nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-021-21667-6 OPEN

Structural and functional characterization of a putative de novo gene in *Drosophila*

Andreas Lange[1,5], Prajal H. Patel[2,5], Brennen Heames[1], Adam M. Damry[3], Thorsten Saenger[4], Colin J. Jackson[3], Geoffrey D. Findlay[2]✉ & Erich Bornberg-Bauer[1]✉

Comparative genomic studies have repeatedly shown that new protein-coding genes can emerge de novo from noncoding DNA. Still unknown is how and when the structures of encoded de novo proteins emerge and evolve. Combining biochemical, genetic and evolutionary analyses, we elucidate the function and structure of *goddard*, a gene which appears to have evolved de novo at least 50 million years ago within the *Drosophila* genus. Previous studies found that *goddard* is required for male fertility. Here, we show that Goddard protein localizes to elongating sperm axonemes and that in its absence, elongated spermatids fail to undergo individualization. Combining modelling, NMR and circular dichroism (CD) data, we show that Goddard protein contains a large central α-helix, but is otherwise partially disordered. We find similar results for Goddard's orthologs from divergent fly species and their reconstructed ancestral sequences. Accordingly, Goddard's structure appears to have been maintained with only minor changes over millions of years.

[1] Institute for Evolution and Biodiversity, University of Münster, Münster, Germany. [2] Department of Biology, College of the Holy Cross, Worcester, MA, USA. [3] Research School of Chemistry, ANU College of Science, Canberra, Australia. [4] Department of Pediatric Kidney, Liver and Metabolic Diseases, Hannover Medical School, Hannover, Germany. [5] These authors contributed equally: Andreas Lange, Prajal H. Patel. ✉email: gfindlay@holycross.edu; ebb@wwu.de

De novo evolved genes are novel genes that arise from previously noncoding DNA[1–4]. Contrary to most other newly evolved genes, which are generated by duplications[5] or rearrangements of existing gene fragments[6], de novo genes are not derived from existing, protein-coding sequences. Accordingly, selection may only act on the functional structure of an encoded protein after it has been born. De novo genes have been confirmed across a wide range of eukaryotes[7–14].

Studies over the last decade have illustrated the key mechanisms underlying these genes' emergence. Many de novo protein-coding genes start out as a noncoding transcript, transcribed from intergenic or intronic regions. As circumstantial events, mutations that create a protein-coding open-reading frame (ORF) can occur in the transcribed region[11,13–17]. The order of these steps can vary, and each of the steps is frequent. Recent evidence showed an abundance of species-specific, newly generated transcripts supporting de novo gene emergence[11,18,19]. While most novel transcripts are quickly lost, those that are retained and encode a polypeptide are exposed to selection, eliminating novel proteins that are deleterious for cell function[11,20]. While the computationally predicted structural properties (such as secondary structure and disorder) of de novo proteins do not appear to change significantly over millions of years[11], relatively little is known about how these properties are acquired upon gene birth. In particular, it remains unclear how de novo genes, which arise from essentially random sequences, are able to form their initial structures, acquire functions, become fixed in a population, and persist beyond several speciation events. Understanding these issues is important, given that de novo genes have already changed our perceptions of how genomic novelties can arise[21]. In particular, it is often proposed that newly evolved genes in general, and de novo genes in particular, are involved in many important processes such as development, stress response, and environmental adaptation[4,22].

Most insights concerning de novo gene evolution stem from large-scale comparative genomic, transcriptomic, and proteomic studies. Less is known about the specific functions and structures of de novo proteins because very few of them have been studied in detail. One example is the "antifreeze glycoprotein" (AFGP), which protects Arctic codfishes from freezing[23,24]. AFGP acquired, probably by convergent evolution, a structure that is similar to an evolutionary unrelated antifreeze protein found in Antarctic notothenoids[25,26]. Another example is Bsc4, a non-essential de novo protein found in *Saccharomyces cerevisiae* and implicated in DNA repair under nutrient-deficient conditions[8,27]. Bsc4 contains large disordered regions[28], but further details regarding its structure, cellular location, and function remain unclear. Recently, a de novo yeast protein was computationally and experimentally shown to progressively evolve properties that place it into the endoplasmic reticulum membrane[29]. Finally, two putative de novo genes, named *goddard* (*gdrd*) and *saturn*, have been identified in fruit fly[30] (note that we stick to a widely used convention in fly genetics, capitalizing protein and putting the gene names in lower case italics). Both genes appear to have arisen from intronic regions at least 50 million years ago (Mya), at the root of the *Drosophila* genus (Supplementary Fig. S1), and preliminary structural features have been predicted computationally. Both genes are expressed specifically in the male reproductive tract, a pattern conserved across many species, and RNA interference tests found that both are essential for male fertility in *D. melanogaster* (*Dmel*)[30].

A recent conceptualization for defining de novo gene functionalisation[31] describes five levels of functional analysis. Our previous work on *gdrd*[30] addressed the gene's expression (a conserved, male-specific pattern across *Drosophila* species) and began to investigate the protein's evolutionary implications (conserved in most fly species, except for *D. willistoni*) and capacities (predicted to have one major α-helix with disordered termini). Here, we present a more detailed analysis of the structure, function, and evolution of the Gdrd protein, which allowed us to describe its molecular and structural properties, its cellular function, and thus its potential, physiological implications. We also further elaborate on its evolutionary implications[31]. We first use computational and experimental approaches to determine the structure of Gdrd protein from *Dmel*. Then, using null and tagged rescue alleles of the *gdrd* gene, we show that Gdrd protein localizes to elongating sperm axonemes and it is required to form individual sperm cells in the postmeiotic testis. Finally, we predict the structures of orthologous Gdrd proteins from other *Drosophila* species and use ancestral sequence reconstruction to infer how this structure might have arisen and subsequently evolved. Gdrd's high degree of structural conservation, coupled with its functional role, suggests that it likely became involved in spermatogenesis early in the evolution of the *Drosophila* genus.

Results

Gdrd is monomeric, soluble, and compact with a helix at its core and disordered termini. To further assess the likelihood that Gdrd (with length 113 residues) forms a stably folded protein, we carried out predictions for a number of biophysical properties (see "Methods"). The *Dmel* Gdrd has a theoretical pI of 4.25 and does not contain any cysteine or tryptophan residues. Secondary structure predictions consistently indicate an α-helix at the core (residues 40–80), as suggested by Gubala et al.[30] (Supplementary Fig. S2a–c). Kyte-Doolittle hydrophobicity indicates 17% hydrophobic residues, which are primarily present in the core α-helix, with no indication of transmembrane regions (Supplementary Fig. S2d). Therefore, Gdrd is likely a stably folded protein. The existence of this core α-helix is further corroborated by the prediction of a coiled-coil formation (CC) between residues 45 and 80. Consistent with the gene's potential de novo origin, CCs can be formed relatively easily, from sequences that are almost random, provided they have at least a clear hydrophobic polar pattern[32]. Generally, many predicted CCs, in particular short ones, have overlapping predictions, e.g., with regions predicted to be disordered[32]. However, this ambivalence may also reflect their true structural state, since CCs are often formed from non-folding structural elements in response to triggers such as binding to another protein[32,33]. A second, shorter helix is also predicted near the N-terminus (residues 10–18), but with lower confidence. The rest of the protein, in particular the termini, is predicted as disordered[30]. Using additional programs (see "Methods") to investigate Gdrd further, we find: (i) that the core α-helix is stably folded, while the termini of the protein appear less ordered, (ii) no indication of toxic aggregation-prone segments, (iii) that solubility is predicted to be high for much of the sequence with the exception of the hydropathic core helix, (iv) finally, that Gdrd is not predicted to have regions likely to undergo liquid–liquid phase separation. Taken together, preliminary structural predictions indicate that Gdrd adopts a soluble (validated by SDS-PAGE, Supplementary Fig. S2f), relatively compact, non-aggregating structure with helices at its core (Supplementary Fig. S2a–d) and N-terminus and partially disordered regions throughout the remainder of the protein (Supplementary Figs. S2 and S3).

