].

**Structure predictions and MD simulations**

Structure predictions were performed with AlphaFold2 (v.2.3.1; Database cut-off date: 2022-12-19) [84]. MD simulations were performed using GROMACS 2021.2 [88] on AlphaFold2 predictions with highest mean predicted local distance difference test (pLDDT) score. Structures were prepared using OPLS-AA/L force field [89] and solvent SOL. The structure was solvated in a cubic box of SPC/E water with 10-Å clearance and the electrostatic charge neutralized by the addition of sodium atoms, followed by energy minimization and equilibration. Triplicate 250ns simulations were run in an NPT ensemble using a V-rescale modified Berendsen thermostat at a temperature of 300 K and a Parrinello-Rahman barostat at a pressure of 1 atm, periodic boundary conditions, and a particle mesh Ewald summation with a grid spacing of 1.6 Å and fourth order interpolation. Custom python and bash scripts were used for analysis of MD simulations (Backbone RMSD, $\alpha$-C RMSF, radius of gyration). Python (v.3.10.13) libraries numpy [90] and matplotlib [91] were used for plotting. For each species, triplicate 250ns MD simulation snapshots were aligned using PyMOL v3.0.5 [92]**.** Structures were visualized using PyMOL v3.0.5 and ChimeraX-1.7.1 [93].

**Acknowledgements**

We would like to thank Dr. Tomer Avidor-Reiss (Department of Biological Sciences, University of Toledo, Toledo, OH) for providing us with the *unc:EGFP* transgenic fly strain and Dr. Emily Rivard for providing feedback on the manuscript. P.H.P, K.L.M and G.D.F were supported by NSF RUI grant (2212972) awarded to G.D.F. E.B.B. and A.L. received funding from Volkswagen foundation grant code 98183. E. B. B. was supported by HFSP (Human Frontiers of Science Programme, RGP0006/2013 and RGP004/2023) and the DFG (Deutsche Forschungsgemeinschaft BO-2544/20-1;503272152). L.A.E. has been supported by EMBO Scientific Exchange Grant 10944. Structure predictions and MD simulations were performed on the HPC cluster PALMA II of the University of Muenster, subsidised by the DFG (INST 211/667-1).

**Supplementary materials**

Scripts for sequence and structural analysis are deposited on GitHub: https://github.com/ArsLeicholt/gdrd_analysis. Supplementary data can be found on Zenodo: https://zenodo.org/records/15173277

**Supplemental Figures**

**Pairwise BLASTP analysis of *melanogaster* Gdrd and orthologs**

| | Full protein BLASTP | N-terminus BLASTP | Central Helix BLASTP | C-terminus BLASTP |
|---|---|---|---|---|
| ***D. simulans*** | 108/113 (96%) | 40/40 (100%) | 35/38 (92%) | 33/35 (94%) |
| ***D. yakuba*** | 86/113 (76%) | 34/40 (85%) | 29/38 (76%) | 23/35 (66%) |
| ***D. ananassae*** | 46/101 (46%) | 20/40 (50%) | 18/38 (47%) | 7/11 (64%) |
| ***D. mojavensis*** | 20/61 (33%) | NHD | 9/29 (31%) | NHD |
| ***D. virilis*** | 10/26 (38%) | NHD | 10/26 (38%) | NHD |
| ***D. grimshawi*** | NHD | NHD | NHD | NHD |

**Figure S1: Pairwise BLASTP analysis of Gdrd orthologs.** Either the full length, N-terminus, central helix, or C-terminus *D. melanogaster* Gdrd protein sequences were used as Query in pairwise BLAST analyses with orthologous Gdrd proteins present throughout the *Drosophila* genus. BLASTP fails to detect any homology between the *D. melanogaster* and *D. grimshawi* proteins. NHD = No Homology Detected.

