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

*Current address: Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853

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

Orphan genes evolve rapidly, raising questions about whether their functions remain conserved or diverge across species. To address this, we investigated *goddard* (*gdrd*), an orphan gene essential for spermatogenesis in *Drosophila melanogaster*. Within the *Drosophila* genus, Gdrd proteins retain a conserved core structure but display substantial variation in length and primary sequence. Here we perform cross-species gene-swap assays in *D. melanogaster* testes to examine how these lineage-specific changes affect Gdrd function. Strikingly, the highly divergent *D. mojavensis* ortholog fully rescues fertility in *gdrd* null flies, suggesting that ancestral Gdrd acted within a conserved spermatogenesis pathway. By contrast, several orthologs, including one from a more closely related species, cannot substitute for the *melanogaster* gene. Cytological analysis shows that all divergent Gdrd orthologs retain some ability to interact with axonemes and ring centrioles, consistent with the protein's structural conservation, but many non-complementing orthologs display weaker axonemal binding. Furthermore, all tested orthologs exhibit divergent localizations to organellar structures. Using computational analyses and molecular dynamics simulations, we identified intrinsic protein qualities that may account for several observations made in the gene swap assays. Rescuing orthologs bear motifs with shared physicochemical properties in their intrinsically disordered regions, while non-rescuing variants exhibit structural instabilities. Taken together, these findings show that while Gdrd's ancestral structure and interactions are conserved, several orthologs have undergone lineage-specific evolutionary changes.

INTRODUCTION

Orphan genes are phylogenetically restricted genes that encode proteins lacking detectable homology to known proteins (Tautz and Domazet-Lošo 2011; Zhao et al. 2024; Pereira et al. 2025). As a group, they comprise roughly 10–20% of eukaryotic genomes, with 30% as an upper limit (Khalturin et al. 2009; Prabh and Rödelsperger 2016). Because orphan genes lack well-studied homologs and characterized structural domains, their functions are difficult to predict (Middendorf et al. 2024). Nonetheless, many are implicated in species-specific adaptations involving behavior, reproduction, host-pathogen interactions, and speciation (Wilson et al. 2005; Khalturin et al. 2009; Fakhar et al. 2023). During their typically brief lifespans—encompassing gene birth, maturation, and eventual loss (Schmid and Aquadro 2001; Tautz and Domazet-Lošo 2011; Palmieri et al. 2014; Iyengar and Bornberg-Bauer 2023; Grandchamp et al. 2024; Lebherz et al. 2024)—low selective constraints facilitate the rapid evolution of their protein-coding sequences (Domazet-Loso and Tautz 2003; Wissler et al. 2013). These changes may refine orphan protein functions and enable their integration into cellular pathways (Domazet-Loso and Tautz 2003; Peng and Zhao 2024).

The *Drosophila melanogaster* gene *goddard* (*gdrd*) was identified as an orphan gene that plays an essential role in male fertility (Zhang et al. 2010; Gubala et al. 2017; Heames et al. 2020; Lange et al. 2021; Peng and Zhao 2024). Although other origination mechanisms cannot be excluded due to the gene's advanced age, Gdrd's testis-biased expression and the lack of structurally similar homologs in the Protein Data Bank suggest a possible *de novo* origin from non-coding sequences (Gubala et al. 2017; Lange et al. 2021; Peng and Zhao 2024). Functional analyses indicate that *gdrd* most likely acts during spermatid elongation, when Gdrd achieves peak expression and localization at the polymerizing axoneme and the insect ring centriole (Lange et al. 2021)). Loss of *gdrd* results in stalled spermatid cysts at late spermatid elongation and the failure to produce mature sperm. At present, the precise role *gdrd* plays at the axoneme or the insect ring centriole remains unknown.

gdrd appears to be phylogenetically restricted to the *Drosophila* genus (Gubala et al. 2017; Heames et al. 2020; Lange et al. 2021; Peng and Zhao 2024)). Within the *melanogaster* species group, it has evolved under purifying selection, indicating functional conservation (Gubala et al. 2017; Peng and Zhao 2024). In more distantly related *Drosophila* species, however, *gdrd* orthologs only weakly align with the *D. melanogaster* protein using BLASTP (Gubala et al. 2017; Lange et al. 2021). Despite these divergences in primary amino acid sequence, all Gdrd orthologs maintain their structural composition: a central helix of conserved length and flanking intrinsically disordered regions (IDRs) (Figure 1A) (Lange et al. 2021). The main structural variation

among extant orthologs is gene length, as orthologs in distant species often contain substantially longer IDRs at their termini (Lange et al. 2021). Whether these differences in gene length or primary sequence affect protein function remains an open question.

Indeed, short protein-coding sequences—a common feature of orphan genes (Toll-Riera et al. 2009; Palmieri et al. 2014; Zhang et al. 2019; Heames et al. 2020)—and protein termini (particularly the first and last 50 amino acids) often evolve at elevated rates (Moutinho et al. 2019; Bricout et al. 2023). Moreover, several studies have documented increased non-synonymous to synonymous substitution ratios (d_N/d_S) in orphan genes across species, suggesting that their amino acid sequences diverge rapidly (Schmid and Tautz 1997; Domazet-Loso and Tautz 2003; Toll-Riera et al. 2009; Palmieri et al. 2014). This enhanced sequence variability appears to stem from intrinsic protein features such as solvent accessibility and intrinsic disorder (Goldman et al. 1998; Guo et al. 2004; Choi et al. 2006).

Interestingly, multiple studies have identified IDRs as a defining feature of *de novo*-emerged orphan genes (Ángyán et al. 2012; Landry et al. 2015; Basile et al. 2017; Heames et al. 2020; Eicholt et al. 2022; Heames et al. 2023; Aubel et al. 2024; Middendorf and Eicholt 2024; Peng and Zhao 2024). Although the precise functions of these regions remain largely unresolved, IDRs generally impart essential protein characteristics such as structural flexibility and accessible interaction sites (Brown et al. 2011). Because IDR properties are primarily physicochemical (e.g., charge, polarity) and largely independent of amino acid position, these regions can tolerate frequent substitutions, insertions, and deletions with minimal impact on function (Uversky 2017; LeBlanc et al. 2024; Jemth 2025). While this primary sequence mutability may mimic evolution under relaxed selection, IDR physicochemical properties remain subject to evolutionary constraint (Brown et al. 2011). Consequently, functionally important physicochemical motifs within IDRs persist across time (Zarin et al. 2021) and can be identified using position-independent, alignment-free homology tools (Chow et al. 2024).

As orphan genes evolve, rapid sequence divergence can have varied consequences, ranging from negligible effects on protein–protein interactions (Vakirlis et al. 2020; Chen et al. 2024; Middendorf et al. 2024) to the refinement of existing functions or even the emergence or loss of function. Because the functions and interaction partners of many orphan genes remain poorly characterized, their evolutionary trajectories are difficult to resolve. Here, we perform assessments of *gdrd* ortholog function and localization in *D. melanogaster* in combination with AlphaFold2 (AF2) predictions, molecular dynamics (MD) simulations, physicochemical analyses of IDRs, and sequence homology analyses to address two questions. Do amino acid changes within a conserved structure alter

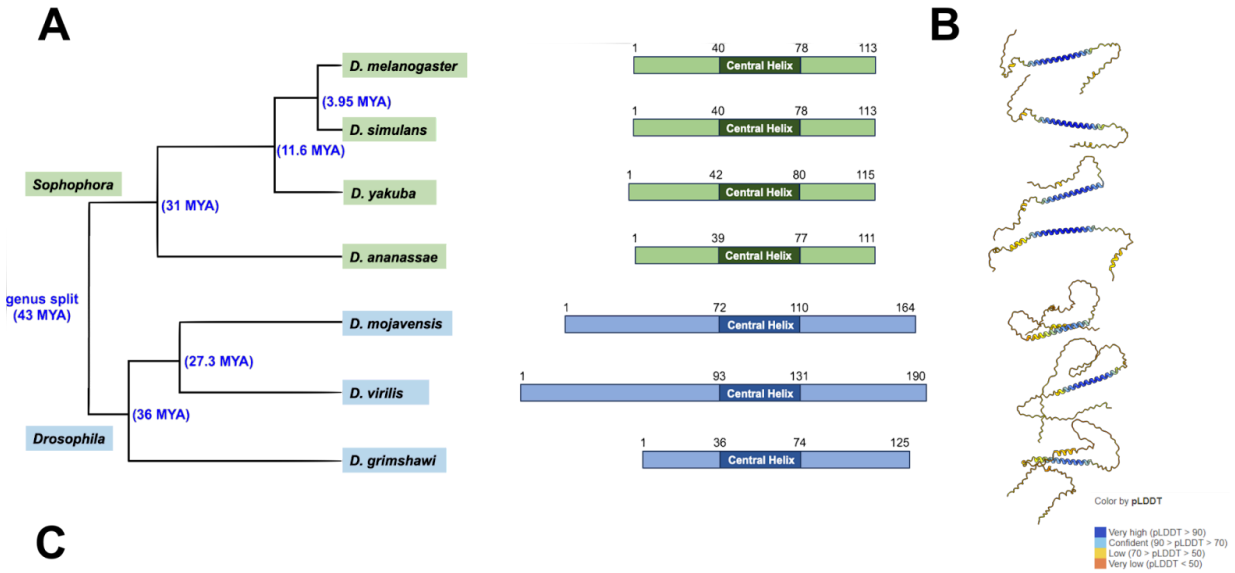
protein function, or are such substitutions largely inconsequential? And, what kinds of functional refinements has the gene undergone during its evolutionary history?

Consistent with Gdrd's structural conservation across the genus, we find that a highly diverged ortholog can functionally replace the *D. melanogaster* gene, suggesting that the ancestral Gdrd protein was integrated into a spermatogenesis-related pathway. In several descendant lineages, however, amino acid changes have led to overall structural destabilization and altered physicochemical properties within terminal IDRs. These changes correlate with an inability to function in *D. melanogaster*. Taken together with observations that several orthologs exhibit distinct subcellular localization patterns when expressed in *D. melanogaster* testes, these findings suggest that Gdrd has retained, lost, and acquired different physicochemical and localization properties during its evolution across the genus.

RESULTS

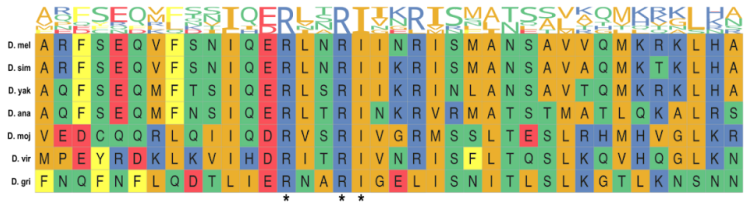
Gdrd orthologs from the *Drosophila* subgenus show substantial primary sequence divergence from the *melanogaster* protein

To investigate the evolutionary fates of *gdrd* orthologs present within the *Drosophila* genus, we selected orthologs from the *Sophophora* and *Drosophila* subgenera (Figure 1A), which split ~40-43 million years ago (MYA) (Russo et al. 1995; Kumar et al. 2017). This included four related orthologs from the subgenus *Sophophora* (inclusive of the *melanogaster* gene) and three divergent orthologs from the subgenus *Drosophila* (Figure 1A). We used AF2 predictions to confirm the structural conservation of these orthologous proteins (Figure 1B).



C

ClustalO alignment of orthologous central helices



% identity of orthologous Gdrd proteins in ClustalO alignments	Full	N-term	Central Helix	C-term
D. mel	100	100	100	100
D. sim	96	100	92	94
D. yak	76	85	76	66
D. ana	42	49	47	27
D. moj	20	15	26	18
D. vir	21	18	26	17
D. gri	11	0	15	13

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:* The Gdrd orthologs share a structure and can be divided into regions of disordered N- and C-termini and a central helix. *Sophophora* subgenus *gdrd*

genes (green) typically have shorter coding sequences compared to their *Drosophila* subgenus (blue) counterparts.
(B) The central helix is conserved in all Gdrd orthologs and predicted with very high confidence by AlphaFold2 (>90 pLDDT). See bottom right for pLDDT score legend.
(C) *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.

To evaluate amino acid sequence conservation among divergent *gdrd* orthologs, we performed pairwise BLASTP comparisons between the *D. melanogaster* protein and the selected *Drosophila* genus orthologs. As expected, *Sophophora* subgenus orthologs display substantial conservation across their full lengths, with sequence identities ranging from 46–96% (Figure S1). In contrast, orthologs from the *Drosophila* subgenus exhibit limited sequence homology. For *D. mojavensis* Gdrd, similarity is confined to the central helix and C-terminus, whereas in *D. virilis*, similarity is restricted to the central helix (Figure S1). The *D. grimshawi* ortholog shows no significant alignment-based similarity to *D. melanogaster* (Figure S1). Pairwise BLASTP analyses between *Drosophila* subgenus species orthologs and the *melanogaster gdrd* termini detects no significant homology within any *Drosophila* subgenus species. This homology detection failure, likely results from the evolution of extended, highly divergent termini in *D. mojavensis* and *D. virilis*, combined with the intrinsic disorder characteristic of these regions (Figure S1).