We further corroborated these predictions with ab initio tertiary structure prediction using the QUARK server[34]. Consistent with the above heuristic methods, a helical core and terminal disorder are predicted (Supplementary Fig. S4). We performed a pairwise root-mean-square deviation (RMSD)

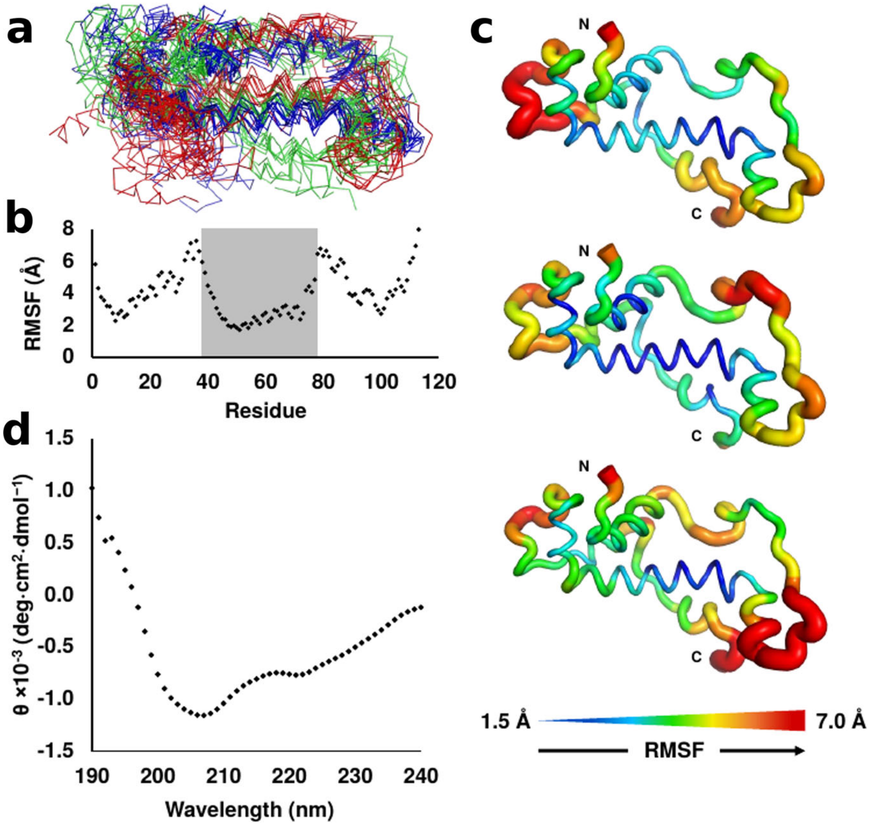

Fig. 1 Molecular dynamics (MD) and circular dichroism (CD) of Gdrd confirm a partially ordered alpha helical structure. a Representative backbone ensemble of the modeled Gdrd structure composed of ten frames sampled every 20 ns from each of three 200-ns MD replicates (shown as green, blue, and red ribbons, respectively). The central helix and a portion of the N-terminal helix remain stably folded across all three simulations despite considerable flexibility in the rest of the protein structure, indicative of a partially ordered structure. **b** Plot of C_α root-mean-square fluctuation (RMSF) versus residue position (averaged over three MD replicates) further demonstrates that the central helix of Gdrd, shaded, is the most conformationally rigid structure in the protein (C_α RMSF = 2–4 Å). **c** Mapping the RMSF values to a representative Gdrd MD structure for each of the three simulations shows similar regions of conformational flexibility for each replicate. **d** CD spectrum of Gdrd demonstrates characteristics typical of helical proteins. A helix minimum at 222 nm that is weaker than the helix minimum at 208 nm is characteristic of a flexible or distorted helix as was observed in the MD simulation[90].

analysis between the top-predicted Gdrd structure and the following four top structures and found that the major predicted structural features of Gdrd are conserved across all five top structures. While these all-atom RMSD values are relatively high, ranging from 11 to 13 Å, such values are reasonable considering the flexible Gdrd C-terminus. In addition, we observe a considerably lower RMSD for the core helix, for which all pairwise RMSD measurements are <3 Å (Supplementary Fig. S4)[35]. We also applied the previous structural prediction methodology to a 6x-His-tagged version of Gdrd from *Dmel* and obtained highly similar results to the untagged Gdrd protein, a finding that supports the use of tagged Gdrd for further experimental work (Supplementary Fig. S5). Using the top-predicted structure from QUARK as an input template (Supplementary Fig. S3b), we then performed three independent molecular dynamics (MD) simulations using GROMACS[36–38] (see "Methods"). Across all three simulations, the structures rapidly diverge from the input template and reach relatively high RMSD values (Supplementary Fig. S3) due to significant disorder in the termini and loop regions. However, the central helix and a portion of the N-terminus remain stably folded (Fig. 1a). A residue-by-residue RMSF analysis confirms these results, demonstrating greater rigidity in the central helix and N-terminus than throughout the rest of the protein across all three simulations (Fig. 1b, c).

In order to assess the novelty of the structures predicted for Gdrd, we used 3D-BLAST and mTM-align to compare our ab initio models of Gdrd to all structures in the Protein Data Bank (PDB)[39–41]. With both methods, we find no clear similarity to any known eukaryotic structures when searching the top five models predicted for Gdrd against the PDB—but note that Gdrd's short length results in a large number of spurious alignments of Gdrd's helix-turn motif with the helical bundles of larger, unrelated proteins.

To confirm our computational predictions, we cloned and overexpressed Gdrd in *Escherichia coli* (strain BL21 Star(DE3)). We note that expression attempts using a range of tags (maltose-binding protein, Strep-tag and the Fh8 tag) and restriction sites were unsuccessful, reminiscent of the complications encountered in expressing Bsc4[28]. For Gdrd, some fractions failed to elute, while others formed inclusion bodies that could not be further purified (see "Methods"). Only the combination of an N-terminal 6x-His-tag with purification over a nickel column yielded exploitable concentrations of >8 mg/mL soluble protein (see Supplementary Fig. S2f). Mass spectrometry (see Supplementary Data, Zenodo https://doi.org/10.5281/zenodo.4476357 and Supplementary Data 1 and Supplementary Data 2) indicates that Gdrd is monomeric in solution. Accordingly, if the helical core region does indeed form a CC, the CC does not mediate homodimerization[33]. However, this does not rule out a role for the CC in heterodimerization with other proteins or binding to small molecules. Experimental far-UV circular dichroism (CD) spectra of purified Gdrd comply with our computational