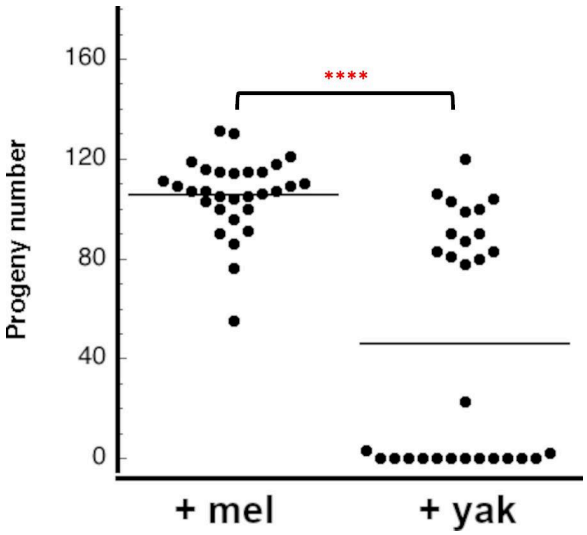



**Figure S2: Replicate fertility assay of *D. yakuba* gene swap males.** *D. yakuba* gene swap males exhibit bimodal fertility, with some males exhibiting near wild-type levels of fertility while other males are either sterile or near sterile. Statistical test: *t*-tests with unequal variance. **** = $p < .0001$

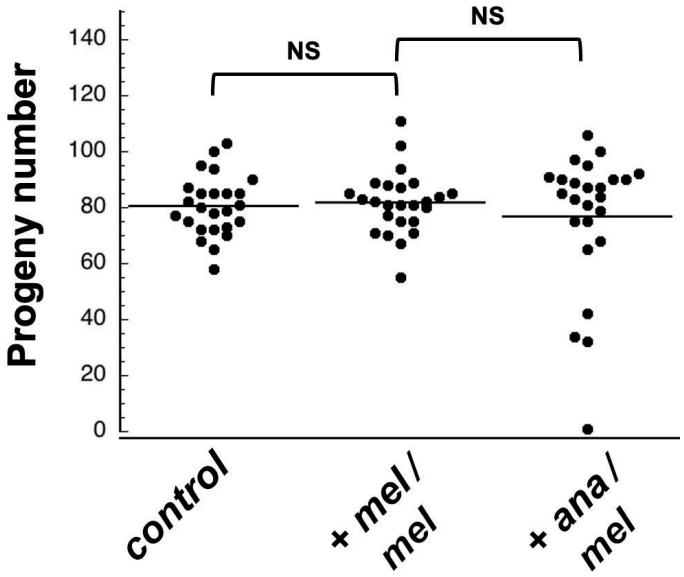



**Figure S3: Spermatid elongation attenuation phenotypes in *D. ananassae* gene swaps are recessive.** Transheterozygous male flies carrying one *D. ananassae gdrd* and one *D. melanogaster gdrd* rescue construct in a *gdrd* null background have wild type levels of fertility. Statistical test: *t*-tests with unequal variance. NS = Not Significant.