To get around these sequence homology detection issues, we performed a multiple sequence alignment of *gdrd* orthologs using Clustal Omega (ClustalO) (Sievers et al. 2011). Among *Sophophora* subgenus orthologs, full-length pairwise alignments with *D. melanogaster* Gdrd shows 42–96% identity (Figure 1C). This range decreases to 15-21% identity for *Drosophila* subgenus orthologs, with *D. grimshawi* being the least conserved. When we analyzed percent identity by protein region (N-terminus, central helix and C-terminus), we observed similar levels of amino acid sequence conservation across the Gdrd proteins, with a few exceptions. Notably, The *D. ananassae* C-terminus and *D. grimshawi* N-terminus show elevated sequence divergence relative to other regions (Figure 1C).

Collectively, these data indicate that Gdrd orthologs across the *Drosophila* genus have retained their overall structure but also experienced lineage-specific amino acid sequence changes. This divergence may stem from lineage-specific adaptations arising

through coevolution with other proteins, from reduced selective constraints on the gene, or both.

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

Cross-species complementation assays offer a powerful framework for testing the functional consequences of protein divergence and have been widely used in *Drosophila* to study the evolution of DNA-binding proteins (Bayes and Malik 2009; Saint-Leandre et al. 2020; Brand and Levine 2022). To examine how sequence divergence influences *gdrd* function, we performed gene swap experiments using a previously described *gdrd* rescue construct that restores fertility to *gdrd* null (i.e., full gene deletion) mutants and expresses C-terminally hemagglutinin (HA)-tagged proteins amenable to subcellular localization studies (Lange et al. 2021).

To ensure consistent expression in *D. melanogaster*, we also codon-optimized each *gdrd* ortholog and integrated all constructs into the same genomic locus via phiC31-mediated recombination before crossing them into a *gdrd* null background. Codon optimization, which alters synonymous sites, has been successfully applied to *Drosophila* orthologs in previous studies (Saint-Leandre et al. 2020; Brand and Levine 2022). Because synonymous substitutions can influence gene expression (Zhang and Qian 2025), optimization could theoretically affect transcript abundance. To test this, we assessed transcript levels using RT-PCR and confirmed robust and comparable transcript abundance across nearly all constructs used in this study (Figure S2). Because the level of protein expression for each ortholog could still vary, our subsequent analyses focus primarily on qualitative differences in orthologous protein localization.

We initiated our analysis by replacing the wild-type *D. melanogaster gdrd* coding sequence in the rescue construct with a codon-optimized version and assessing its expression in testes. To improve upon previous whole-mount testes localization studies, we assessed Gdrd expression and subcellular localization in isolated cysts, which facilitate developmental staging and imaging. Each cyst consists of descendants from a single germline stem cell division and two somatically derived cyst cells that enclose the developing germ cells. Within each cyst, germ cells progress synchronously through spermatogenesis, beginning with coordinated mitotic divisions during spermatogonial phases, pre-meiotic growth during spermatocyte stages, meiosis, and then several post-meiotic spermiogenic stages, which include spermatid elongation, individualization, and spermiation (Fabian and Brill 2012). Because Gdrd localizes to the basal body, transition zone, and axoneme during spermatid elongation (Lange et al. 2021), we examined its expression relative to Unc, an early marker of centriole-to-basal body

conversion (Baker et al. 2004). These Unc-positive structures are crucial for the building of the sperm axoneme.

Our analysis of codon-optimized *D. melanogaster* Gdrd expression and localization in isolated cysts recapitulates previous whole-mount findings (Lange et al. 2021) (Figure 2). Gdrd expression initiates in mature spermatocytes, where the protein exhibits diffuse cytoplasmic localization (Figure 2A). At this stage, spermatocytes contain four Unc-positive basal bodies (Figure 2A) that anchor cilia within a plasma membrane cap (Baker et al. 2004; Riparbelli et al. 2012). Besides acetylated tubulin and basal body components (Basiri et al. 2014), these structures also contain components of the transition zone/ ciliary gate, including MKS1, Cep290, BD91/ BD92, and Cby (Pratt et al. 2016). Gdrd shows no enrichment at these ciliary structures, suggesting that the protein 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 early stages of spermatid elongation, Gdrd exhibits cytoplasmic localization (i.e., nuclear and mitochondrial exclusion) and colocalization with Unc at the basal body, transition zone, and axoneme (Figure 2B). As cyst elongation proceeds, this pattern becomes restricted. Unc remains at the basal body and transition zone but no longer associates with the axoneme (Figures 2C–2D). By contrast, Gdrd enriches distally to the Unc-positive basal body, to either a distinct basal body or axonemal 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). In late elongated cysts, Gdrd colocalizes with Unc at the transition zone (Figure 2D), a structure that persists until the onset of sperm individualization (Vieillard et al. 2016). Our immunohistochemical analysis of Gdrd reveals that the protein disappears prior to this disassembly (Figure 2D).

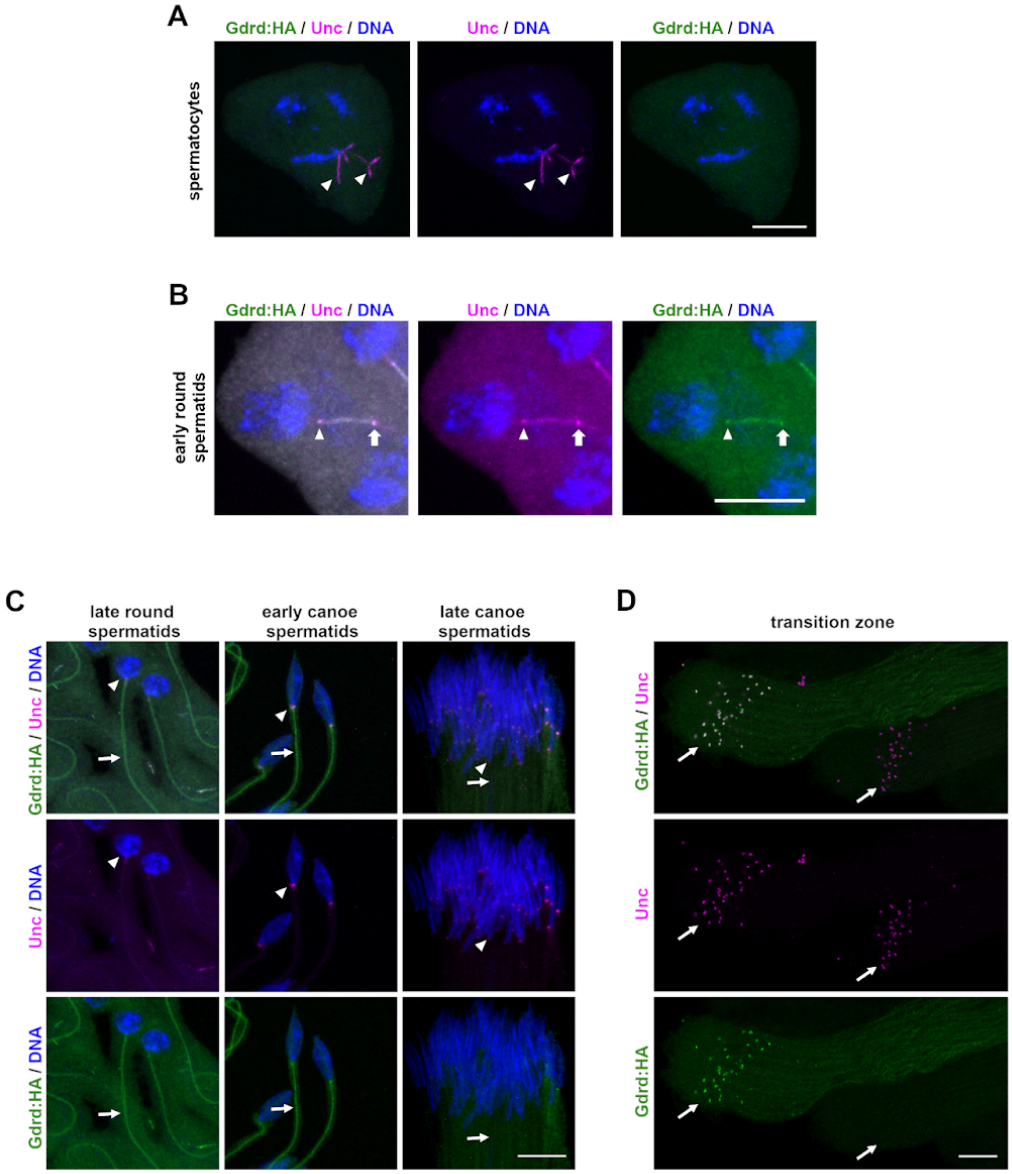


Figure 2: Gdrd associates with the cytoplasmic flagella during spermatid elongation.
Genotype in all images: *unc:EGFP rescue, codon-optimized (CO) D. mel gdrd:HA rescue, Δgdrd*. All images are composed of confocal z-stacks.
(A) Codon-optimized *D. melanogaster* (*D. mel*) Gdrd localizes throughout the cytoplasm but shows no enrichment at cilia-like basal bodies (arrow heads).
(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. Arrow heads or arrows label basal bodies or axonemes, respectively, within a single developing spermatid.

(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). Transition zones are labeled with arrows.
Scale bar for all images = 10 µm.

In many insect species, the sperm axoneme is principally cytoplasmic, with only its distal tip sequestered within a membranous ciliary cap (Johnson 1922; Phillips 1970; Basiri et al. 2014). This compartmentalization is mediated by the insect ring centriole, an electron dense structure that lacks centriole-like organization (Phillips 1970). An analogous murine structure, the annulus, shares both protein components and functions with the insect ring centriole (Kwitny et al. 2010; Basiri et al. 2014; Hoque et al. 2024). In flies, this compartment is thought to promote efficient tubulin polymerization for axoneme elongation (Basiri et al. 2014). Although Gdrd decorates the axoneme throughout elongation, it never localizes beyond the transition zone, suggesting that the compartmentalized distal tip is either biochemically distinct from the cytoplasmic axoneme or that the ciliary gate bars Gdrd entry (Figures 2B, 2D). Altogether, these findings indicate that Gdrd associates with the cytoplasmic flagellum until a time point just before the start of spermatid individualization and that the codon-optimized *D. melanogaster* Gdrd faithfully reproduces prior expression and localization patterns.

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

Sequence analyses revealed that the disordered N- and C-termini of Gdrd orthologs vary in length (*D. mojavensis*, *D. virilis*) and amino acid composition (*D. ananassae*, *D. grimshawi*) (Figure 1C). To assess the functional relevance of these terminal IDRs and contextualize their evolutionary divergence, we performed structure-function analyses using modified *D. melanogaster* rescue constructs lacking either terminal region. The N-terminal truncation (GdrdΔN) removed 36 amino acids after the initial methionine, and the C-terminal truncation (GdrdΔC) removed the final 35 amino acids (Figure 3A). MD simulations based on AF2 predictions predict that both truncated proteins remain structurally stable and retain their central α-helices (Figure 3B). Each construct was then introduced into *D. melanogaster* and assessed for functional complementation. The full-length codon-optimized construct restores fertility to wild-type levels, indistinguishable from the non–codon-optimized rescue (n = 25 per genotype, unequal variance *t*-test, P = 0.08) (Figure 3C). By contrast, both truncations fail to rescue fertility, indicating that the N- and C-terminal regions are essential for Gdrd function (n = 25 per genotype, unequal variance *t*-test, P < 0.0001) (Figure 3C).