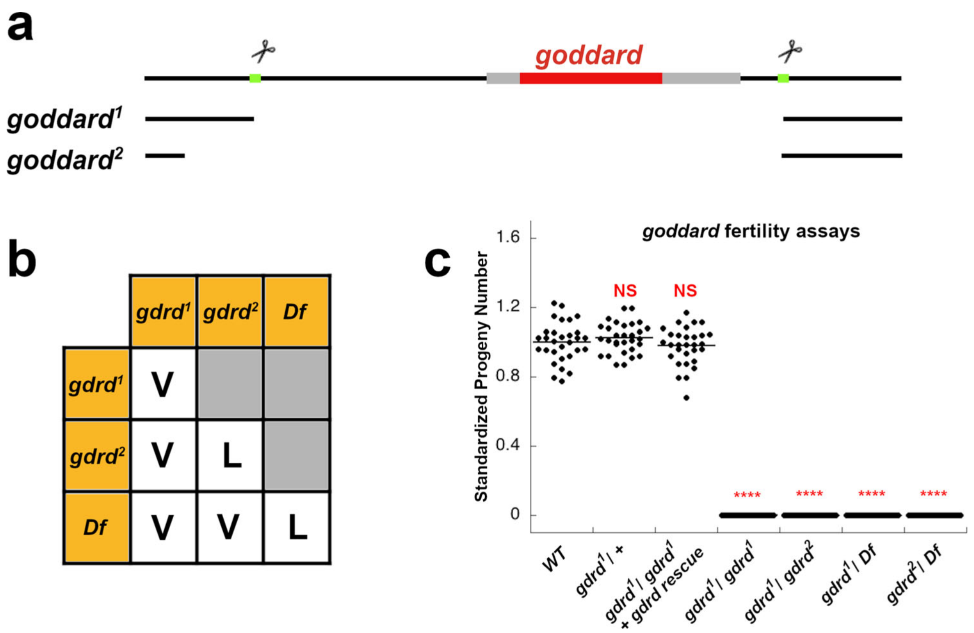

Fig. 2 *gdrd* null alleles are viable but cause male sterility. a *gdrd* genomic locus. Two CRISPR-Cas9 target sites (green) that flank the *gdrd* CDS (red) were selected to create null alleles of the gene. The *gdrd*^{1} allele results from a precise deletion at both sites, whereas the *gdrd*^{2} allele deletes an additional 153 bp upstream of the 5' target site. **b** Complementation test using the *gdrd* alleles and a large, molecularly defined deficiency *Df(3L)ED4543* that spans the *gdrd* locus shows that loss of *gdrd* has no effect on organismal viability; V = viable and L = lethal. **c** *gdrd* mutants as homozygotes or in heterozygous combinations with each other or a deficiency *Df(3L)ED4543* are all 100% sterile (single replicate; $n = 30$ for all genotypes; P value = 1.83E-30). Standardized average progeny number (progeny number/average progeny number) was 1.00 ± 0.110 (±s.e.) for WT flies (w^{1118}), 1.03 ± 0.087 (±s.e.) for *gdrd*/+ heterozygous flies (single replicate; $n = 30$; P value = 0.388), and 0.98 ± 0.111 (±s.e.) for *gdrd* rescue flies (homozygous *gdrd* mutants carrying a single copy of the *gdrd* rescue construct; single replicate; $n = 30$; P value = 0.436). For all *gdrd* loss of function backgrounds, the average progeny number was 0 ± 0 (±se). Statistical analysis: two-sample T test (two-sided), ****$P < 0.0001$. Source data are provided as a Source Data file.

predictions: the two minima at 222 nm and 208 nm indicate a flexible α-helical character, while the absence of a minimum around 218 nm suggests a lack of β-sheet regions (Fig. 1d). CD results also confirm that Gdrd is at most partially disordered since highly disordered proteins are expected to show negative ellipticity below 210 nm, with a minimum at 195 nm[42,43]. For Gdrd, the CD signal shows positive ellipticity below 200 nm. Indeed, computational interpretation of the CD spectrum using K2D3 suggests an α-helical content of 85%[44], further supporting the non-transient nature of a shorter N-terminal helix in addition to the longer helix at the core of Gdrd. Also consistent with both computational predictions (QUARK, MD) and CD spectroscopy, the ^{15}N-HMQC NMR spectrum indicates that a large region of Gdrd is partially disordered, as indicated by broadened or missing peaks, with the remainder of the protein showing a spectrum characteristic of a low-diversity fold, as would be expected for a primarily helical rigid region (Supplementary Fig. S6). Last, we performed a thermal unfolding experiment (thermal shift assay, TSA) using SYPRO orange (Supplementary Fig. S7). Consistent with CD and NMR, the TSA of Gdrd shows a thermal unfolding transition at 47.3 ± 0.9 °C, indicating that Gdrd possesses a fold that can be denatured, exposing additional dye-binding sites. Characterization by fluorescence and refolding experiments were not performed due to the lack of tryptophan or cysteine residues, and β-sheet content.

***gdrd* is required for male fertility**. Having examined the structure of Gdrd, we next attempted to clarify its "physiological implications" and "interactions"[31]. Initial phenotypic characterization of *gdrd* using testis-specific knockdown revealed that *gdrd* was required for the production of mature sperm[30]. This initial finding suggested that *gdrd* functions during spermatogenesis. To better understand the functional role of *gdrd* in fertility and bypass any potential drawbacks of RNAi, such as partial gene knockdown, we generated null alleles using CRISPR-Cas9 genome editing[45]. Using two guide RNAs to simultaneously target loci 627 bp upstream of the *gdrd* start site and 278 bp downstream of the stop codon, we generated two independent alleles, both of which completely deletes the Gdrd protein-coding region (Fig. 2a).

The *gdrd*^{1} allele deletes a 1.27-kb segment of the genome, reflecting a deletion generated by precise double-stranded cuts at the targeted sites (Fig. 2a). A second allele, *gdrd*^{2}, removes an extra 153 bp upstream of the 5' target site, forming an even larger deletion (Fig. 2a). We found that *gdrd*^{1} homozygous mutant flies are viable, while *gdrd*^{2} homozygous mutant flies are lethal (Fig. 2b). Flies carrying either mutant allele in combination with a molecularly defined deficiency that spans the *gdrd* gene or in combination with each other are, however, viable, suggesting that the chromosome bearing the *gdrd*^{2} mutation also has a second site lethal mutation not associated with the *gdrd* gene. We further confirmed that *gdrd*^{1} specifically affects the *gdrd* locus by analyzing the mRNA levels of both *gdrd* and a neighboring locus, *CG5048* (Supplementary Fig. S8). Altogether, these data suggest that *gdrd* is not required for organismal viability, consistent with the observation that the gene is primarily expressed in the testis.

We next used various combinations of *gdrd* mutant alleles and the defined deficiency line to replicate the results of the previous RNAi-based fertility assay[30]. The previous assay showed that depletion of *gdrd* transcripts within the male germ cell lineage causes complete sterility. Likewise, our *gdrd* alleles, either as homozygotes or in combination with each other or a deficiency, are also fully sterile, suggesting that our mutations are functionally null alleles (Fig. 2c).