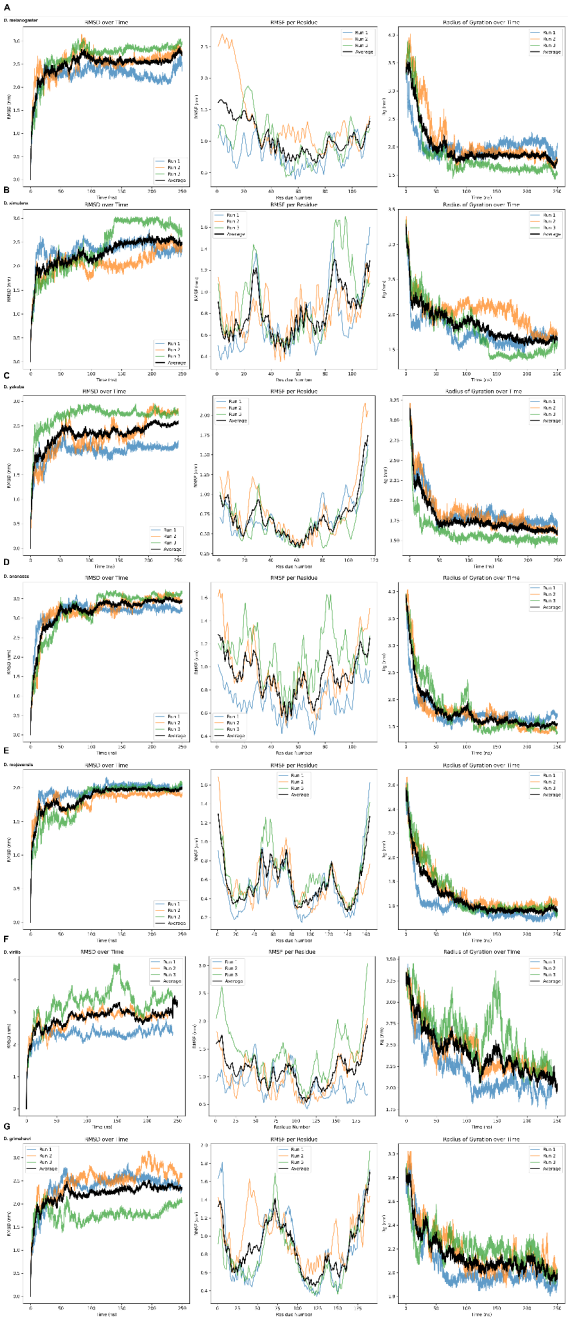


**Figure S4: Triplicate MD runs of gdrd orthologs backbone RMSD, $\alpha$-C RMSF and radius of gyration (left to right) for all triplicate MD runs of gdrd orthologs and the average of those triplicates.** (A) *D. melanogaster* (B) *D. simulans* (C) *D. yakuba* (D) *D. ananasse* (E) *D. mojavensis* (F) *D. virilis* (G) *D. grimshawi.*