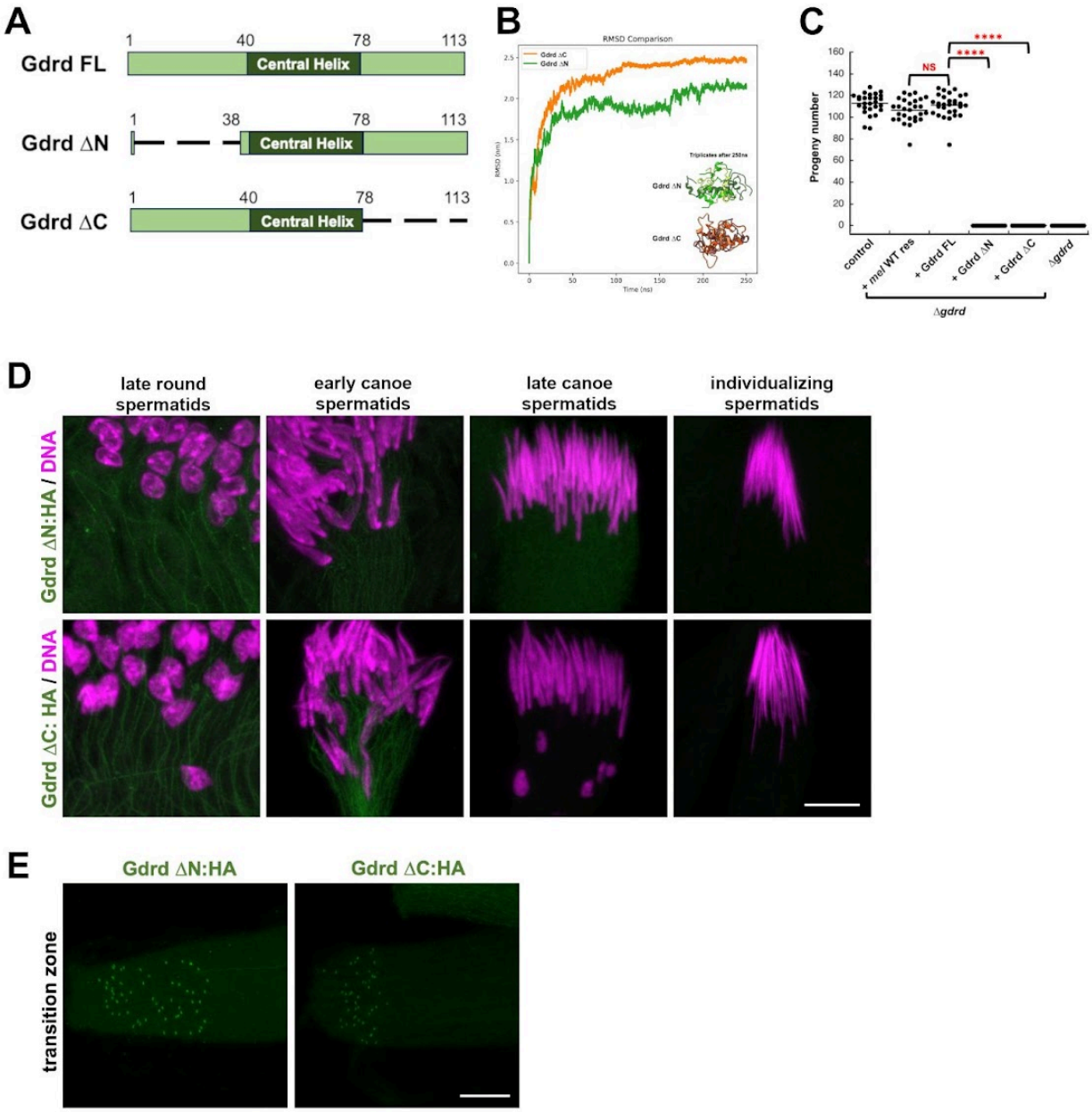


Figure 3: Gdrd's central helix mediates localization to axonemes and transition zones.
(A) Diagram of full length codon-optimized *D. mel* Gdrd (Gdrd FL), *D. mel* Gdrd N-terminal truncation (Gdrd ΔN), and *D. mel* Gdrd C-terminal truncation (Gdrd ΔC).
(B) Averaged backbone RMSD over triplicates of MD 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 S3).
(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 examined the expression and localization patterns of the N- and C-terminally truncated Gdrd proteins to determine whether their inabilities to rescue fertility are caused by altered subcellular distributions. Despite robust transcript levels (Figure S2), both truncated proteins are difficult to detect, suggesting either instability or reduced solubility due to the truncation of the terminal IDRs. Indeed, previous studies have demonstrated the role IDRs play in protein solubility (Tretyachenko et al. 2017; Uversky 2017; Borcherds et al. 2021; Eicholt et al. 2022). Of the two truncations, the C-terminally truncated protein appears the least stable and becomes undetectable in late elongating cysts (Figure 3D). The diminished levels of both truncated proteins most likely accounts for the inability of either construct to restore fertility to *gdrd* mutant males. Nevertheless, even at low protein levels, both truncated proteins localize to axonemes in late round and early canoe spermatids (Figure 3D) and are detectable at the transition zone (Figure 3E). Altogether, these observations suggest that Gdrd's central helix, which is common to both constructs, mediates its localization to the axoneme and insect ring centriole.

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 orthologs from *D. simulans*, *D. yakuba*, *D. ananassae*, *D. mojavensis*, *D. virilis*, and *D. grimshawi*. Since Gdrd-HA is undetectable on western blots, we used RT-PCR to confirm that all constructs produce approximately equal levels of transcript (Figure S2). To analyze protein expression, we performed immunohistochemistry on whole testes and found that all exogenously introduced *gdrd* orthologs are expressible in *D. melanogaster* testes (Figure 4). As all orthologs are driven by the *D. melanogaster* gene's regulatory elements, including its 5' and 3' UTRs, their expression in testes closely parallels that of the native gene, with maximal protein accumulation in post-meiotic round spermatids (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 appears weaker, suggesting that these proteins accumulate at lower levels, have reduced stability, 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, Figure S4). Expression of the *D. grimshawi*

ortholog, however, is robust at early spermatid stages, and the protein appears to associate with the plasma membrane (Figure S4). Altogether, these data indicate that Gdrd orthologs can be expressed in *D. melanogaster* testes.

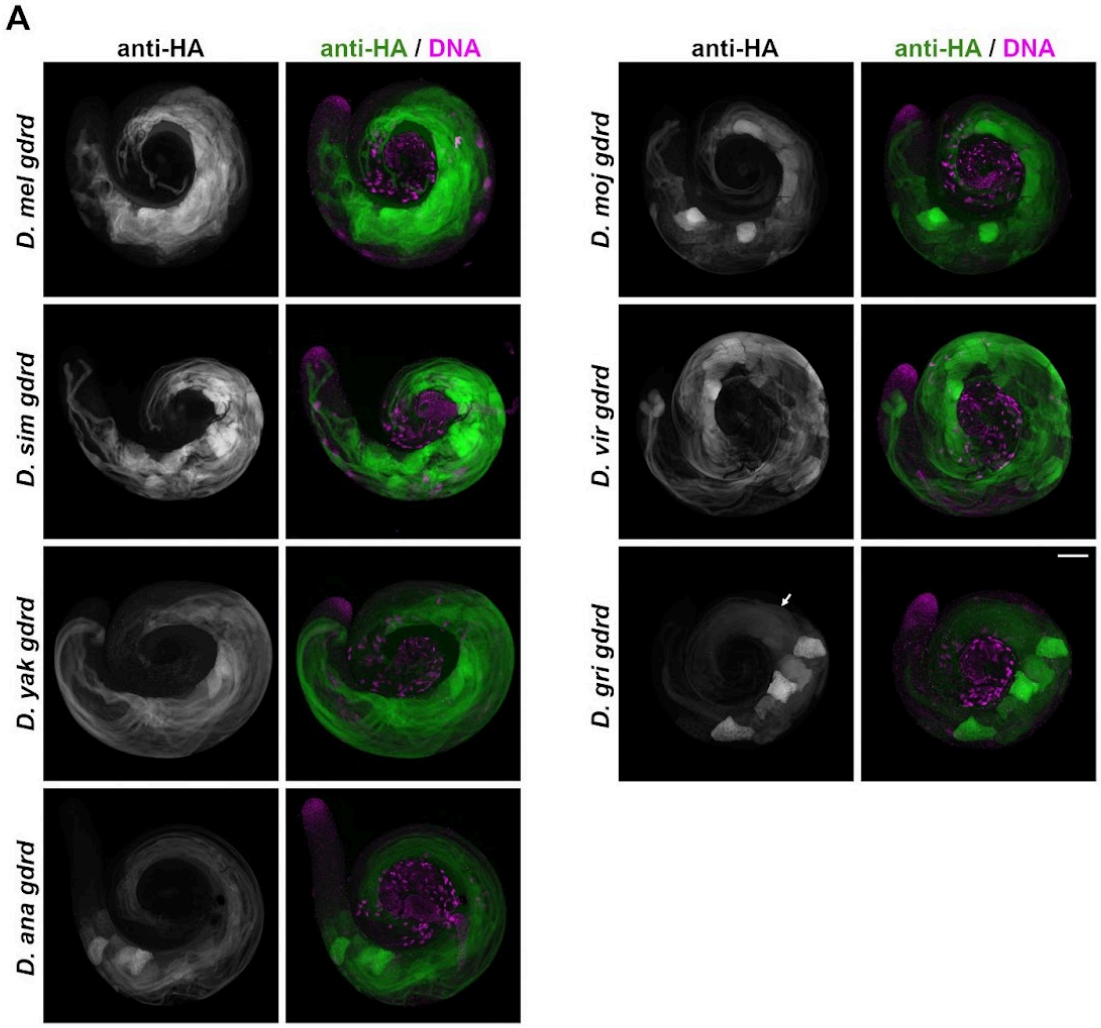


Figure 4: Exogenous expression of *gdrd* orthologs in *D. melanogaster* testes appears largely robust.

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 each 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.

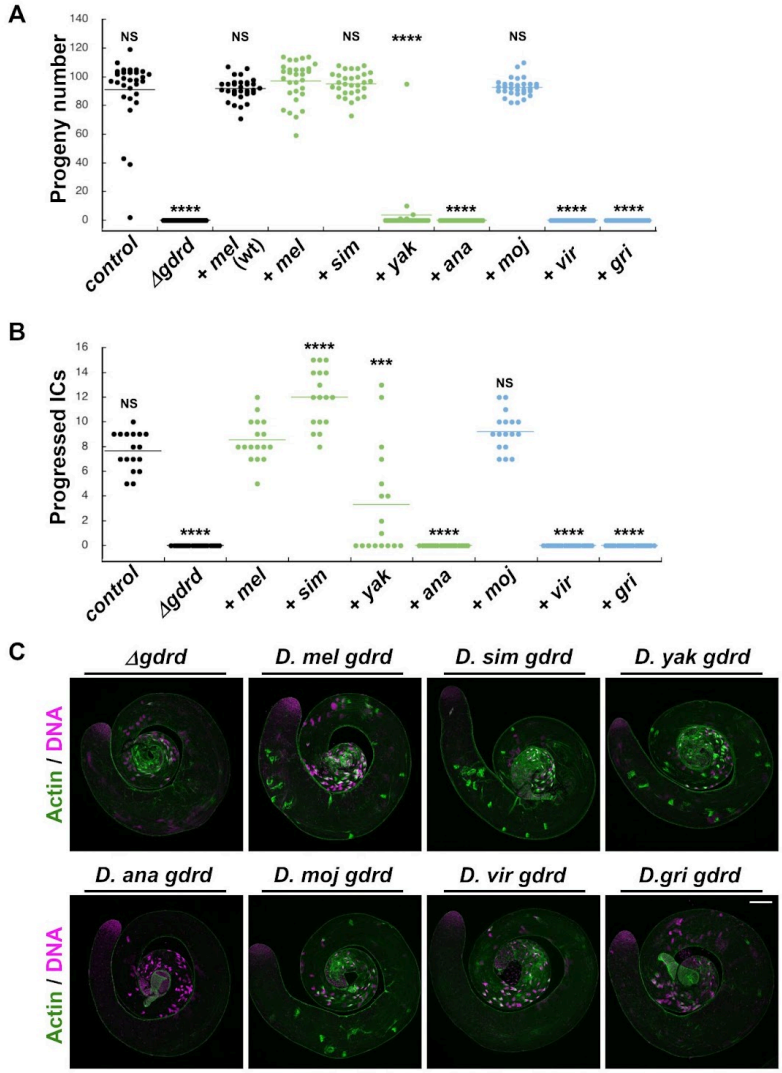


Figure 5: A *gdrd* ortholog that diverged over 40 million years ago fully restores fertility to *gdrd* null *D. melanogaster* males, while several other orthologs of equal or lesser divergence do not.

(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* (*ana*), *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, **** = $P < 0.0001$. One-way ANOVA analysis ($F_{9,290}$ = 579.1, $P < 10^{-178}$) gives equivalent results, with pairwise Tukey-Kramer tests grouping all of the highly fertile genotypes in one group (all $P > 0.57$), and the low-fertility *yakuba* ortholog and all non-rescuing genotypes into a second group (all $P > 0.95$).

(B) Testes from either non-rescuing (*ana*, *vir, gri*) or poorly rescuing (*yak*) ortholog gene swaps display either no or diminished progressed ICs, respectively, 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). Two-sample *t*-tests with unequal variances show significantly reduced levels of IC progression for the *ana*, *vir*, and *gri* orthologs (all $P < 0.0001$) and the *yak* ortholog (P = 0.00015), and significantly increased numbers of progressed ICs for the *sim* ortholog ($P < 0.0001$). In the graph, NS = not significant, *** = $P < 0.001$, **** = $P < 0.0001$. A one-way ANOVA ($F_{8,144} = 111.7$, $P < 10^{-57}$) gives equivalent results, with the genotypes grouping as follows in pairwise Tukey-Kramer tests: *sim* ($p < 0.002$ for all pairwise comparisons to other genotypes); w^{1118}, *mel*, *moj*; *yak* ($P < 0.0001$ for all pairwise comparisons to other genotypes); *ana, vir, gri*.
(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 evaluate the functional conservation of *gdrd* orthologs, we tested their ability to rescue fertility defects in *gdrd* mutants. Successful complementation was interpreted as evidence of functional conservation. Negative results in these assays, however, can have several non-mutually exclusive explanations, including loss or diminished functional capabilities, functional divergence, or lineage-specific co-evolution with binding partners.