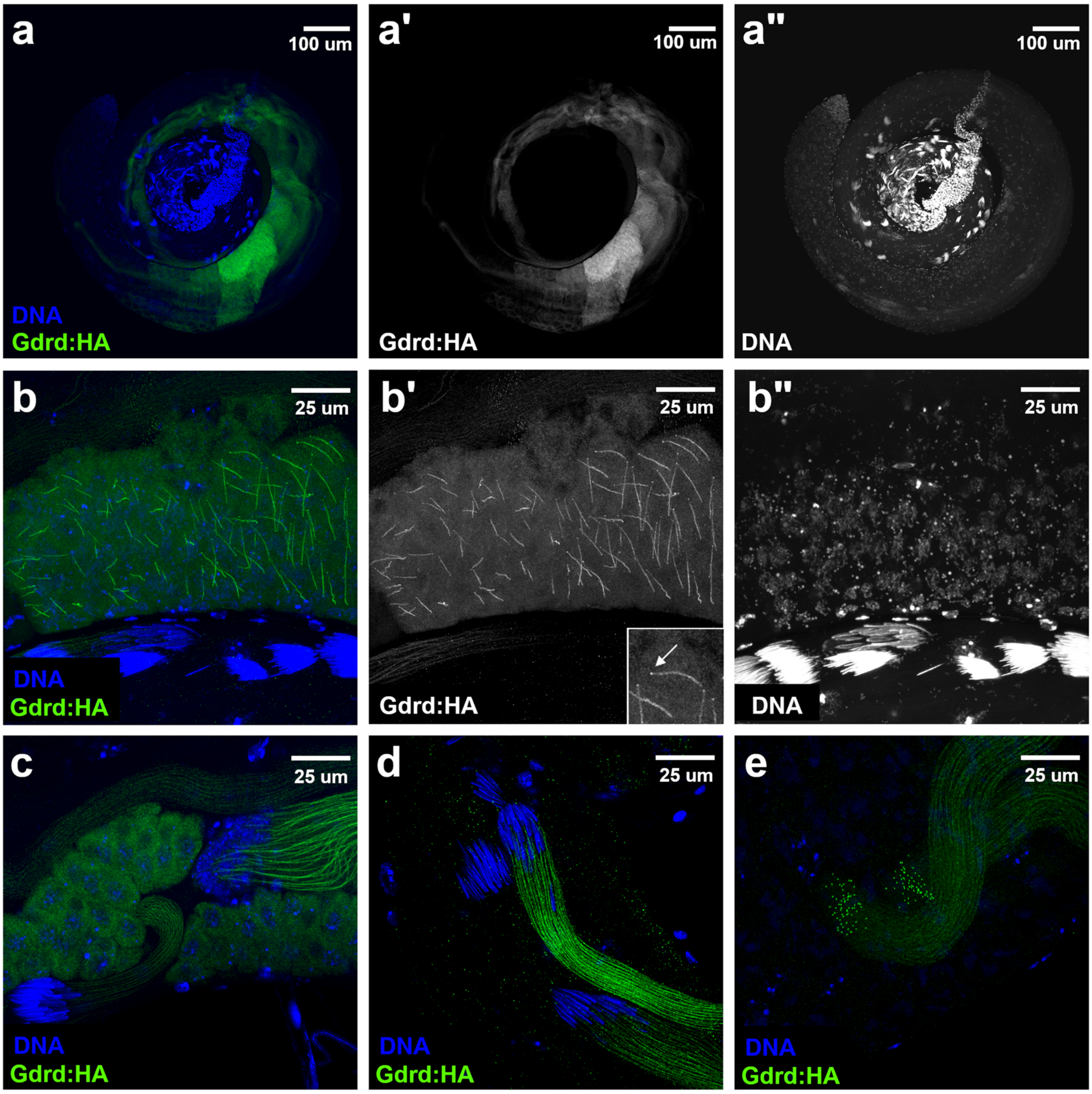

Fig. 3 Gdrd protein expresses during spermatid elongation and localizes to the growing axoneme. a–**a**″ Gdrd protein expression turns on during spermatid elongation. HA-tagged Gdrd protein expressed from a functional rescue construct in a *gdrd* mutant background is present weakly in mature spermatocytes and peaks in the early spermatids. Gdrd expression, however, is not present at the basal end of the testes, indicating that the protein is not present in individualizing or mature spermatids. **b**–**b**″ Gdrd localizes to the growing axoneme during early stages of spermatid elongation. Inset in (**b**′) shows that Gdrd localizes to the axoneme as well as a distally located punctate structure (arrow) that is reminiscent of proteins that localize to the ring centriole. **c** Gdrd localization in spermatid bundles with round nuclei (upper right) or canoe-shaped nuclei (lower left) becomes exclusive to the axonemes. **d** Spermatid bundles at the basal end of the testes with canoe-shaped nuclei express Gdrd, whereas later-staged spermatid bundles (with needle-shaped nuclei) lose Gdrd expression. **e** Distal ends of two elongating spermatid bundles. Gdrd localizes to both the growing axonemes and distally located punctate structures. The above images were derived from two independent replicates.

Gdrd is expressed in elongating spermatid cysts and associates with growing axonemes. To investigate the expression pattern and subcellular localization of the Gdrd protein in vivo, we generated a hemagglutinin (HA)-tagged *gdrd* genomic rescue construct containing both the Gdrd-coding region and its upstream and downstream regulatory elements. This construct, when placed into a *gdrd* mutant background, is sufficient to restore fertility to wild-type levels, indicating that the construct produces functional Gdrd proteins at sufficient levels and in the correct spatiotemporal pattern during spermatogenesis (Fig. 2c). Using an antibody (AB) that recognizes the HA epitope, we then visualized HA-tagged Gdrd proteins in whole-mount testes. Gdrd expression most likely starts in mature spermatocytes and peaks in early spermatids (Fig. 3a–a”). Interestingly, Gdrd expression is not observed at the basal end of the testis, where mature individualized sperm are present, indicating that Gdrd is transiently expressed primarily during sperm-tail-elongation stages of spermatogenesis (Fig. 3a–a”).

An analysis of the protein’s subcellular localization indicates that Gdrd is primarily cytoplasmic. In mature spermatocytes, the protein shows a nuclear exclusion pattern, though a low level of protein may exist in the nucleus as well (Fig. 3a–a”). Later during spermatogenesis, at the onset of sperm-tail elongation, the localization of the Gdrd protein starts to shift from the cytosol to the growing axoneme (Fig. 3b–b”). Indeed, in highly elongated spermatid bundles, Gdrd is undetectable in the cytosol and is