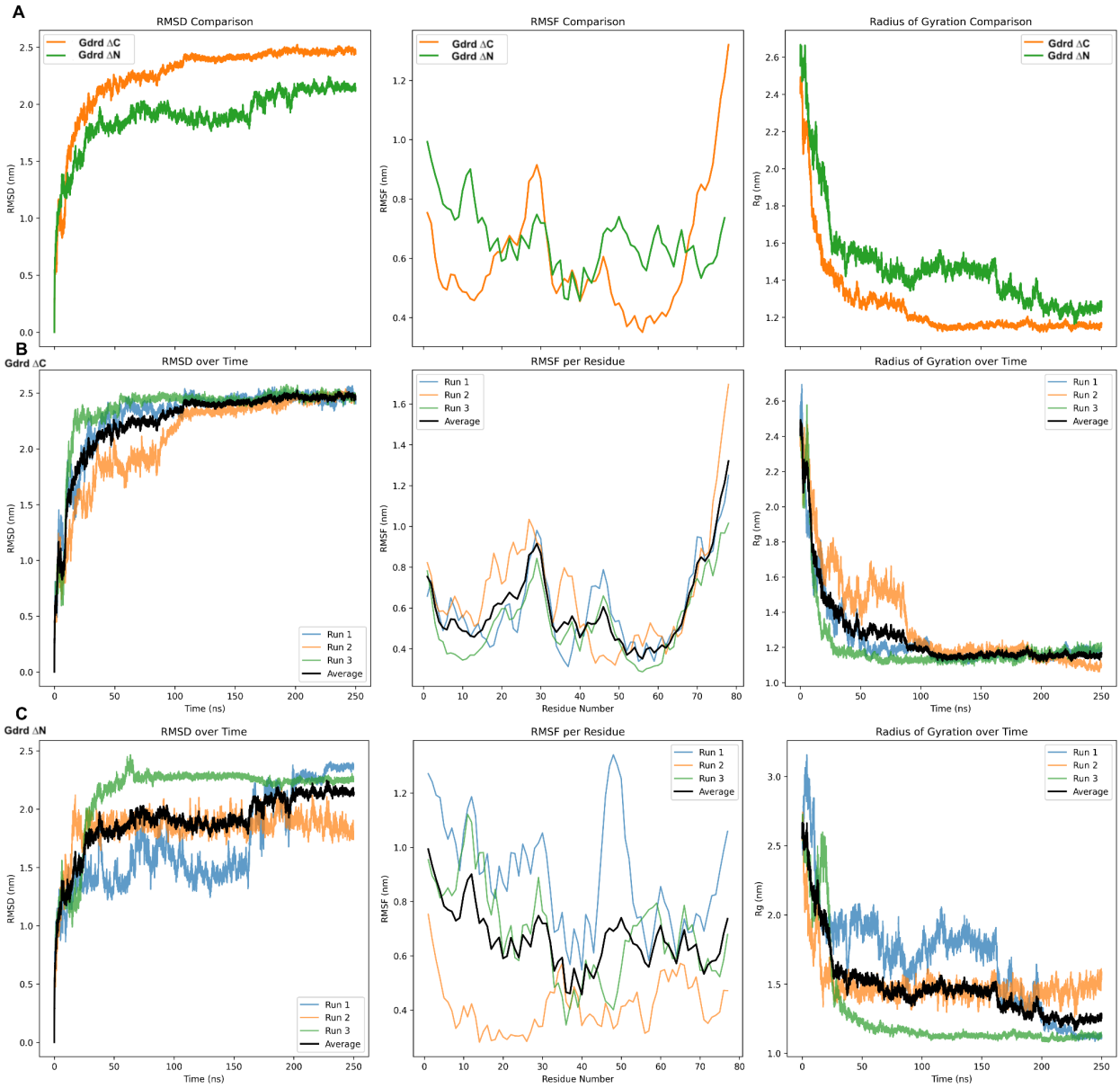



**Figure S5: Triplicate MD runs of *D. melanogaster* gdrd truncations backbone RMSD, $\alpha$-C RMSF and radius of gyration (left to right) for all triplicate MD runs of Gdrd ΔC (B), Gdrd ΔN (C) and the average of those triplicates (A).**

**References**


1. Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12: 692–702.

2. Pereira AB, Marano M, Bathala R, Zaragoza RA, Neira A, Samano A, et al. Orphan genes are not a distinct biological entity. Bioessays. 2025;47: e2400146.

3. Zhao L, Svetec N, Begun DJ. De Novo Genes. Annu Rev Genet. 2024;58: 211–232.

4. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TCG. More than just orphans: are taxonomically-restricted genes important in evolution? Trends Genet. 2009;25: 404–413.

5. Prabh N, Rödelsperger C. Are orphan genes protein-coding, prediction artifacts, or non-coding RNAs? BMC Bioinformatics. 2016;17: 226.

6. Middendorf L, Ravi Iyengar B, Eicholt LA. Sequence, Structure, and Functional Space of Drosophila De Novo Proteins. Genome Biol Evol. 2024;16. doi:10.1093/gbe/evae176

7. Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D. Orphans as taxonomically restricted and ecologically important genes. Microbiology (Reading). 2005;151: 2499–2501.

8. Fakhar AZ, Liu J, Pajerowska-Mukhtar KM, Mukhtar MS. The Lost and Found: Unraveling the Functions of Orphan Genes. J Dev Biol. 2023;11. doi:10.3390/jdb11020027

9. Schmid KJ, Aquadro CF. The evolutionary analysis of “orphans” from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics. 2001;159: 589–598.

10. Palmieri N, Kosiol C, Schlötterer C. The life cycle of Drosophila orphan genes. Elife. 2014;3: e01311.

11. Iyengar BR, Bornberg-Bauer E. Neutral Models of De Novo Gene Emergence Suggest that Gene Evolution has a Preferred Trajectory. Mol Biol Evol. 2023;40. doi:10.1093/molbev/msad079

12. Grandchamp A, Czuppon P, Bornberg-Bauer E. Quantification and modeling of turnover dynamics of de novo transcripts in Drosophila melanogaster. Nucleic Acids Res. 2024;52: 274–287.

13. Lebherz MK, Fouks B, Schmidt J, Bornberg-Bauer E, Grandchamp A. DNA Transposons Favor De Novo Transcript Emergence Through Enrichment of Transcription Factor Binding Motifs. Genome Biol Evol. 2024;16. doi:10.1093/gbe/evae134

14. Domazet-Loso T, Tautz D. An evolutionary analysis of orphan genes in Drosophila. Genome Res. 2003;13: 2213–2219.

15. Wissler L, Gadau J, Simola DF, Helmkampf M, Bornberg-Bauer E. Mechanisms and dynamics of orphan gene emergence in insect genomes. Genome Biol Evol. 2013;5: 439–455.

16. Peng J, Zhao L. The origin and structural evolution of de novo genes in Drosophila. Nat Commun. 2024;15: 810.

17. Vakirlis N, Carvunis A-R, McLysaght A. Synteny-based analyses indicate that sequence divergence is not the main source of orphan genes. Elife. 2020;9. doi:10.7554/eLife.53500

18. Anna Grandchamp, Margaux Aubel, Lars A Eicholt, Paul Roginski, Victor Luria, Amir Karger, Elias Dohmen. De Novo Gene Emergence: Summary, Classification, and Challenges of Current Methods. ecoevorxiv. 2025. doi:10.32942/X2DP88

19. Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, et al. Origin of primate orphan genes: a comparative genomics approach. Mol Biol Evol. 2009;26: 603–612.