Codon-optimized *D. melanogaster gdrd* restores fertility to wild-type levels (n = 30 per genotype, unequal variance *t*-test, P = 0.25) (Figure 5A). Although *Sophophora* Gdrd orthologs share sequence and structural conservation (Figure 1), they exhibit distinct capacities to restore fertility in *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). Despite sharing 76% amino acid identity with *D. melanogaster*, the *D. yakuba* ortholog restores only 4–43% of wild-type fertility levels across replicates (n = 30 per genotype for each replicate, unequal variance *t*-tests, $P < 0.0001$) (Figure 5A, Figure S5). 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 abilities to restore fertility in *D. melanogaster gdrd* mutants. Much to our surprise, the *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 (Figure 1C).

Second, because the common ancestor of *D. mojavensis* and *D. melanogaster* lies at the base of the *Drosophila* genus, the ancestral *gdrd* gene likely already performed the core functions found in extant *D. melanogaster*. Third, the lengthening of the terminal IDRs in *D. mojavensis* (Figures 1A & 5A) appears to have little impact on the protein's central functions. On the other hand, the equally divergent orthologs from *D. virilis* (21% identity) 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 in *D. melanogaster*, amino acid sequence divergence has diminished the functional capabilities of some *gdrd* orthologs while being inconsequential in others.

Because *gdrd* loss-of-function mutations disrupt sperm individualization (Lange et al. 2021), we examined this process in gene-swapped testes by quantifying progressed individualization complexes (ICs). Gene swap testes expressing non-complementing orthologs had no progressed ICs, indicating complete loss of function in this assay (Figure 5B). In contrast, *gdrd* mutant testes expressing the *D. yakuba* ortholog, which partially restores fertility, exhibit progressing ICs, though in significantly reduced numbers compared with those expressing the *D. melanogaster* protein (*D. melanogaster*: 8.5 ± 0.4; *D. yakuba*: 3.3 ± 1.1; $n = 15$, unequal variance *t*-test, $P < 0.0001$) (Figure 5B). Interestingly, *D. simulans* ortholog gene swap testes show an increased number of progressing ICs despite restoring fertility to wild-type levels (Figure 5B). By contrast, *gdrd* mutant testes expressing the *D. ananassae* ortholog exhibit more severe defects than *gdrd* null mutants. Whereas IC formation is reduced but detectable in *gdrd* mutants (Lange et al. 2021), most testes expressing the *D. ananassae* ortholog lacked ICs entirely, suggesting that spermatid elongation never reaches completion (Figure 5D). Because this phenotype appears in two independently generated insertion strains, it is unlikely to result from secondary mutations. Thus, expression of the *D. ananassae* ortholog in *D. melanogaster* may interfere with spermatogenic processes, thereby attenuating elongation. To test whether this effect is dominant or dominant negative, we measured fertility in flies carrying one copy each of the *D. ananassae* and *D. melanogaster* orthologs in a *gdrd* mutant background. These flies displayed fertility levels comparable to controls (Figure S6), indicating that the *D. ananassae* ortholog acts recessively and that the attenuation effect is mild.

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 shows weak expression and no axonemal localization (Figure 4, S4), we exclude this ortholog in all remaining analyses.

As previously shown (Figure 2C), codon-optimized *D. melanogaster* Gdrd decorates the axoneme in late round spermatids, a stage marked by extensive axoneme elongation but limited nuclear re-modeling. This localization persists through late canoe stages and disappears in individualizing cysts (Figure 6A). Both *D. simulans* and *D. yakuba* Gdrd, which restore fertility either fully or partially, respectively, show robust axonemal localization (Figure 6A). In contrast, the *D. ananassae* ortholog, which fails to rescue fertility, localizes weakly to axonemes. Similarly, both *D. virilis* and *D. mojavensis* orthologs localize to axonemes, though *D. virilis* Gdrd appears fainter and more cytoplasmic (Figure 6A). Overall, we find that all orthologs retain some axonemal localization, but those with diminished axonemal binding correspondingly fail to restore fertility in *gdrd* mutant flies.

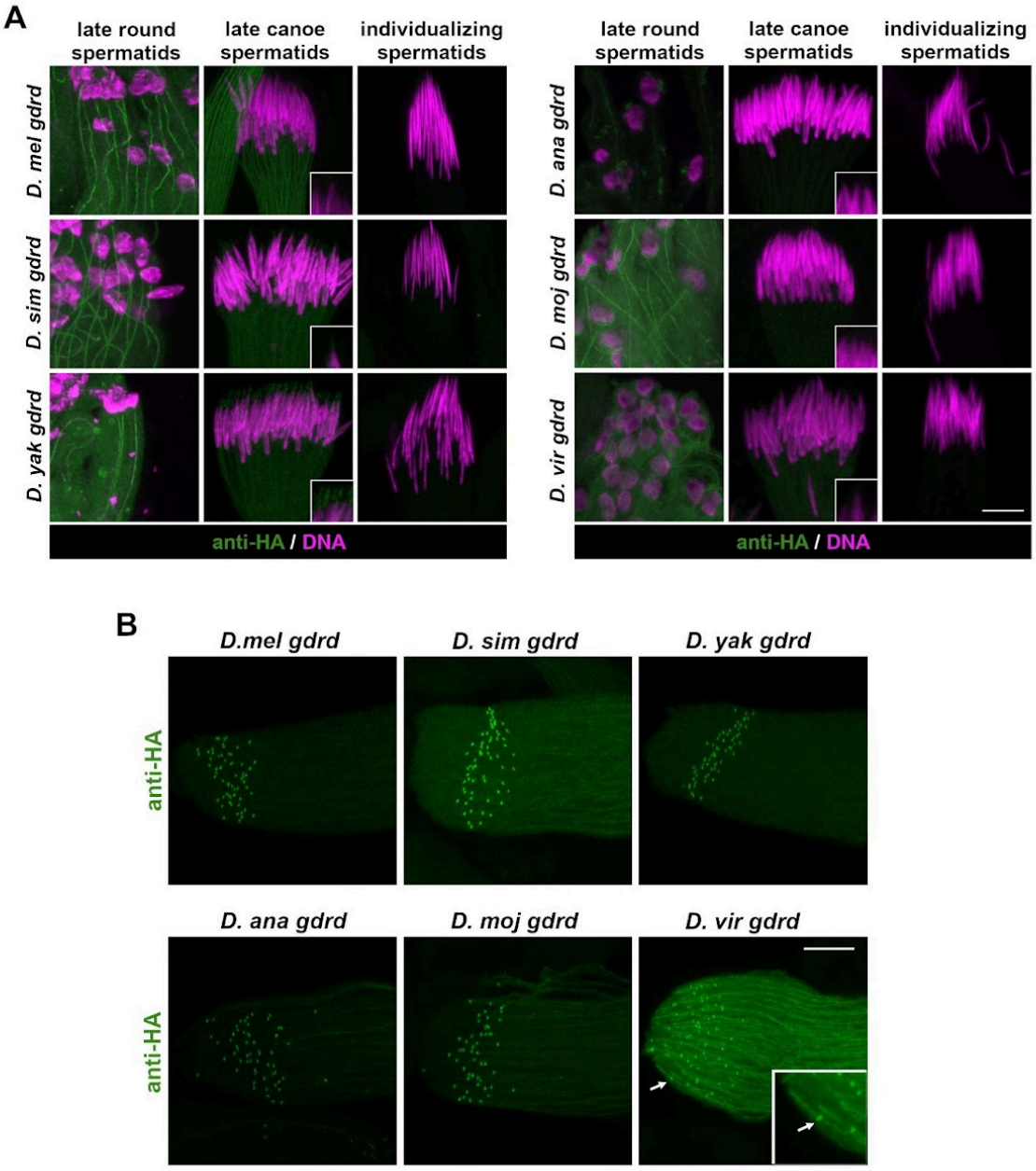


Figure 6: Gdrd orthologs show variable localization to axonemes and often exhibit divergent subcellular localization patterns in late spermatids.
In all images, anti-HA (green) and DAPI (magenta) detect HA-tagged Gdrd orthologs and DNA, respectively. (A) Comparisons of ortholog localizations during late 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. The *D. moj* protein localizes to the apical surface of the round nuclei (arrowhead).
(B) All Gdrd orthologs localize to the transition zone. The *D. vir* ortholog breaches the ciliary gate (inset). Scale bar = 10 µm.

Interestingly, several tested orthologs also exhibit localization patterns that diverge from the *D. melanogaster* protein. In late canoe stage spermatids, *D. simulans* and *D. yakuba* Gdrd both localize near the acrosome at the basal end of spermatid nuclei (Figure 6A). Additionally, *D. yakuba* Gdrd associates with a round organellar structure in late round spermatids (Figure 6A). The *D. ananassae* ortholog, by contrast, decorates large oblong organelles adjacent to round nuclei or the axoneme (Figure 6A). The *D. mojavensis* Gdrd ortholog localizes along the apical side of the nuclear membrane, resembling proteins that associate with the nuclear cap (Figure 6A). Taken together, these data raise the possibility that divergent Gdrd orthologs might have evolved interactions with additional partners.

We next examined if orthologous Gdrds also localize to the transition zone. 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 shows 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 could reflect a loss of ciliary gate function inherent to the *D. melanogaster gdrd* mutant.

These divergent localization patterns, however, are not restricted to elongating spermatid stages, when the Gdrd protein most likely functions, but are often mirrored in earlier spermatogenesis stages, namely 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 observed divergent localization patterns in post-meiotic early round spermatid cysts. During these stages, *D. melanogaster* Gdrd displays dynamic expression and localization. In newly post-meiotic spermatids, Gdrd expression is initially weak but

increases during the onion and leaf blade stages, when the nebenkern forms and unfurls, respectively (Figure 7B). Across all three stages, *D. melanogaster* Gdrd remains primarily cytoplasmic, excluded from nuclei and mitochondrial derivatives (Figure 7B). At the leaf blade stage, when axoneme formation begins, Gdrd localizes to the basal body, growing axoneme, and transition zone/ciliary gate (Figure 2B, Figure 7B).

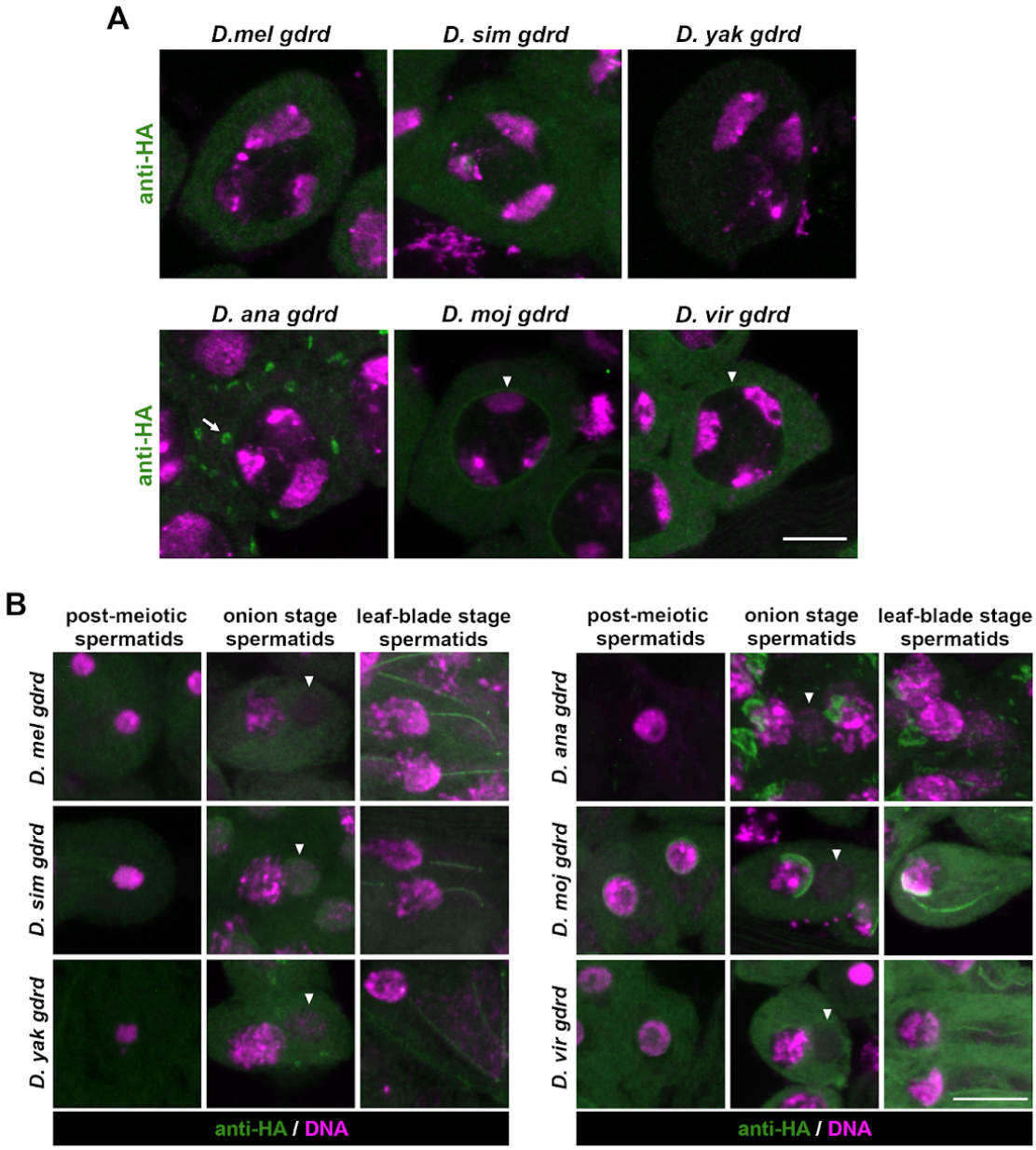


Figure 7: Gdrd orthologs exhibit divergent subcellular localization patterns in spermatocytes and early round spermatids. In all images, anti-HA (green) and DAPI (magenta) detect HA-tagged Gdrd orthologs and DNA, respectively.