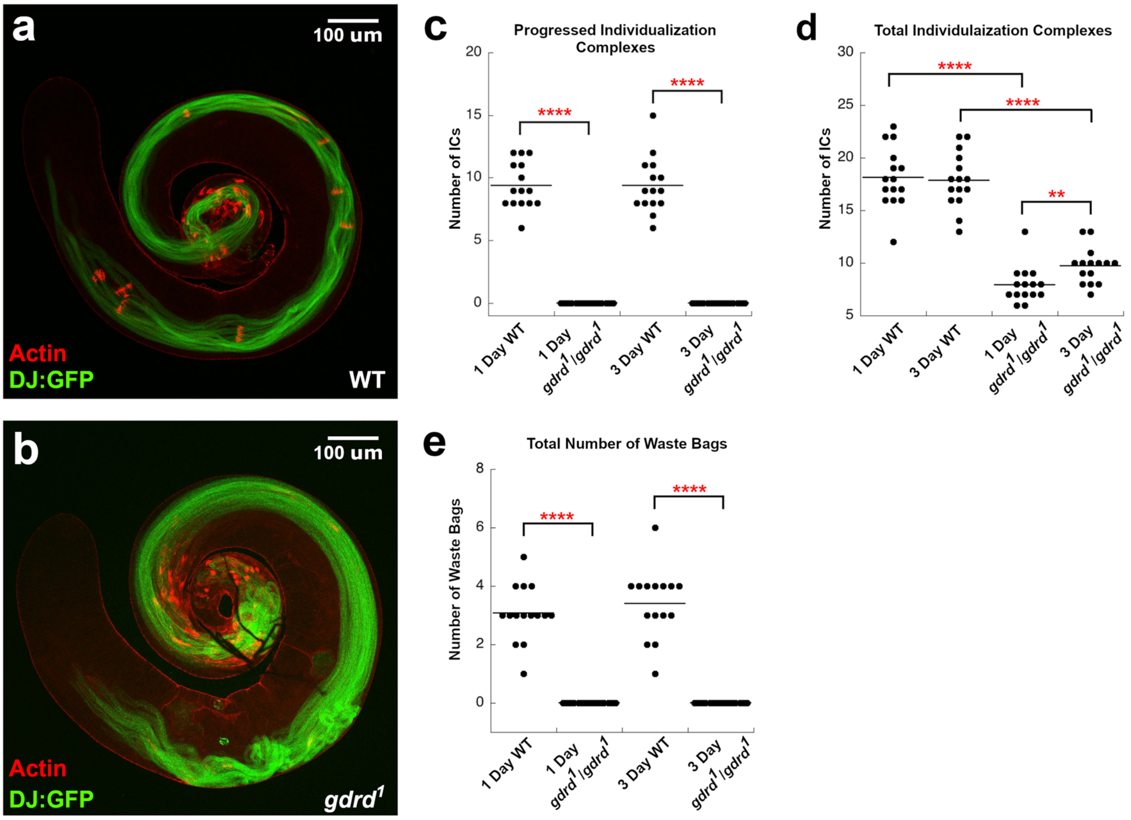

Fig. 4 *gdrd* mutant sperm fail to undergo individualization. a Wild-type testis (w^{1118}/*Y*; *dj*: *GFP*/+) and **b** *gdrd* mutant testis (w^{1118}/*Y*; *dj*: *GFP*/+; $gdrd^1$/$gdrd^1$) were labeled with phalloidin to mark actin-rich individualization complexes (ICs). Dj:GFP (green) marks elongated/elongating sperm. Scale bars in (**a**) and (**b**) = 100 µm. **c** ICs assemble but fail to progress in *gdrd* mutant testes. The analysis was conducted on day 1 post eclosion and on day 3 post eclosion. An average number of progressed ICs was 9.4 ± 0.48 (±s.e.) and 9.4 ± 0.54 (±s.e.) in WT day 1 and day 3 testes, respectively ($n = 15$). The average number of progressed ICs was 0 ± 0 (±s.e.) and 0 ± 0 (±s.e.) in *gdrd* day 1 and day 3 testes, respectively (single replicate; $n = 15$; *P* values = 3.12E-7 and 3.26E-7). **d** Assembly of ICs is decreased in *gdrd* mutant testes. The average number of total ICs was 18.1 ± 0.74 (±s.e.) and 17.9 ± 0.68 (±s.e.) in WT day 1 and day 3 testes, respectively ($n = 15$). The average number of total ICs was 7.9 ± 0.44 (±s.e.) and 9.7 ± 0.44 (± s.e.) in *gdrd* day 1 and day 3 testes, respectively (single replicate; $n = 15$; *P* values = 1.85E-6 and 1.81E-6). **e** Waste bags are absent in *gdrd* mutant testes. The average number of waste bags was 3.1 ± 0.25 (±s.e.) and 3.4 ± 0.31 (±s.e.) in WT day 1 and day 3 testes, respectively ($n = 15$). The average number of waste bags was 0 ± 0 (±s.e.) and 0 ± 0 (±s.e.) in *gdrd* day 1 and day 3 testes, respectively (single replicate; $n = 15$; *P* values = 2.55E-7 and 2.75E-7). Statistical analysis: Mann–Whitney *U* test (one-sided), $^{**}P < 0.01$, $^{****}P < 0.0001$. Source data are provided as a Source Data file.

exclusively associated with growing axonemes (Fig. 3c–d). During these later stages of elongation, the initially round spermatid nuclei undergo a morphological change, stretching first into a canoe shape and then finally into a thin, needle-like structure[46]. Using these shape changes to developmentally stage spermatid cysts, we find that there is a gradual reduction in the intensity of Gdrd staining in spermatid cysts with canoe-shaped nuclei when compared to less mature cysts with round nuclei (Fig. 3c–d). This may reflect either the active degradation of Gdrd or the titration of the protein as cyst size and axoneme length increase. We also find that elongating spermatid bundles at the needle stage no longer have Gdrd expression (Fig. 3c–d).

Besides this axonemal localization, we also observe a punctate Gdrd-positive structure at the distal end of each growing axoneme during both early and late spermatid elongation (Fig. 3b–b’, e). The localization pattern is highly reminiscent of proteins present in the ring centriole/transition zone, a site important for axoneme growth and remodeling[47,48].

***gdrd* is required for sperm individualization**. We next determined the stage at which *gdrd* is required during spermatogenesis. We crossed into the $gdrd^1$ background a transgene expressing Don Juan:GFP (Dj:GFP), a sperm-tail marker that is activated during the early stages of sperm-tail elongation[49]. Dj:GFP-positive elongating/elongated sperm tails are present in both wild-type and *gdrd* mutant testes, suggesting that post-meiotic spermatids are, at the very least, able to undergo sperm-tail elongation (Fig. 4a, b). We also find that sperm coiling at the basal end of the testis occurs normally in both wild-type and *gdrd* null flies, indicating that the absence of sperm in the seminal vesicle is not due to a defect in sperm packaging (Fig. 4a, b).

Simultaneously, we tested whether loss of *gdrd* affects sperm individualization. Individualization complexes (ICs) assemble normally at the apical end of nuclear bundles in *gdrd* mutant testes, but they fail to travel down the length of the sperm, indicating a failure in sperm individualization (Fig. 4a, b). This defect occurs in both 1-day-old and 3-day-old mutants (Fig. 4c). Furthermore, *gdrd* mutant testes show a major reduction in the total number of ICs, suggesting either a delay in IC formation or that spermatogenesis has halted in the tissue due to failure of IC translocation in older cysts (Fig. 4d). We next quantified the number of nuclear bundles associated with ICs. We find that 57% of wild-type nuclear bundles are associated with investment cones while only 23% of nuclear bundles are undergoing individualization in *gdrd* mutant testes (*z* test, $P < 0.00001$). Finally, we found that waste bags, extraneous cytoplasm culled during