20. Zhang W, Gao Y, Long M, Shen B. Origination and evolution of orphan genes and de novo genes in the genome of Caenorhabditis elegans. Sci China Life Sci. 2019;62: 579–593.

21. 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: 382–398.

22. Schmid KJ, Tautz D. A screen for fast evolving genes from Drosophila. Proc Natl Acad Sci U S A. 1997;94: 9746–9750.

23. Moutinho AF, Trancoso FF, Dutheil JY. The Impact of Protein Architecture on Adaptive Evolution. Mol Biol Evol. 2019;36: 2013–2028.

24. Bricout R, Weil D, Stroebel D, Genovesio A, Roest Crollius H. Evolution is not Uniform Along Coding Sequences. Mol Biol Evol. 2023;40. doi:10.1093/molbev/msad042

25. Goldman N, Thorne JL, Jones DT. Assessing the impact of secondary structure and solvent accessibility on protein evolution. Genetics. 1998;149: 445–458.

26. Guo HH, Choe J, Loeb LA. Protein tolerance to random amino acid change. Proc Natl Acad Sci U S A. 2004;101: 9205–9210.

27. Choi SS, Vallender EJ, Lahn BT. Systematically assessing the influence of 3-dimensional structural context on the molecular evolution of mammalian proteomes. Mol Biol Evol. 2006;23: 2131–2133.

28. Ángyán AF, Perczel A, Gáspári Z. Estimating intrinsic structural preferences of de novo emerging random-sequence proteins: is aggregation the main bottleneck? FEBS Lett. 2012;586: 2468–2472.

29. Landry CR, Zhong X, Nielly-Thibault L, Roucou X. Found in translation: functions and evolution of a recently discovered alternative proteome. Curr Opin Struct Biol. 2015;32: 74–80.

30. Basile W, Sachenkova O, Light S, Elofsson A. High GC content causes orphan proteins to be intrinsically disordered. PLoS Comput Biol. 2017;13: e1005375.

31. Eicholt LA, Aubel M, Berk K, Bornberg-Bauer E, Lange A. Heterologous expression of naturally evolved putative de novo proteins with chaperones. Protein Sci. 2022;31: e4371.

32. Heames B, Buchel F, Aubel M, Tretyachenko V, Loginov D, Novák P, et al. Experimental characterization of de novo proteins and their unevolved random-sequence counterparts. Nat Ecol Evol. 2023;7: 570–580.

33. Middendorf L, Eicholt LA. Random, de novo, and conserved proteins: How structure and disorder predictors perform differently. Proteins. 2024;92: 757–767.

34. Aubel M, Buchel F, Heames B, Jones A, Honc O, Bornberg-Bauer E, et al. High-throughput Selection of Human de novo-emerged sORFs with High Folding Potential. Genome Biol Evol. 2024;16. doi:10.1093/gbe/evae069

35. Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. Evolution and disorder. Curr Opin Struct Biol. 2011;21: 441–446.

36. Uversky VN. Paradoxes and wonders of intrinsic disorder: Stability of instability. Intrinsically Disord Proteins. 2017;5: e1327757.

37. LeBlanc C, Stefani J, Soriano M, Lam A, Zintel MA, Kotha SR, et al. Conservation of function without conservation of amino acid sequence in intrinsically disordered transcriptional activation domains. bioRxiv. 2024. doi:10.1101/2024.12.03.626510

38. Jemth P. Protein binding and folding through an evolutionary lens. Curr Opin Struct Biol. 2025;90: 102980.

39. Zarin T, Strome B, Peng G, Pritišanac I, Forman-Kay JD, Moses AM. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. Elife. 2021;10. doi:10.7554/eLife.60220

40. Chow CFW, Ghosh S, Hadarovich A, Toth-Petroczy A. SHARK enables sensitive detection of evolutionary homologs and functional analogs in unalignable and disordered sequences. Proc Natl Acad Sci U S A. 2024;121: e2401622121.

41. 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: 781.

42. Chen J, Li Q, Xia S, Arsala D, Sosa D, Wang D, et al. The Rapid Evolution of De Novo Proteins in Structure and Complex. Genome Biol Evol. 2024;16. doi:10.1093/gbe/evae107

43. Gubala A, Schmitz JF, Kearns MJ, Vinh T, Bornberg-Bauer E, Wolfner M, et al. The Goddard and Saturn genes are essential for Drosophila male fertility and may have arisen DE Novo. Mol Biol Evol. 2017;34: 1066–1082.