(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 (arrow), while the *D. moj* and *D. vir* proteins are enriched at the nuclear perimeter (arrow heads).
(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 (arrow heads), 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 *D. simulans* Gdrd protein localizes to the nebenkern at the onion stage but loses this mitochondrial derivative association as axoneme elongation begins at the leaf blade stage (Figure 7B). The *D. yakuba* ortholog shows a similar trend. At the onion stage, *D. yakuba* Gdrd localizes to round organellar structures resembling those observed later in development (Figure 6A), but shifts to the axoneme at the leaf blade stage (Figure 7B). This pattern suggests that when Gdrd orthologs fail to interact with developing axonemal components, they may mislocalize or form alternate interactions. Alternatively, interacting partners present at the nebenkern or round organelles may be lost at later stages. In contrast, *D. ananassae* Gdrd is weakly expressed in post-meiotic spermatids (Figure 7B), localizing to the developing acrosome in both onion and leaf blade stages. This potentially suggests that its vesicular localization in spermatocytes corresponds to Golgi structures (Figure 7A). Notably, *D. ananassae* Gdrd shows minimal axonemal localization, implying preferential interaction with the acrosome. Likewise, the related *D. mojavensis* and *D. virilis* orthologs also display similar dynamics. Both initially localize to the nuclear periphery after meiosis, then shift asymmetrically to the apical nuclear surface at the onion stage, forming a distinct nuclear cap that persists through the leaf blade stage (Figure 7B). The *D. virilis* protein, however, remains largely cytoplasmic throughout, consistent with its weakened affinity for the *D. melanogaster* axoneme.

To the extent that orthologous protein functions and localizations can be assessed in *D. melanogaster* cells, these data suggest that, despite structural conservation, amino acid sequence divergence among Gdrd orthologs has major effects on both these properties.

Gdrd orthologs exhibit sublineage-specific alterations in physicochemical composition of terminal IDRs

Amino acid sequence divergence can potentially lead to alterations to a protein's intrinsic properties. Because our AlphaFold2 predictions, structure-function analyses, and localization studies highlight the structural and functional conservation of the central helix, we next examined whether amino acid sequence divergence within the terminal IDRs contributes to diminished function or altered localization in ortholog gene swaps. As highly disordered regions evolve rapidly due to weak structural constraints, we applied k-mer–based, alignment-free tools, SHARK-Dive (Chow et al. 2024) and SHARK-Capture (Chow et al. 2025), to identify conserved motifs based on physicochemical similarity. These analyses show that the *D. mojavensis* termini more closely resemble those of *Sophophora* orthologs (Figure 8A), sharing N-terminal motifs (PLDTSED in *Sophophora*; DTTHRSED and PISESIETG in *D. mojavensis*) which are absent in *D. grimshawi* and *D. virilis* orthologs (Figure 8B). Additionally, the net charge of the *D. mojavensis* termini mirrors that of *Sophophora* orthologs, whereas the *D. grimshawi* and *D. virilis* termini differ strongly (Figure 8A). Since the non-rescuing *D. virilis* ortholog and the rescuing *D. mojavensis* ortholog differ primarily in the strength of localization to the axoneme, nuclear cap, and insect ring centriole (Figure 7B), these findings suggest the possibility that divergence of the terminal IDRs may modulate Gdrd's binding affinity/ stability at these subcellular sites.

A

	SHARK-Dive score of orthologous Gdrd proteins termini	
	N-term	C-term
D. mel	0.88	0.88
D. sim	0.88	0.87
D. yak	0.87	0.87
D. ana	0.87	0.86
D. moj	0.81	0.85
D. vir	0.75	0.8
D. gri	0.78	0.79

	SHARK-Dive score (normalized) of orthologous Gdrd proteins termini	
	N-term	C-term
D. mel	1	1
D. sim	1	0.98
D. yak	0.98	0.98
D. ana	0.98	0.76
D. moj	0.49	0.72
D. vir	0	0.09
D. gri	0.25	0

	Overall charge of orthologous Gdrd proteins termini	
	N-term	C-term
D. mel	-12	-5
D. sim	-12	-6
D. yak	-12	-4
D. ana	-8	-3
D. moj	-13	-5
D. vir	-24	-8
D. gri	-3	-1

B

SHARK-Capture detected motifs in Gdrd orthologous termini

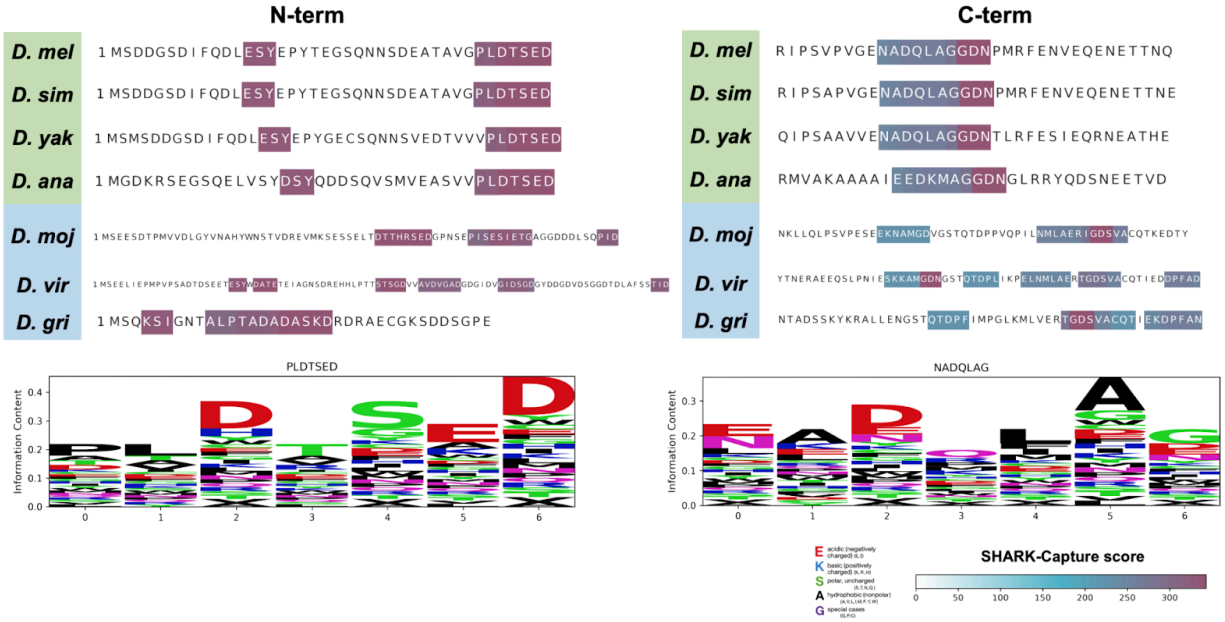


Figure 8: Alignment-free and charge analysis reveal that *D. mojavensis* terminal IDRs resemble those of *Sophophora* group species.

(A) 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*.

(B) 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 ortholog termini (*bottom*). See bottom right for SHARK scores and biophysical properties of amino acids.

MD simulations show that Goddard's conserved central α-helix remains stable across lineages, while the N- and C-termini remain disordered

We next investigated whether the ability of an ortholog to function in *D. melanogaster* correlated with its predicted structural stability. To do so, we assessed the stability of the AF2-predicted protein structures (Figure 1B) 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 (Aubel et al. 2023; Liu et al. 2023; Middendorf and Eicholt 2024), and MD simulations can test the stability of the predicted structures (Lange et al. 2021; Middendorf et al. 2024; Peng and Zhao 2024).

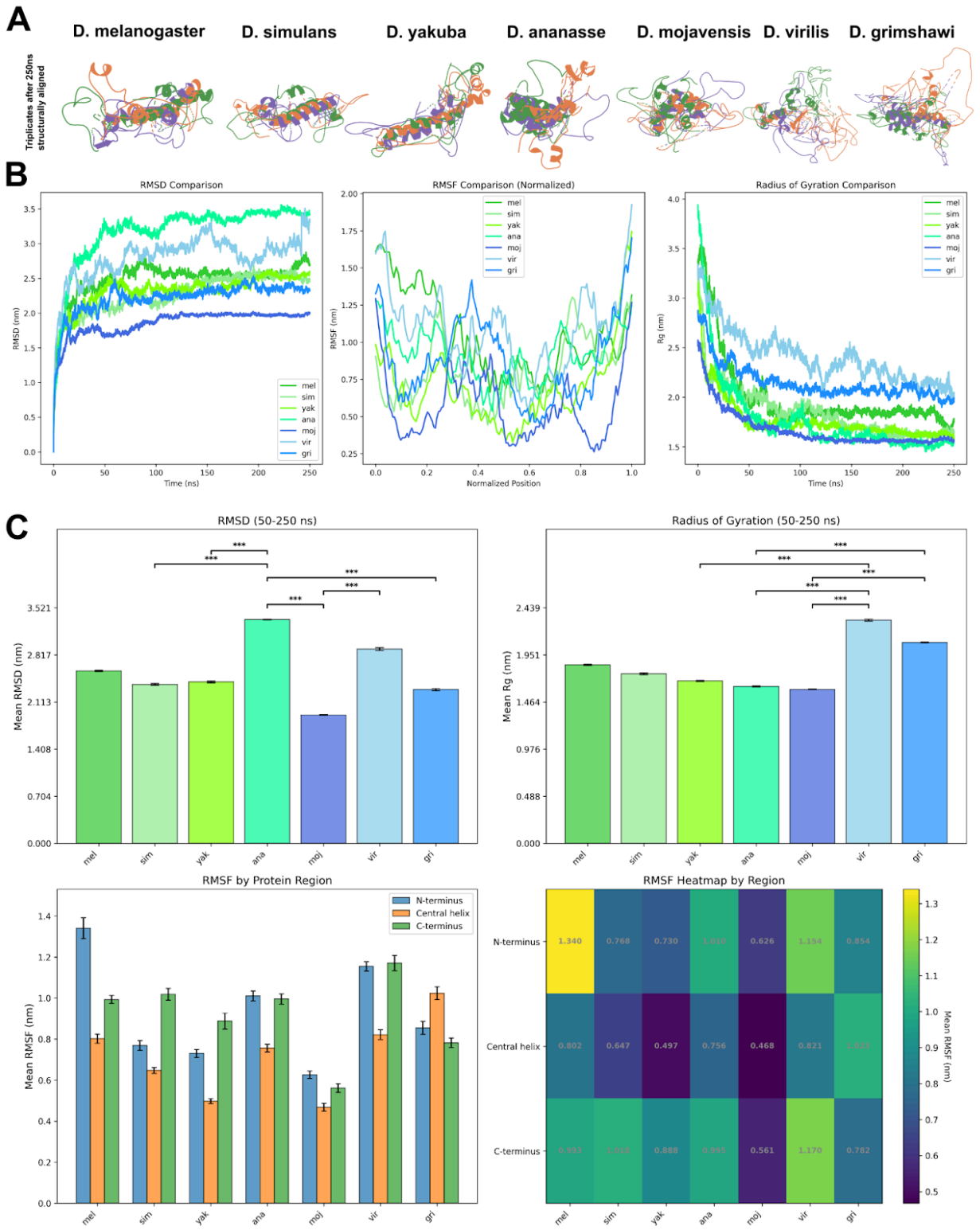


Figure 9: MD simulations of Gdrd orthologs over 250ns.
(A) Structural alignments of triplicate MD simulations after 250 ns highlight the dynamic structures of Gdrd orthologs, while the central helix remains stable. (B) Averaged backbone RMSD over time (left), α-C RMSF per normalized residue position (center) and radius of gyration over time (right) of all Gdrd

orthologs. *D. ananassae* shows by far the highest RMSD, indicating relative instability. α-C RMSF is high at the termini for all orthologs, in accordance with their disorder, while central residues fluctuate less and remain a structured helix overall. The trend of radii of gyration shows that all orthologs become more compact over time, likely forming molten globule-like structures. (C) Mean RMSD, radius of gyration, and RMSF values averaged over the equilibrated 50–250 ns region. Kruskal–Wallis H-tests detected significant interspecies differences (*** = $p < 0.001$), followed by Dunn's post-hoc tests with Bonferroni correction for 21 pairwise comparisons (see Supplement). Only the top 5 most significant comparisons are shown here. *D. ananassae* exhibited the highest RMSD, whereas *D. virilis* showed the largest radius of gyration. Flexibility was highest in the termini, while the conserved central helix remained stably folded across all orthologs. Single runs and α-C RMSF per residue available in supplementary materials (Figures S7, S8).