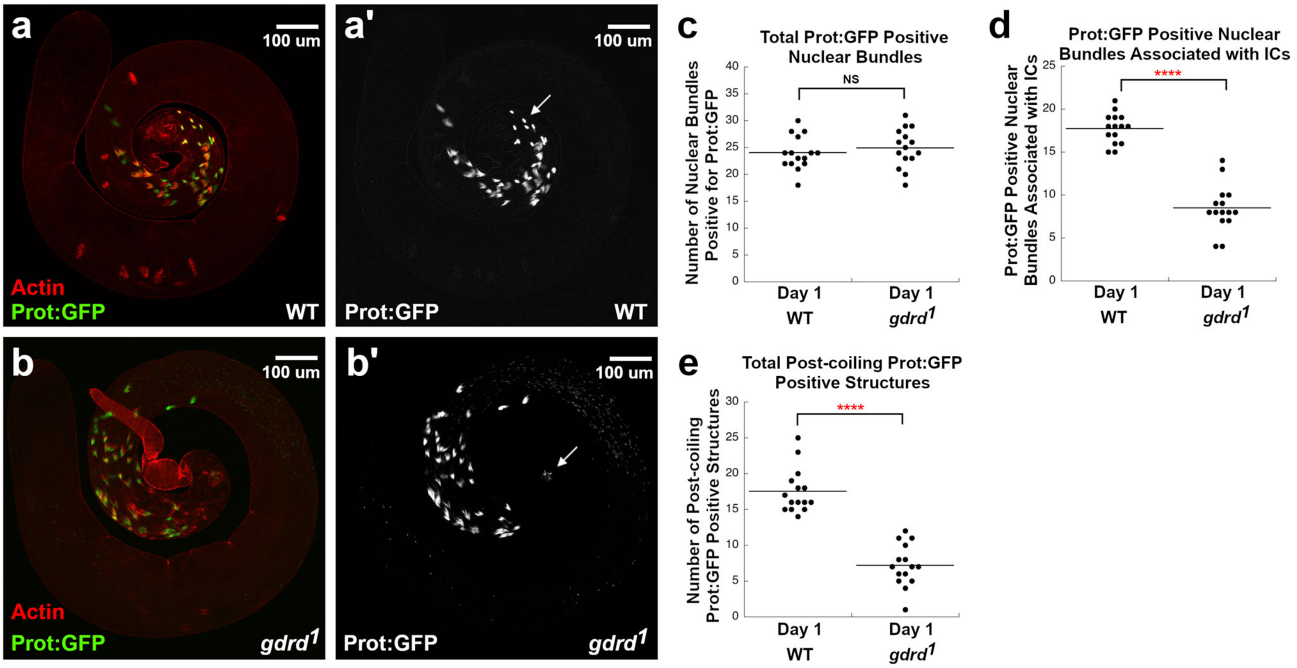

Fig. 5 Nuclear compaction is normal in *gdrd* mutant testis, but sperm bundles are potentially targeted for degradation post-coiling. **a** Wild-type testis (*prot:GFP/Y; +/+*) and **b** *gdrd* mutant testis (*prot:GFP/Y;gdrd^1/gdrd^1*) labeled with phalloidin (red) to mark actin-rich individualization complexes (ICs). Prot:GFP (green) labels protaminated nuclear bundles undergoing DNA compaction. **a**', **b**' Protamination of nuclei at the basal end of *gdrd* mutant testis is unaffected. At the very basal end of nuclei, compact protamine-positive structures associated with coiled sperm (arrows) are present in wild-type (**a**') testes but are mostly absent in *gdrd* mutant testes (**b**'). In *gdrd* mutant testis (**b**'), protamine-positive remnants are present indicating that the bundled sperm may have undergone degradation. **c** Total number of protamine-positive nuclear bundles is unaffected in *gdrd* mutants indicating that nuclear compaction is unaltered in *gdrd* mutants. The average number of protamine-positive nuclear bundles was 24 ± 0.80 (±s.e.) and 24.9 ± 0.93 (±s.e.) in WT and *gdrd* day 1 testes, respectively (single replicate; $n = 15$; P value = 0.815). **d** Association between protamine-positive nuclear bundles and individualization complexes is decreased in *gdrd* mutant testis. The average number of protamine-positive nuclear bundles associated with actin was 17.7 ± 0.45 (±s.e.) and 8.5 ± 0.70 (±s.e.) in WT and *gdrd* day 1 testes, respectively (single replicate; $n = 15$; P value = 1.53E-6). **e** The number of coiled sperm bundles is decreased in *gdrd* mutant testis, indicating that these structures are targeted for degradation. The n value for each sample was 15. The average number of post-coiling protamine-positive structures was 17.5 ± 0.80 (±s.e.) and 7.2 ± 0.76 (±s.e.) in WT and *gdrd* day 1 testes, respectively (single replicate; $n = 15$; P value = 1.60E-6). Statistical analysis: Mann–Whitney U test (one-sided), ****$P < 0.0001$. Source data are provided as a Source Data file.

individualization, are completely absent in both 1- and 3-day-old *gdrd* null testes, consistent with our observation that the individualization complexes fail to move down the length of the elongated spermatids (Fig. 4e). Taken together, these data show that in the absence of *gdrd*, sperm individualization can sometimes initiate but consistently fails to complete.

Another feature of late spermatogenesis is nuclear condensation, whereby histones are stripped from chromatin and replaced with protamines, thus allowing for the compaction of the paternal genome[50,51]. In *Dmel*, nuclear condensation is associated with nuclear reshaping, a process that is coordinated with both sperm elongation and IC assembly[52]. We observe that protamination of nuclear DNA occurs at the basal end of the testis in both wild-type and *gdrd* mutants (Fig. 5a, b), and that mutant testes and controls showed no significant difference in the number of nuclear bundles marked with protamine-GFP (Fig. 5c). Consistent with our findings above, we do, however, observe a decrease in the number of protamine-positive nuclear bundles associated with ICs (Mann–Whitney U test, P value < 0.00001), further indicating that individualization is lost in *gdrd* mutant testes Fig. 5d). We also noted a significant (Mann–Whitney U test, P value < 0.00001) decrease in fully condensed, protamine-GFP-positive sperm bundles at the far basal end of the testis, suggesting the possibility that some mutant sperm are targeted for destruction (Fig. 5e). Consistent with this idea, we frequently observed protamine-GFP-positive nuclear remnants at the very basal end (Fig. 5a, b). This finding is consistent with the observation that there are no mature sperm in the seminal vesicle.

Structural properties of Gdrd have changed little since its birth and are conserved across the *Drosophila* genus. After looking into the possible structure of Gdrd and its function within *Dmel*, we next attempted to understand (i) if the properties of Gdrd have changed since its emergence and if so, (ii) how these changes might have influenced its function. Accordingly, we compared the structure of *Dmel* Gdrd to its orthologs from four *Drosophila* species: *D. ananassae* (*Dana*), *D. virilis* (*Dvir*), *D. mojavensis* (*Dmoj*), and *D. grimshawi* (*Dgri*) (Fig. 6 and Supplementary Fig. S9). We showed previously that Gdrd orthologs in the first three species listed are male-specific and testis-biased in expression, supporting a conserved role in male fertility[30]. Alignment of these sequences (Fig. 6 and Supplementary Fig. S9) demonstrates a conserved helical core between residues 40 and 79 of Gdrd, with an average pairwise sequence similarity of 27.6% across the whole alignment and 51.6% within the core α-helix (excluding *Dgri*). To investigate the origins of this conserved core helix, we carried out ancestral sequence reconstruction (ASR) of Gdrd using orthologs from across the *Drosophila* clade (see "Methods"), and predicted the structures of the most likely ancestral sequences of (i) *Dmel/Dana*, (ii) *Dvir/Dmoj/Dgri*, and (iii) the common ancestor of all five species using QUARK (Fig. 6). Interestingly, the ancestor of *Dmel* and *Dana* is very similar to the two extant

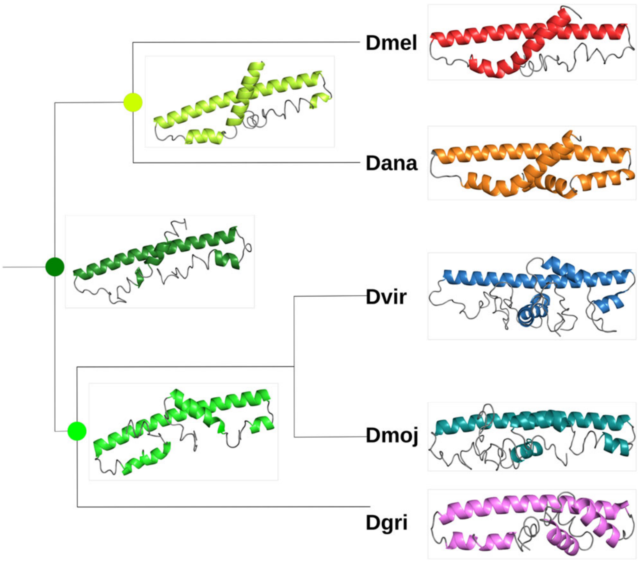

Fig. 6 Structure prediction on the ancestral reconstruction of *gdrd* and its orthologs using QUARK. Orthologs are from *Dmel*, *Dana*, *Dvir*, *Dmoj*, and *Dgri*. Additionally, predictions for the most likely sequences for reconstructed ancestors of *Dmel*/*Dana* (bright green), *Dvir*/*Dmoj*/*Dgri* (green), and their most recent common ancestor (dark green) are shown. Helices are shown with a different color in each species. PyMOL[91] was used to make protein cartoon structures (branch lengths are not meaningful).

structures (115 aa, sequence similarity 78%). We repeated predictions of hydropathy, aggregation propensity, folding, and structural properties for all additional extant sequences and the three ancestors, with results almost identical to those of extant *Dmel* Gdrd (Supplementary Figs. S1, S9, and S10). Compared to Gdrd of *Dmel* and *Dana* the extended N- and C-terminal regions in *Dgri*, *Dmoj*, and *Dvir* are predicted to display shorter helices in addition to the core α-helix. Interestingly, the ancestral central α-helix is predicted to be 10 aa shorter (28 vs. 38 aa) than the extant ones, suggesting that this helix has been gradually extended by the accumulation of helix-stabilizing mutations.