44. 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.

45. Russo CA, Takezaki N, Nei M. Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol. 1995;12: 391–404.

46. Kumar S, Stecher G, Suleski M, Hedges SB. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. Mol Biol Evol. 2017;34: 1812–1819.

47. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7: 539.

48. Chow CFW, Lenz S, Scheremetjew M, Ghosh S, Richter D, Jegers C, et al. SHARK-capture identifies functional motifs in intrinsically disordered protein regions. Protein Sci. 2025;34: e70091.

49. Bayes JJ, Malik HS. Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. Science. 2009;326: 1538–1541.

50. Saint-Leandre B, Christopher C, Levine MT. Adaptive evolution of an essential telomere protein restricts telomeric retrotransposons. Elife. 2020;9. doi:10.7554/eLife.60987

51. Brand CL, Levine MT. Cross-species incompatibility between a DNA satellite and the Drosophila Spartan homolog poisons germline genome integrity. Curr Biol. 2022;32: 2962–2971.e4.

52. Baker JD, Adhikarakunnathu S, Kernan MJ. Mechanosensory-defective, male-sterile unc mutants identify a novel basal body protein required for ciliogenesis in Drosophila. Development. 2004;131: 3411–3422.

53. Riparbelli MG, Callaini G, Megraw TL. Assembly and persistence of primary cilia in dividing Drosophila spermatocytes. Dev Cell. 2012;23: 425–432.

54. Basiri ML, Ha A, Chadha A, Clark NM, Polyanovsky A, Cook B, et al. A migrating ciliary gate compartmentalizes the site of axoneme assembly in Drosophila spermatids. Curr Biol. 2014;24: 2622–2631.

55. Pratt MB, Titlow JS, Davis I, Barker AR, Dawe HR, Raff JW, et al. Drosophila sensory cilia lacking MKS proteins exhibit striking defects in development but only subtle defects in adults. J Cell Sci. 2016;129: 3732–3743.

56. Vieillard J, Paschaki M, Duteyrat J-L, Augière C, Cortier E, Lapart J-A, et al. Transition zone assembly and its contribution to axoneme formation in Drosophila male germ cells. J Cell Biol. 2016;214: 875–889.

57. Johnson HH. PERIPHERAL MIGRATION OF A CENTRIOLE DERIVATIVE IN THE SPERMATOGENESIS OF OEligCANTHUS. Science. 1922;56: 759–760.

58. Phillips DM. Insect sperm: their structure and morphogenesis. J Cell Biol. 1970;44: 243–277.

59. Kwitny S, Klaus AV, Hunnicutt GR. The annulus of the mouse sperm tail is required to establish a membrane diffusion barrier that is engaged during the late steps of spermiogenesis. Biol Reprod. 2010;82: 669–678.

60. Hoque M, Li F-Q, Weber WD, Chen JJ, Kim EN, Kuo P-L, et al. The Cby3/ciBAR1 complex positions the annulus along the sperm flagellum during spermiogenesis. J Cell Biol. 2024;223. doi:10.1083/jcb.202307147

61. Tretyachenko V, Vymětal J, Bednárová L, Kopecký V Jr, Hofbauerová K, Jindrová H, et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. Sci Rep. 2017;7: 15449.

62. Borcherds W, Bremer A, Borgia MB, Mittag T. How do intrinsically disordered protein regions encode a driving force for liquid-liquid phase separation? Curr Opin Struct Biol.

2021;67: 41–50.

63. Liu J, Yuan R, Shao W, Wang J, Silman I, Sussman JL. Do “Newly Born” orphan proteins resemble “Never Born” proteins? A study using three deep learning algorithms. Proteins. 2023;91: 1097–1115.

64. Aubel M, Eicholt L, Bornberg-Bauer E. Assessing structure and disorder prediction tools for emerged proteins in the age of machine learning. F1000Res. 2023;12: 347.

65. Seffernick JT, Ren H, Kim SS, Lindert S. Measuring Intrinsic Disorder and Tracking Conformational Transitions Using Rosetta ResidueDisorder. J Phys Chem B. 2019;123: 7103–7112.

66. Giménez-Andrés M, Čopič A, Antonny B. The Many Faces of Amphipathic Helices. Biomolecules. 2018;8. doi:10.3390/biom8030045

67. 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: 1626–1632.