Our modeling and MD simulations revealed that all orthologous Gdrd proteins exhibit structural flexibility (Figure 9). Notably, we consistently observe flexible terminal regions characterized by disorder and transient short helix formations (Figure 9A). In contrast, the central helical region remained largely stable across all simulations (Figure 9). Root Mean Square Deviation (RMSD) analyses confirm this intrinsic flexibility through high deviation values indicative of dynamic behavior (Seffernick et al. 2019) (Figure 9B, left). Among the studied orthologs, *D. ananassae* had the highest RMSD values after relaxation, suggesting it is the most flexible or least stable, followed by *D. virilis* and *D. grimshawi*, while *D. mojavensis* appeared the most stable (Figure 9C, top left).

An examination of residue-specific flexibility using Root Mean Square Fluctuation (RMSF) analysis (Figure 9B, center) confirms that the stable central helix exhibits substantially fewer fluctuations than the disordered termini. The N-terminus of the *D. melanogaster* ortholog exhibits the highest fluctuations during simulations, while the *D. mojavensis* protein shows the least fluctuations along the whole protein (Figure 9C, bottom). Additionally, we used Radius of Gyration (Rg) analysis to determine the overall compactness and stability across orthologs (Figure 9B, right). Consistent with RMSD and RMSF observations, orthologs display decreasing radii of gyration over simulation time (Figure 9B), indicative of initial structural relaxation followed by stable yet flexible conformations, characteristic of molten globule-like structures. After initial relaxation (first 50 ns), the *D. virilis* and *D. grimshawi* proteins display a significantly higher mean Rg than other orthologs (Figure 9C, top right). Collectively, our results show that Gdrd orthologs retain a conserved, stably folded core but differ in the flexibility and compaction of their disordered termini, pointing to evolutionary fine-tuning of terminal IDRs rather than wholesale structural changes as the likely source of functional divergence.

DISCUSSION

Through a combination of orthologous gene swaps, sequence analyses, structural modeling, and molecular dynamics simulations, we investigated when a putative *de novo*-evolved orphan gene might have acquired essential roles in fertility and how interspecific variation in protein sequence and structural properties affects the ability of the protein to perform its essential function in *D. melanogaster*. Our findings reveal that *gdrd* orthologs differ in their ability to function in *D. melanogaster* spermatogenesis, and this variation does not align consistently with phylogenetic distance. Notably, a highly diverged ortholog from *D. mojavensis* in the *Drosophila* group fully restores fertility, suggesting that the ancestral *gdrd* gene had already evolved a critical role in sperm production roughly 40-43 million years ago.

This work extends previous studies emphasizing structural conservation in *de novo* genes (Schmitz et al. 2018; Dowling et al. 2020; Lange et al. 2021; Chen et al. 2024; Middendorf et al. 2024; Peng and Zhao 2024) by showing that functional conservation can be equally long-standing. In addition to the conserved central helix and terminal IDRs (Figures 1 & 8), our findings suggest that the ancestral Gdrd protein likely shared other key traits with extant orthologs, including associations with the axoneme and insect ring centriole (Figure 6). MD simulations further indicate that these structural elements are not only conserved but dynamically stable: the central helix remains stably folded across lineages, whereas the disordered termini retain high flexibility and compact over time into molten-globule–like conformations (Figure 9), reflecting conserved structural plasticity.

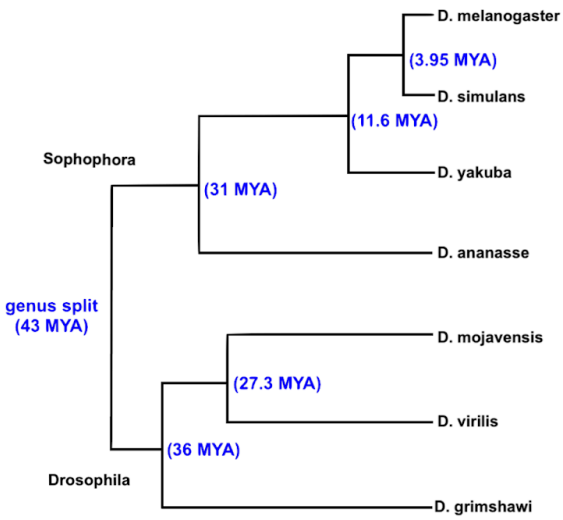
Structure-function analyses of the *D. melanogaster* protein indicate that its interactions with the axoneme and insect ring centriole are mediated by the conserved central helix, while the IDRs may modulate their strength or facilitate additional interactions (Figure 3). Comparing rescuing and non-rescuing orthologs further highlights the features critical for Gdrd function in *D. melanogaster*: orthologs that localize strongly to the axoneme can substitute for the native gene, suggesting Gdrd acts at this structure. A third feature that might have been present in the genus ancestor is the presence of several conserved physicochemical motifs within the IDRs. K-mer-based SHARK-Capture analysis revealed several conserved physicochemical motifs within the IDRs of rescuing orthologs (*D. mojavensis*, *melanogaster* group) that are absent in *D. virilis*, potentially explaining the *D. virilis* protein's weaker axonemal and nuclear cap localization (Figures 7-8). Interestingly, the non-rescuing *D. ananassae* ortholog, despite being more closely related to the *melanogaster* group than *D. mojavensis* and containing the same conserved terminal motifs as rescuing species (Figure 8), shows the highest instability during MD simulations (Figure 9), suggesting that excessive

structural lability could potentially impair its ability to complement the loss of the native gene.

In our ortholog gene swap experiments, we observed distinct subcellular localization patterns among several Gdrd orthologs. These novel patterns could stem either from lineage-specific evolution leading to functional innovation in their native species or from mislocalization in *D. melanogaster* caused by weakened interactions with host proteins. However, we favor the former explanation for the *D. mojavensis*, *D. virilis*, and *D. ananassae* orthologs, the stable and distinct localizations of which during spermatid elongation suggest the emergence of lineage-specific functions. Interestingly, expanded Gdrd localization frequently involves membranous structures, such as the nuclear envelope in *D. mojavensis* and *D. virilis* or, the acroblast/ Golgi in *D. ananassae* and the plasma membrane in *D. grimshawi*. The *D. yakuba* protein also associates with an unidentified spherical organelle during elongation. This pattern may reflect an intrinsic membrane affinity of Gdrd, consistent with reports showing that intrinsically disordered proteins often mediate membrane binding (MacAinsh et al. 2025). In the context of *de novo* gene evolution, rapid sequence and physicochemical shifts within IDRs may drive divergence in Gdrd's membrane or organellar associations, which vary in their chemical and physical properties. Supporting this idea, studies in yeast suggest that membranes can act as evolutionary incubators for *de novo* evolved gene products by providing stable environments that prevent protein aggregation or purge while allowing for the refinements of protein-protein interactions (Vakirlis et al. 2020; Houghton et al. 2024). Indeed, two other *Drosophila* orphan genes, *saturn* and *katherine johnson*, also encode membrane-associated proteins essential for fertility (Gubala et al. 2017; Guay et al. 2025), suggesting that this evolutionary strategy may be commonly used in *Drosophila*. Interestingly, *goddard* orthologs that restore *D. melanogaster* fertility show broad localization to tubulin-associated, membrane-proximal structures such as the axoneme, ring centriole, and spermatid nuclear envelope, raising the possibility that the ancestral Gdrd could have functioned at membrane-linked microtubular interfaces. Whether nuclear cap localization represents an ancestral feature or a lineage-specific expansion in *repleta* and *virilis* group species, however, remains unresolved.

Collectively, our data suggest that Gdrd was likely fully functional in the common ancestor but that its biophysical properties were maintained in some lineages and altered or lost in others (Figure 10). This pattern may be broadly representative of orphan genes that experience rapid sequence evolution and contain IDRs. Just as the gain and loss of *de novo* genes occur along a continuum (Carvunis et al. 2012; Heames et al. 2020; Iyengar and Bornberg-Bauer 2023), their functional trajectories likely follow a similar path, shaped by the binding promiscuity and polyvalency inherent to their IDRs. Such flexibility may promote coevolution with diverse interaction partners across

lineages, enabling rapid adaptation and integration into cellular networks (Bornberg-Bauer et al. 2021; Jemth 2025). This framework may explain why *Drosophila* subgenus orthologs – except *D. mojavensis*, which retains similar motifs and charge profiles (Figure 8) – cannot restore fertility despite preserving overall structure. These orthologs might have coevolved with distinct molecular partners, resulting in divergent IDR properties and altered 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 10: 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 also suggest that amino acid sequence divergence in orphan genes can have major functional consequences, as exogenously expressed orthologs often show diminished or lost activity and altered subcellular localizations. While structural conservation commonly suggests functional conservation, a case somewhat comparable to *gdrd* is seen with *bag of marbles* (*bam*), a well-studied orphan gene essential for *D. melanogaster* male and female fertility (McKearin and Spradling 1990). Like Gdrd, Bam lacks identifiable binding domains and contains intrinsically disordered regions (Lamb et al. 2020; Bubnell et al. 2022; Farfán-Pira et al. 2023; Baker et al. 2024). However, CRISPR/Cas9 knockouts in multiple *Drosophila* species have shown that *bam*'s essential role in fertility has been independently lost in several *melanogaster* group lineages (Lamb et al. 2020; Bubnell et al. 2022; Farfán-Pira et al. 2023; Baker et al. 2024), even though the gene has an equivalent function to *D.*

melanogaster in a divergent member of the *Drosophila* subgenus, *D. americana (Arnce et al. 2025)*.

While our gene swap experiments suggest that amino acid sequence divergence can lead to diminished or lost function, this approach has inherent limitations that must be considered. One potential shortcoming is that rescue constructs may not achieve expression levels equivalent to the endogenous locus. However, several lines of evidence suggest that this effect is likely minimal in our study. First, expression of all of our rescue constructs is driven by the transcriptional regulatory regions of *D. melanogaster gdrd*, and the level of expression achieved is sufficient for full rescue by three orthologs (Figures 3, 5). Second, transcript levels are broadly similar across most orthologs (Figure S2). Third, cytological analyses indicate that most orthologs are expressed at levels comparable to, or higher than, the rescuing *D. mojavensis* ortholog. To more definitively assess functional differences, future studies could employ CRISPR/Cas9 to insert orthologs directly into the native *gdrd* locus. (Adikusuma et al. 2017; Saint-Leandre et al. 2020; Brand and Levine 2022). This approach would clarify whether the *D. virilis* and *D. ananassae* orthologs are entirely nonfunctional in *D. melanogaster* or capable of partial rescue when expressed at more endogenous levels. Conversely, higher expression of the *D. ananassae* gene might exacerbate spermatid elongation defects, further complicating interpretation.

A second limitation arises from the exogenous nature of gene swap assays. Although testis architecture and function are broadly conserved across *Drosophila*, interspecies physiological differences persist (Schärer et al. 2008; Lüpold et al. 2016; Alpern et al. 2019). Addressing this limitation will require direct examination of orthologous Gdrd proteins in their native testes, using CRISPR-generated knockouts in combination with either specific antibodies or HA-tagged knock-ins in non-*melanogaster* species (Lamb et al. 2020; Bubnell et al. 2022; Farfán-Pira et al. 2023; Baker et al. 2024; Arnce et al. 2025).

In conclusion, our findings suggest that the common ancestor of the *Sophophora* and *Drosophila* subgenera expressed a Gdrd protein that already possessed the core interactions and activities required for spermatogenesis in present-day *D. melanogaster*. The conservation of transition zone localization and the association of axonemal localization with function underscore that proper Gdrd positioning at these structures is critical for its role. However, our gene swap analyses also indicate that Gdrd function might have diverged across lineages. Orthologs from *D. mojavensis* and *D. virilis* show distinct and expanded localization patterns in *D. melanogaster*, suggesting broader subcellular roles in their native species, while the conserved *D. ananassae* ortholog

may have evolved new interaction preferences. Future studies will determine whether these altered localization patterns correspond to 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 (Guay et al. 2025) at 25°C without a light-dark cycle.