Taken together, these results suggest that (i) the initial structure of Gdrd already featured a core α-helix upon gene emergence which (ii) was gradually extended during its early evolution but (iii) was largely unmodified over the last circa 15 million years, and (iv) terminal extensions forming short helices have been added in some but not all species.

Discussion

Since their initial discovery more than a decade ago[7,8,53], de novo genes and the mechanisms underlying their emergence have been studied intensely. However, concerns regarding the reliability of their computational identification[54–56], and if and how they code for functional proteins, have also been raised. We showed previously[30] that *gdrd* is likely a de novo evolved gene, supported by its absence from the syntenic regions of outgroup species, the lack of detectably similar proteins in any other taxa, its intronic location, and its high level of intrinsic disorder. Here, we combined multiple approaches, including computational phylogenomic and structure predictions, experimental structural analyses, and cell biological assays, to further understand Gdrd's structure, evolution, and importance in spermatogenesis.

In deducing the structure of Gdrd (based on ab initio protein structure prediction and MD, in combination with CD and NMR), we further confirmed the protein's likely de novo origin, as structure-based homology searches using our models for Gdrd against all structures in the PDB detected no significant similarity in *Drosophila* and any other known eukaryotic species. We next observed that there is no transposable element (TE) nearby to the *gdrd* locus (see "Methods"). A nearby TE remnant would indicate that *gdrd* could potentially be a strongly diverged transposed duplicate of another protein-coding gene, which can also not be found in the genomes of outgroup species due to the disrupted synteny. Emergence from an intronic region thus remains the most plausible mechanism for the emergence of *gdrd*.

How did the *gdrd* gene come to acquire its essential function in *D. melanogaster*? We can gain clues from its structure and its generally high degree of evolutionary conservation within the genus. We found that the inferred ancestral form of the Gdrd protein has several intrinsically disordered (ID) regions, in addition to a helical folding core with a high predicted CC propensity. Rather than impeding protein functionality, ID is now recognized as an important structural feature that can mediate binding to a wide range of biomolecules and is occasionally essential for protein function[57]. Accordingly, the disordered termini of Gdrd may have helped the protein gain further interactions. Otherwise, Gdrd resembles what is believed to be a functional protein with rather average biochemical properties given that it has a folded core (supported by MD, CD, TSA, and NMR), appears to be soluble, is not involved in phase separation (see "Methods"), and is neither aggregating nor multimeric. Terminal extensions of proteins with ID via loss and replacement of start or stop codons have been described before[58,59] and may evolve into conserved stretches with domain-like properties over even longer time scales[17,60]. Tretyachenko et al.[61] have also demonstrated that both ID and secondary structure can emerge from random sequences, much as de novo genes do. Likewise, ID may, to an extent, counter maladaptive aggregation[61]. Given that the protein properties of de novo proteins show a high degree of overlap with those of conserved and foldable proteins, it is therefore plausible that de novo proteins with properties such as (or similar to) Gdrd emerge from intergenic or intronic regions without prior adaptation. Based on its conserved structure (described here) and conserved male-biased expression pattern[30], we hypothesize that Gdrd was likely functional at or shortly after its emergence at the base of the *Drosophila* phylogeny, at least with respect to its expression and capacities. These levels have been described as the lowest level of function of de novo gene emergence[31].

The next levels of this model describe a new protein gaining interactions with other cellular components (e.g., proteins or membranes) and the acquisition of physiological implications for a specific biological process. Based on protein expression alone, *gdrd* most likely functions during the spermatid elongation phase of spermatogenesis. The protein's expression starts in late spermatocytes and quickly peaks in early spermatids. While the initial localization of the Gdrd protein is predominantly cytosolic, the protein begins to associate with the growing axoneme at the start of spermatid elongation, suggesting a potential role for *gdrd* in regulating axonemal assembly. We also find that Gdrd expression is lost at the onset of individualization, indicating that Gdrd protein is not an integral component of the mature flagellum. Altogether, this expression pattern and localization suggest that Gdrd's physical interactions are most likely restricted with proteins expressed during spermatid elongation.

Our genetic analyses, which attempt to address the physiological role *gdrd* plays, indicate that the protein functions at or prior to spermatid individualization. Loss of *gdrd* leads to the arrest of spermatid cysts at the onset of individualization. While protamine expression and nuclear shape changes occur normally in *gdrd* mutant testes, fewer spermatid nuclear bundles are associated with ICs. This suggests that *gdrd* may be required to trigger individualization or its loss might affect some aspect of spermatid elongation itself such as axoneme growth, stability, and structure. Indeed genes that affect these processes often impair spermatid individualization[48,62].

There are several cellular events that also correlate with the onset of individualization including IC assembly, activation of caspases, and the disassembly of the ring centrioles[63–65]. Interestingly, Gdrd localizes to a structure reminiscent of the ring centriole, a specialized area at the distal end of axonemes that coordinates axoneme growth and stability[65]. This localization thus raises the possibility that *gdrd* may function in regulating this structure.

Another event that occurs during the transition from spermatid elongation to individualization is a switch from detyrosinated tubulin to polyglycated tubulin[48,66]. Gdrd likewise associates with axonemes during elongation and rapidly disappears at the onset of individualization, suggesting that Gdrd is most likely only associated with axonemes with detyrosinated tubulin, a marker of microtubule assembly. Hence, one possible avenue for future analyses will be to determine if *gdrd* functions in axoneme growth or stability. A major aspect of spermatogenesis that varies across *Drosophila* species is sperm-tail length[67,68]. While *gdrd* is not required for spermatid elongation, the gene may be required to generate or maintain long axonemes. Interestingly, we observed that the predicted structure of Gdrd protein in *Dmel* is largely unchanged from the predicted structure of Gdrd at the base of the *melanogaster* group, but differs from the predicted structure in outgroup species. One difference in sperm between the *melanogaster* group and its immediate outgroup, *obscura*, is that sperm are longer in the former set of species[67]. Thus, it is possible that Gdrd structural refinement is correlated with changes in overall sperm-tail length.

The final level of functional analysis of de novo genes is a consideration of their evolutionary implications[31]. By definition, the ancestor in which Gdrd initially arose must have had the ability to produce sperm, so Gdrd was unlikely to be required for this process at its birth. In extant *Dmel*, however, the gene is completely essential for any sperm production. Furthermore, the gene is present in all species analyzed except for *D. willistoni*, and its structure appears to be largely conserved since its origin. These data are consistent with two possibilities. First, Gdrd might have quickly evolved an essential function in late-stage spermatogenesis but became dispensable in the *D. willistoni* lineage because of lineage-specific changes to this process in the ancestor of this species. Less is known about spermatogenesis in this species, though ultrastructural studies of the process indicate that it is broadly similar to *D. melanogaster*[69]. Further mechanistic investigation of spermatogenesis in *D. willistoni* may help generate hypotheses about why Gdrd was lost specifically in this lineage, while it was retained in other divergent *Drosophila* species. Second, it is possible that in ancestors of the *melanogaster* group, Gdrd played a neutral or slightly beneficial role, consistent with its maintenance. Gdrd might have, at that point, fully integrated into the cellular interaction network, possibly via binding to other proteins, for example through the formation of a coiled-coil. Gdrd may not, however, have immediately carried out a function essential for sperm production. Such a gain of function might have arisen later, possibly modulated by changes at Gdrd's termini.