68. Dowling D, Schmitz JF, Bornberg-Bauer E. Stochastic Gain and Loss of Novel Transcribed Open Reading Frames in the Human Lineage. Genome Biol Evol. 2020;12: 2183–2195.

69. Masel J. Cryptic genetic variation is enriched for potential adaptations. Genetics. 2006;172: 1985–1991.

70. Rajon E, Masel J. Evolution of molecular error rates and the consequences for evolvability. Proc Natl Acad Sci U S A. 2011;108: 1082–1087.

71. Wilson BA, Masel J. Putatively noncoding transcripts show extensive association with ribosomes. Genome Biol Evol. 2011;3: 1245–1252.

72. 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: 370–374.

73. Bornberg-Bauer E, Hlouchova K, Lange A. Structure and function of naturally evolved de novo proteins. Curr Opin Struct Biol. 2021;68: 175–183.

74. Adikusuma F, Pederick D, McAninch D, Hughes J, Thomas P. Functional Equivalence of the SOX2 and SOX3 Transcription Factors in the Developing Mouse Brain and Testes. Genetics. 2017;206: 1495–1503.

75. Schärer L, Da Lage J-L, Joly D. Evolution of testicular architecture in the Drosophilidae: a role for sperm length. BMC Evol Biol. 2008;8: 143.

76. Lüpold S, Manier MK, Puniamoorthy N, Schoff C, Starmer WT, Luepold SHB, et al. How sexual selection can drive the evolution of costly sperm ornamentation. Nature. 2016;533: 535–538.

77. Alpern JHM, Asselin MM, Moehring AJ. Identification of a novel sperm class and its role in fertilization in Drosophila. J Evol Biol. 2019;32: 259–266.

78. Lamb AM, Wang Z, Simmer P, Chung H, Wittkopp PJ. affects pigmentation divergence and cuticular hydrocarbons in and. Front Ecol Evol. 2020;8. doi:10.3389/fevo.2020.00184

79. Bubnell JE, Ulbing CKS, Fernandez Begne P, Aquadro CF. Functional Divergence of the bag-of-marbles Gene in the Drosophila melanogaster Species Group. Mol Biol Evol. 2022;39. doi:10.1093/molbev/msac137

80. Farfán-Pira KJ, Martínez-Cuevas TI, Evans TA, Nahmad M. A cis-regulatory sequence of the selector gene vestigial drives the evolution of wing scaling in Drosophila species. J Exp Biol. 2023;226. doi:10.1242/jeb.244692

81. Baker CA, Guan X-J, Choi M, Murthy M. The role of in specifying courtship behaviors across divergent species. Sci Adv. 2024;10: eadk1273.

82. Guay SY, Patel PH, Thomalla JM, McDermott KL, O'Toole JM, Arnold SE, et al. An orphan gene is essential for efficient sperm entry into eggs in Drosophila melanogaster. Genetics. 2025;229. doi:10.1093/genetics/iyaf008

83. Tamura K, Stecher G, Kumar S. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. Mol Biol Evol. 2021;38: 3022–3027.

84. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596: 583–589.

85. Elias Dohmen, Margaux Aubel, Lars A. Eicholt, Paul Roginski, Victor Luria, Amir Karger, Anna Grandchamp. DeNoFo: a file format and toolkit for standardised, comparable de novo gene annotation. biorxiv. 2025. doi:10.1101/2025.03.31.644673

86. Sitaram P, Hainline SG, Lee LA. Cytological analysis of spermatogenesis: live and fixed preparations of Drosophila testes. J Vis Exp. 2014; e51058.

87. Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, et al. Fiji: an open-source platform for biological-image analysis. Nat Methods. 2012;9: 676–682.

88. Páll S, Zhmurov A, Bauer P, Abraham M, Lundborg M, Gray A, et al. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. J Chem Phys. 2020;153: 134110.

89. Shivakumar D, Williams J, Wu Y, Damm W, Shelley J, Sherman W. Prediction of Absolute Solvation Free Energies using Molecular Dynamics Free Energy Perturbation and the OPLS Force Field. J Chem Theory Comput. 2010;6: 1509–1519.

90. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585: 357–362.

91. Hunter JD. Matplotlib: A 2D Graphics Environment. Comput Sci Eng. 2007;9: 90–95.

92. Schrödinger LLC. PyMOL The PyMOL Molecular Graphics System, Version 1.8, Schrödinger, LLC. 2015.

93. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, et al. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. 2021;30: 70–82.