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 (Lange et al. 2021), 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 (Sievers et al. 2011) and analyzed using ggmsa and ggplot2 in R v4.2.2 (Wickham 2011; Zhou et al. 2022). Truncated versions of *D. melanogaster* Gdrd were designed based on AlphaFold2 (Jumper et al. 2021) predictions, experimental structure data (Lange et al. 2021), 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 (Dohmen et al. 2025) to annotate the methods used to identify Gdrd in earlier studies (Gubala et al. 2017; Lange et al. 2021). The DeNoFo annotation file available in the supplementary material.

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 (Chow et al. 2024). Similarity was normalized with D. *melanogaster* Gdrd self-comparison set as 1. Shared motifs were identified using SHARK-Capture in default mode (Chow et al. 2025). 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 (Lange et al. 2021). For ortholog gene swaps, the *D. melanogaster*, *D. simulans* (LOC6738055), *D. yakuba* (LOC6534118), *D. ananassae* (LOC6507245), *D. mojavensis*, *D. virilis*, and *D. grimshawi gdrd* coding sequences were first codon-optimized using GENEius software (Eurofins MWG Operon). Amino acid sequences for *D. mojavensis*, *D. virilis*, and *D. grimshawi* Gdrd orthologs were previously documented (Gubala et al. 2017). 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. RT-PCR was used to evaluate transcript levels in sexually mature males of carrying each transgene in the *gdrd* null background. Primers located in the 5' and 3' UTRs of *gdrd* were used to amplify the full length protein-coding sequence (including the 3xHA tag) of each transgene from cDNA as previously described (Rivard et al. 2021). As a control for cDNA synthesis and to detect any contaminating genomic DNA, the housekeeping control gene *RpL19* was also amplified from each cDNA sample. 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*

921 virgin females. All males were collected as virgins and then aged in uncrowded
922 conditions for four to six days prior to mating. The age range of virgin females fell
923 between four to eight days. Likewise, virgin females used in the assay were isolated in
924 uncrowded condition and reared on a high yeast diet prior to mating to increase
925 fecundity. Mating and egg laying proceeded for 48 hours post-crossing before parents
926 were discarded. Progeny number was determined by counting the number of pupal
927 cases on the side of each vial 10 days (~216 hours) after setting the cross. Final sample
928 sizes for each genotype were n = 25-30 males, depending on the experiment. Progeny
929 count data were analyzed in Microsoft Excel using *t*-tests with unequal variances.
930

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

933 Antibody labeling of whole testis and testis squashes were performed as previously
934 described in Lange *et al*. (Lange et al. 2021) and Sitaram *et al*. (Sitaram et al. 2014),
935 respectively. In brief, processed tissues were fixed following dissection in 4%
936 paraformaldehyde/ 1X phosphate buffered saline (PBS) for 20 minutes. After washing
937 the fixative away, tissues were blocked for 1-2 hours in 1X PBS containing 3% bovine
938 serum albumin, 5% normal goat serum, and 0.1% Triton-X. The samples were then
939 incubated overnight in rabbit anti:HA (C29F4, Cell Signaling Technologies) diluted at
940 1:100 in blocking solution. The following day, the antibody was washed off before
941 applying goat anti-rabbit Alexa Fluor 488 conjugated secondary (A-11008,
942 ThermoFisher Scientific) diluted at 1:200 in blocking solution. The secondary was then
943 washed away prior to mounting samples in Vectashield Plus Antifade Mounting Medium
944 Plus DAPI (Vector labs). To analyze individualization complexes during
945 spermatogenesis, fixed whole testes were incubated overnight with TRITC conjugated
946 phalloidin (R415, Life Technologies/ Molecular Probes) diluted at 1:200 in PBS
947 containing 0.1% Triton-X. Microscopy was performed on an SP8 X confocal microscope
948 (Leica Microsystems) using HC PL APO CS2 20x/0.75 ILL and HC PL APO CS2
949 63x/1.40 oil objectives. DAPI, Alexa Fluor 488, and TRITC were excited using 405,
950 488, and 546 wavelengths, respectively. Post-acquisition processing was performed
951 using ImageJ Fiji (version 1.0) (Schindelin et al. 2012).
952

953 **Structure predictions and MD simulations**

954 Structure predictions were performed with AlphaFold2 (v.2.3.1; Database cut-off date:
955 2022-12-19) (Jumper et al. 2021). MD simulations were performed using GROMACS
956 2021.2 (Páll et al. 2020) on AlphaFold2 predictions with highest mean predicted local
957 distance difference test (pLDDT) score. Structures were prepared using OPLS-AA/L
958 force field (Shivakumar et al. 2010) and solvent SOL. The structure was solvated in a
959 cubic box of SPC/E water with 10-Å clearance and the electrostatic charge neutralized
960 by the addition of sodium atoms, followed by energy minimization and equilibration.

Triplicate 250 ns 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, α-C RMSF, radius of gyration). For each species, triplicate 250 ns MD simulation snapshots were aligned using PyMOL v3.0.5 (Schrödinger 2015). Structures were visualized using PyMOL v3.0.5 and ChimeraX-1.7.1 (Pettersen et al. 2021). Statistical analyses and plotting were performed using Python v.3.10.13 (libraries: numpy (Harris et al. 2020), pandas (Mckinney 2011), scipy (Virtanen et al. 2020), scikit-posthocs (Terpilowski 2019), matplotlib (Hunter 2007)). For RMSD and radius of gyration, we analyzed equilibrated trajectories 50-250 ns subsampled at 100-frame intervals, yielding ~600 observations per species (200 timepoints × 3 replicates). For RMSF, all residues within each region (N-terminus, core, C-terminus) across three replicates were analyzed (n=100-300 per species). Kruskal-Wallis H-tests assessed overall differences, followed by Dunn's post-hoc tests with Bonferroni correction (α=0.05, 21 comparisons). NaN values were excluded.

Use of Artificial Intelligence

After completing data collection and analysis and drafting and editing the full manuscript, we used GPT-5 to condense the text. We then manually edited again prior to submission.

Acknowledgements

We 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 RGP0006/2013) 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).

Data Availability Statement

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. Data underlying figures from the manuscript is provided in supplemental material.

1003 **References**

1004 Adikusuma F, Pederick D, McAninch D, Hughes J, Thomas P. 2017. Functional
1005 Equivalence of the SOX2 and SOX3 Transcription Factors in the Developing Mouse
1006 Brain and Testes. *Genetics* 206:1495–1503.

1007 Alpern JHM, Asselin MM, Moehring AJ. 2019. Identification of a novel sperm class and
1008 its role in fertilization in Drosophila. *J Evol Biol* 32:259–266.

1009 Ángyán AF, Perczel A, Gáspári Z. 2012. Estimating intrinsic structural preferences of de
1010 novo emerging random-sequence proteins: is aggregation the main bottleneck?
1011 *FEBS Lett* 586:2468–2472.

1012 Arnce LR, Bubnell JE, Aquadro CF. 2025. Comparative functional and evolutionary
1013 analysis of essential germline stem cell genes across the genus Drosophila and
1014 two outgroup species. *G3 (Bethesda)* [Internet]. Available from:
1015 http://dx.doi.org/10.1093/g3journal/jkaf224

1016 Aubel M, Buchel F, Heames B, Jones A, Honc O, Bornberg-Bauer E, Hlouchova K.
1017 2024. High-throughput Selection of Human de novo-emerged sORFs with High
1018 Folding Potential. *Genome Biol Evol* [Internet] 16. Available from:
1019 http://dx.doi.org/10.1093/gbe/evae069

1020 Aubel M, Eicholt L, Bornberg-Bauer E. 2023. Assessing structure and disorder
1021 prediction tools for emerged proteins in the age of machine learning. *F1000Res*
1022 12:347.

1023 Baker CA, Guan X-J, Choi M, Murthy M. 2024. The role of in specifying courtship
1024 behaviors across divergent species. *Sci Adv* 10:eadk1273.

1025 Baker JD, Adhikarakunnathu S, Kernan MJ. 2004. Mechanosensory-defective,
1026 male-sterile unc mutants identify a novel basal body protein required for
1027 ciliogenesis in Drosophila. *Development* 131:3411–3422.

1028 Basile W, Sachenkova O, Light S, Elofsson A. 2017. High GC content causes orphan
1029 proteins to be intrinsically disordered. *PLoS Comput Biol* 13:e1005375.

1030 Basiri ML, Ha A, Chadha A, Clark NM, Polyanovsky A, Cook B, Avidor-Reiss T. 2014. A
1031 migrating ciliary gate compartmentalizes the site of axoneme assembly in
1032 Drosophila spermatids. *Curr Biol* 24:2622–2631.

1033 Bayes JJ, Malik HS. 2009. Altered heterochromatin binding by a hybrid sterility protein
1034 in Drosophila sibling species. *Science* 326:1538–1541.

1035 Borcherds W, Bremer A, Borgia MB, Mittag T. 2021. How do intrinsically disordered
1036 protein regions encode a driving force for liquid-liquid phase separation? *Curr Opin*
1037 *Struct Biol* 67:41–50.

1038 Bornberg-Bauer E, Hlouchova K, Lange A. 2021. Structure and function of naturally
1039 evolved de novo proteins. *Curr Opin Struct Biol* 68:175–183.

1040 Brand CL, Levine MT. 2022. Cross-species incompatibility between a DNA satellite and
1041 the Drosophila Spartan homolog poisons germline genome integrity. *Curr Biol*
1042 32:2962–2971.e4.

1043 Bricout R, Weil D, Stroebel D, Genovesio A, Roest Crollius H. 2023. Evolution is not
1044 Uniform Along Coding Sequences. *Mol Biol Evol* [Internet] 40. Available from:
1045 http://dx.doi.org/10.1093/molbev/msad042

1046 Brown CJ, Johnson AK, Dunker AK, Daughdrill GW. 2011. Evolution and disorder. *Curr*
1047 *Opin Struct Biol* 21:441–446.

1048 Bubnell JE, Ulbing CKS, Fernandez Begne P, Aquadro CF. 2022. Functional Divergence
1049 of the bag-of-marbles Gene in the Drosophila melanogaster Species Group. *Mol*
1050 *Biol Evol* [Internet] 39. Available from: http://dx.doi.org/10.1093/molbev/msac137

1051 Carvunis A-R, Rolland T, Wapinski I, Calderwood MA, Yildirim MA, Simonis N,
1052 Charloteaux B, Hidalgo CA, Barbette J, Santhanam B, et al. 2012. Proto-genes and
1053 de novo gene birth. *Nature* 487:370–374.

1054 Chen J, Li Q, Xia S, Arsala D, Sosa D, Wang D, Long M. 2024. The Rapid Evolution of
1055 De Novo Proteins in Structure and Complex. *Genome Biol Evol* [Internet] 16.
1056 Available from: http://dx.doi.org/10.1093/gbe/evae107

1057 Choi SS, Vallender EJ, Lahn BT. 2006. Systematically assessing the influence of
1058 3-dimensional structural context on the molecular evolution of mammalian
1059 proteomes. *Mol Biol Evol* 23:2131–2133.

1060 Chow CFW, Ghosh S, Hadarovich A, Toth-Petroczy A. 2024. SHARK enables sensitive
1061 detection of evolutionary homologs and functional analogs in unalignable and
1062 disordered sequences. *Proc Natl Acad Sci U S A* 121:e2401622121.

1063 Chow CFW, Lenz S, Scheremetjew M, Ghosh S, Richter D, Jegers C, von Appen A,
1064 Alberti S, Toth-Petroczy A. 2025. SHARK-capture identifies functional motifs in
1065 intrinsically disordered protein regions. *Protein Sci* 34:e70091.

1066 Dohmen E, Aubel M, Eicholt LA, Roginski P, Luria V, Karger A, Grandchamp A. 2025.
1067 DeNoFo: a file format and toolkit for standardized, comparable de novo gene
1068 annotation. *Bioinformatics* 41:btaf539.

1069 Domazet-Loso T, Tautz D. 2003. An evolutionary analysis of orphan genes in
1070 Drosophila. *Genome Res* 13:2213–2219.

1071 Dowling D, Schmitz JF, Bornberg-Bauer E. 2020. Stochastic Gain and Loss of Novel
1072 Transcribed Open Reading Frames in the Human Lineage. *Genome Biol Evol*
1073 12:2183–2195.

1074 Eicholt LA, Aubel M, Berk K, Bornberg-Bauer E, Lange A. 2022. Heterologous
1075 expression of naturally evolved putative de novo proteins with chaperones. *Protein*
1076 *Sci* 31:e4371.

1077 Fabian L, Brill JA. 2012. Drosophila spermiogenesis: Big things come from little
1078 packages. *Spermatogenesis* 2:197–212.

1079 Fakhar AZ, Liu J, Pajerowska-Mukhtar KM, Mukhtar MS. 2023. The Lost and Found:
1080 Unraveling the Functions of Orphan Genes. *J Dev Biol* [Internet] 11. Available from:
1081 http://dx.doi.org/10.3390/jdb11020027

1082 Farfán-Pira KJ, Martínez-Cuevas TI, Evans TA, Nahmad M. 2023. A cis-regulatory
1083 sequence of the selector gene vestigial drives the evolution of wing scaling in
1084 Drosophila species. *J Exp Biol* [Internet] 226. Available from:
1085 http://dx.doi.org/10.1242/jeb.244692

1086 Goldman N, Thorne JL, Jones DT. 1998. Assessing the impact of secondary structure
1087 and solvent accessibility on protein evolution. *Genetics* 149:445–458.

1088 Grandchamp A, Czuppon P, Bornberg-Bauer E. 2024. Quantification and modeling of
1089 turnover dynamics of de novo transcripts in Drosophila melanogaster. *Nucleic Acids*
1090 *Res* 52:274–287.

1091 Guay SY, Patel PH, Thomalla JM, McDermott KL, O'Toole JM, Arnold SE, Obrycki SJ,
1092 Wolfner MF, Findlay GD. 2025. An orphan gene is essential for efficient sperm entry
1093 into eggs in Drosophila melanogaster. *Genetics* [Internet] 229. Available from:
1094 http://dx.doi.org/10.1093/genetics/iyaf008

1095 Gubala A, Schmitz JF, Kearns MJ, Vinh T, Bornberg-Bauer E, Wolfner M, Findlay GD.
1096 2017. The Goddard and Saturn genes are essential for Drosophila male fertility and
1097 may have arisen DE Novo. *Mol. Biol. Evol.* 34:1066–1082.

1098 Guo HH, Choe J, Loeb LA. 2004. Protein tolerance to random amino acid change. *Proc*
1099 *Natl Acad Sci U S A* 101:9205–9210.

1100 Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D,
1101 Wieser E, Taylor J, Berg S, Smith NJ, et al. 2020. Array programming with NumPy.
1102 *Nature* 585:357–362.

1103 Heames B, Buchel F, Aubel M, Tretyachenko V, Loginov D, Novák P, Lange A,
1104 Bornberg-Bauer E, Hlouchová K. 2023. Experimental characterization of de novo
1105 proteins and their unevolved random-sequence counterparts. *Nat Ecol Evol*
1106 7:570–580.

1107 Heames B, Schmitz J, Bornberg-Bauer E. 2020. A Continuum of Evolving De Novo
1108 Genes Drives Protein-Coding Novelty in Drosophila. *J Mol Evol* 88:382–398.

1109 Hoque M, Li F-Q, Weber WD, Chen JJ, Kim EN, Kuo P-L, Visconti PE, Takemaru K-I.

2024. The Cby3/ciBAR1 complex positions the annulus along the sperm flagellum during spermiogenesis. *J Cell Biol* [Internet] 223. Available from: http://dx.doi.org/10.1083/jcb.202307147

Houghton CJ, Coelho NC, Chiang A, Hedayati S, Parikh SB, Ozbaki-Yagan N, Wacholder A, Iannotta J, Berger A, Carvunis A-R, et al. 2024. *De novo* proteins integrate cellular systems using ancient protein targeting and degradation pathways. *bioRxivorg* [Internet]. Available from: http://biorxiv.org/lookup/doi/10.1101/2024.08.28.610198

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

Iyengar BR, Bornberg-Bauer E. 2023. Neutral Models of De Novo Gene Emergence Suggest that Gene Evolution has a Preferred Trajectory. *Mol Biol Evol* [Internet] 40. Available from: http://dx.doi.org/10.1093/molbev/msad079

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

Johnson HH. 1922. Peripheral migration of a centriole derivative in the spermatogenesis of Oeligcanthus. *Science* 56:759–760.

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583–589.

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

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

Kwitny S, Klaus AV, Hunnicutt GR. 2010. 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* 82:669–678.

Lamb AM, Wang Z, Simmer P, Chung H, Wittkopp PJ. 2020. affects pigmentation divergence and cuticular hydrocarbons in and. *Front Ecol Evol* [Internet] 8. Available from: http://dx.doi.org/10.3389/fevo.2020.00184

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

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

Lebherz MK, Fouks B, Schmidt J, Bornberg-Bauer E, Grandchamp A. 2024. DNA Transposons Favor De Novo Transcript Emergence Through Enrichment of Transcription Factor Binding Motifs. *Genome Biol Evol* [Internet] 16. Available from: http://dx.doi.org/10.1093/gbe/evae134

LeBlanc C, Stefani J, Soriano M, Lam A, Zintel MA, Kotha SR, Chase E, Pimentel-Solorio G, Vunnum A, Flug K, et al. 2024. Conservation of function without conservation of amino acid sequence in intrinsically disordered transcriptional activation domains. *bioRxiv* [Internet]. Available from: http://dx.doi.org/10.1101/2024.12.03.626510

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

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

MacAinsh M, Muhammedkutty FNK, Prasad R, Zhou H-X. 2025. Membrane association of intrinsically disordered proteins. *Annu. Rev. Biophys.* 54:275–302.

McKearin DM, Spradling AC. 1990. bag-of-marbles: a Drosophila gene required to initiate both male and female gametogenesis. *Genes Dev.* 4:2242–2251.

Mckinney W. 2011. pandas: a Foundational Python Library for Data Analysis and Statistics. *Python for high performance and scientific computing* 14.9:1–9.

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

Middendorf L, Ravi Iyengar B, Eicholt LA. 2024. Sequence, Structure, and Functional Space of Drosophila De Novo Proteins. *Genome Biol Evol* [Internet] 16. Available from: http://dx.doi.org/10.1093/gbe/evae176

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

Páll S, Zhmurov A, Bauer P, Abraham M, Lundborg M, Gray A, Hess B, Lindahl E. 2020. Heterogeneous parallelization and acceleration of molecular dynamics simulations in GROMACS. *J Chem Phys* 153:134110.

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

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

1181 Pereira AB, Marano M, Bathala R, Zaragoza RA, Neira A, Samano A, Owoyemi A,
1182 Casola C. 2025. Orphan genes are not a distinct biological entity. *Bioessays*
1183 47:e2400146.

1184 Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin
1185 TE. 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and
1186 developers. *Protein Sci* 30:70–82.

1187 Phillips DM. 1970. Insect sperm: their structure and morphogenesis. *J Cell Biol*
1188 44:243–277.

1189 Prabh N, Rödelsperger C. 2016. Are orphan genes protein-coding, prediction artifacts,
1190 or non-coding RNAs? *BMC Bioinformatics* 17:226.

1191 Pratt MB, Titlow JS, Davis I, Barker AR, Dawe HR, Raff JW, Roque H. 2016. Drosophila
1192 sensory cilia lacking MKS proteins exhibit striking defects in development but only
1193 subtle defects in adults. *J Cell Sci* 129:3732–3743.

1194 Riparbelli MG, Callaini G, Megraw TL. 2012. Assembly and persistence of primary cilia
1195 in dividing Drosophila spermatocytes. *Dev Cell* 23:425–432.

1196 Rivard EL, Ludwig AG, Patel PH, Grandchamp A, Arnold SE, Berger A, Scott EM, Kelly
1197 BJ, Mascha GC, Bornberg-Bauer E, et al. 2021. A putative de novo evolved gene
1198 required for spermatid chromatin condensation in Drosophila melanogaster. *PLoS*
1199 *Genet.* 17:e1009787.

1200 Russo CA, Takezaki N, Nei M. 1995. Molecular phylogeny and divergence times of
1201 drosophilid species. *Mol Biol Evol* 12:391–404.

1202 Saint-Leandre B, Christopher C, Levine MT. 2020. Adaptive evolution of an essential
1203 telomere protein restricts telomeric retrotransposons. *Elife* [Internet] 9. Available
1204 from: http://dx.doi.org/10.7554/eLife.60987

1205 Schärer L, Da Lage J-L, Joly D. 2008. Evolution of testicular architecture in the
1206 Drosophilidae: a role for sperm length. *BMC Evol Biol* 8:143.

1207 Schindelin J, Arganda-Carreras I, Frise E, Kaynig V, Longair M, Pietzsch T, Preibisch S,
1208 Rueden C, Saalfeld S, Schmid B, et al. 2012. Fiji: an open-source platform for
1209 biological-image analysis. *Nat Methods* 9:676–682.

1210 Schmid KJ, Aquadro CF. 2001. The evolutionary analysis of “orphans” from the
1211 Drosophila genome identifies rapidly diverging and incorrectly annotated genes.
1212 *Genetics* 159:589–598.

1213 Schmid KJ, Tautz D. 1997. A screen for fast evolving genes from Drosophila. *Proc Natl*
1214 *Acad Sci U S A* 94:9746–9750.

1215 Schmitz JF, Ullrich KK, Bornberg-Bauer E. 2018. Incipient de novo genes can evolve

1216 from frozen accidents that escaped rapid transcript turnover. *Nat Ecol Evol*
1217 2:1626–1632.

1218 Schrödinger LLC. 2015. PyMOL The PyMOL Molecular Graphics System, Version 1.8,
1219 Schrödinger, LLC.

1220 Seffernick JT, Ren H, Kim SS, Lindert S. 2019. Measuring Intrinsic Disorder and
1221 Tracking Conformational Transitions Using Rosetta ResidueDisorder. *J Phys Chem*
1222 *B* 123:7103–7112.

1223 Shivakumar D, Williams J, Wu Y, Damm W, Shelley J, Sherman W. 2010. Prediction of
1224 Absolute Solvation Free Energies using Molecular Dynamics Free Energy
1225 Perturbation and the OPLS Force Field. *J Chem Theory Comput* 6:1509–1519.

1226 Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H,
1227 Remmert M, Söding J, et al. 2011. Fast, scalable generation of high-quality protein
1228 multiple sequence alignments using Clustal Omega. *Mol Syst Biol* 7:539.

1229 Sitaram P, Hainline SG, Lee LA. 2014. Cytological analysis of spermatogenesis: live
1230 and fixed preparations of Drosophila testes. *J Vis Exp*:e51058.

1231 Tautz D, Domazet-Lošo T. 2011. The evolutionary origin of orphan genes. *Nat Rev*
1232 *Genet* 12:692–702.

1233 Terpilowski M. 2019. scikit-posthocs: Pairwise multiple comparison tests in Python. *J.*
1234 *Open Source Softw.* 4:1169.

1235 Toll-Riera M, Bosch N, Bellora N, Castelo R, Armengol L, Estivill X, Albà MM. 2009.
1236 Origin of primate orphan genes: a comparative genomics approach. *Mol Biol Evol*
1237 26:603–612.

1238 Tretyachenko V, Vymětal J, Bednárová L, Kopecký V Jr, Hofbauerová K, Jindrová H,
1239 Hubálek M, Souček R, Konvalinka J, Vondrášek J, et al. 2017. Random protein
1240 sequences can form defined secondary structures and are well-tolerated in vivo.
1241 *Sci Rep* 7:15449.

1242 Uversky VN. 2017. Paradoxes and wonders of intrinsic disorder: Stability of instability.
1243 *Intrinsically Disord Proteins* 5:e1327757.

1244 Vakirlis N, Acar O, Hsu B, Castilho Coelho N, Van Oss SB, Wacholder A,
1245 Medetgul-Ernar K, Bowman RW 2nd, Hines CP, Iannotta J, et al. 2020. De novo
1246 emergence of adaptive membrane proteins from thymine-rich genomic sequences.
1247 *Nat. Commun.* 11:781.

1248 Vieillard J, Paschaki M, Duteyrat J-L, Augière C, Cortier E, Lapart J-A, Thomas J,
1249 Durand B. 2016. Transition zone assembly and its contribution to axoneme
1250 formation in Drosophila male germ cells. *J Cell Biol* 214:875–889.

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat. Methods* 17:261–272.

Wickham H. 2011. Ggplot2: Ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* 3:180–185.

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

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

Zarin T, Strome B, Peng G, Pritišanac I, Forman-Kay JD, Moses AM. 2021. Identifying molecular features that are associated with biological function of intrinsically disordered protein regions. *Elife* [Internet] 10. Available from: http://dx.doi.org/10.7554/eLife.60220

Zhang J, Qian W. 2025. Functional synonymous mutations and their evolutionary consequences. *Nat. Rev. Genet.*:1–16.

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

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

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

Zhou L, Feng T, Xu S, Gao F, Lam TT, Wang Q, Wu T, Huang H, Zhan L, Li L, et al. 2022. Ggmsa: A visual exploration tool for multiple sequence alignment and associated data. *Brief. Bioinform.* 23:bbac222.