Our structural, functional, and evolutionary analyses provide novel insights into the early evolution of a putative de novo evolved gene and highlight the subtle changes it underwent as it evolved toward its current, essential role in *Dmel* spermatogenesis. Our work may therefore serve as a blueprint for future investigations into the phenomenon of de novo gene emergence and the functionalization of the proteins they encode. Our results are consistent with and complementary to several large-scale studies[11,14,16,70] that show that after their initial birth and gain of expression, many de novo proteins evolve slowly with only minor structural changes. Future studies may advance our understanding of how de novo genes evolve their functions by focusing on shorter evolutionary time scales, including population-level data, well-resolved structures, and a broader spectrum of functional conditions under which not-yet-adapted de novo proteins are accommodated by highly complex and well-established cellular networks. Such knowledge will improve our understanding of the evolution of proteins in general and may aid in devising new strategies for their design in the lab.

Methods

Computational methods

Structural prediction and homology detection. For the prediction of protein disorder and secondary structure, we used the programs s2D[71], as well as PSIPRED and Quick2D[72,73] as implemented in the MPI bioinformatics toolkit[74]. Ab initio tertiary structures were predicted using the QUARK server (https://zhanglab.ccmb.med.umich.edu/QUARK)[34,75]. The top five predicted models from QUARK were aligned using SALIGN, and RMSD values calculated using the *res_cur* command in PyMOL[76]. Kyte-Doolittle hydrophobicity was calculated with ExPASy's ProtScale[77] using a window size of 19 residues. Aggregation propensity was predicted using TANGO[78] and solubility was predicted using CamSol[79]. Phase separation was predicted using PLAAC, taking background amino acid frequencies from the *Dmel* proteome interpolated at 50% with experimental *S. cerevisea* frequencies, and a minimum domain length of 40 aa[80]. To investigate the likelihood of our predicted structures for Gdrd representing diverged forms of existing homologs, which have already been structurally solved, we took the top five predicted models from QUARK and searched for similar structures in the nr-PDB-90 database using 3D-BLAST with an E-value threshold of 1E-15[39,40]. For additional structural searches, we used the mTM-align alignment server with default settings (https://yanglab.nankai.edu.cn/mTM-align)[41].

Ancestral sequence reconstruction. Orthologs of Gdrd in other *Drosophila* species were gathered by three iterations of PSI-BLAST with an E-value threshold of 0.005[81]. Sequences from *Dvir*, *Dmoj*, and *Dgri*, previously identified by Gubala et al.[30], were subsequently included. T-COFFEE v8.97[82] was used to carry out sequence alignment using default settings. A species tree was downloaded from timetree.org, and RAxML v8.2.11 was used to carry out the ancestral sequence and gap reconstruction (RAxML command *-f A*)[83]. Sequences were reconstructed under the PROTGAMMAJTT model, and gaps were reconstructed separately under the BINCAT model, following the methods described by Aadland et al.[84]. In both cases, the most probable ancestral sequence states were computed using RAxML before being combined into a single gapped alignment.

MD simulation. MD simulations were performed using GROMACS 2018.1[36–38] using the top-predicted Gdrd structure from the QUARK webserver as an input structure. Structures were prepared following the standard procedure outlined in the GROMACS manual and tutorial. Prior to simulation, the structure was solvated in a cubic box of SPC/E water with 10-Å clearance and the electrostatic charge neutralized by the addition of sodium atoms, followed by energy minimization and equilibration in GROMACS. Three 200 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. Simulation trajectories were analyzed using GROMACS and the VMD package[85].

Experimental methods. A table of all primers used can be found in the supporting information (Supplementary Table S1).

In vivo tests of Gdrd

Fly stocks. Flies were raised at 25 °C on standard media. Fly stocks: w^{1118}, *dj-GFP./CyO* (BL5417), *Df(3L)ED4543* (BL8073), and *Vas-Cas9* (BL51323) were obtained from the Bloomington *Drosophila* Stock Center. *ProtB-GFP* (*Mst35b-GFP*[86]) was used to construct the *Prot:GFP*, +/+ and *Prot:GFP*; *gdrd*^{1}/*TM3* lines used in this paper.

CRISPR-Cas9. 5′- and 3′-flanking CRISPR-Cas9 gene-editing target sites, GGTGGAACGGGTGGACGGAATGG and CCAAACTTGCTTTCATTCGGTCC respectively, were identified using CRISPR Target Finder[87]. Guide RNAs were constructed by cloning annealed primers into pU6-3-gRNA vector (*Drosophila* Genomics Resource Center; Kate O'Connor-Giles). We then used the co-CRISPR technique as described in Ge et al.[88] to generate and screen for mutations at the *gdrd* locus (Rainbow Transgenics).

Fertility assay. Single virgin males were collected and aged for 6 days before they were mated individually to three Canton-S females. Both males and females were removed after 3 days of mating. Progeny number was determined by counting the number of pupal cases on the side of each vial 10 days after setting the cross.

Gene expression. RNA preps of w^{1118} and *gdrd* mutant flies were prepared using TRIzol (Life Technologies), followed by RQ1 DNase treatment (Promega) and cDNA synthesis using the SmartScribe kit (Clontech) and oligo-dT primers. The following primer pairs were used to detect each gene: *Gdrd* RT F/R; *CG5048* RT F/R; *RPL32*RT F/R.

Transgene. We used Gibson Assembly (NEB) to generate the HA-tagged Gdrd rescue construct. Putative upstream regulatory regions *gdrd* CDS and putative downstream regulatory regions were PCR amplified using Q5 High Fidelity Polymerase (NEB) and the Gdrd Rescue F1/R1 and Gdrd Rescue F3/R3 primer pairs (Supplementary Table S1), respectively. The 3X HA tag was amplified using pTWH (*Drosophila* Genomics Resource Center; T. Murphy) using Gdrd Rescue F2/R2 primers. PCR fragments were then assembled into a XbaI/AscI-linearized w +attB plasmid (Sekelsky, Addgene plasmid 30326). Tagged rescue construct was then phiC31 integrated into the PBac{y^{+}-attP-9A}VK00020 (BL24867) docking site (Rainbow Transgenics).

Immunostaining, phalloidin labeling, and microscopy. Testes were dissected in PBS, fixed for 20 min in 4% paraformaldehyde diluted in PBS, and subsequently permeabilized with PBX (PBS with 0.1% Triton-X). HA-tagged Gdrd was detected using rabbit anti-HA (Cell Signaling Technology, C29F4) diluted at 1:100 in PBX + 5% normal goat serum. Following overnight incubation in primary AB, samples were washed with PBX and then incubated with anti-rabbit Alexa 488 conjugated secondary antibody diluted 1:200 in PBX + 5% NGS (Life Technologies, A11008). Secondary AB has washed away with PBX. Actin-based structures were visualized by incubating fixed samples for 2 h with TRITC-conjugated phalloidin (1:200; Molecular Probes) diluted in PBX. Nuclear DNA was visualized by incubating tissues with ToPro-3 Iodide (1 mM solution diluted to 1:1000; Invitrogen) for 15 min followed by PBS washes. Samples were mounted in Vectashield mounting medium (Vector Laboratories). Images were acquired using a Leica SP5 confocal microscope (CTR6000; Leica Microsystems) with its accompanying software using N PLAN 20.0 × 0.40 DRY, HCX PL APO CS 40.0 × 1.25 OIL UV, and HCX APO CS 63.0 × 1.40 oil UV objectives (Leica Microsystems). Images were processed and analyzed using ImageJ Fiji (version 1.0)[89]. Data were compiled into Microsoft Excel for Mac (version 16.16.27) for statistical analysis and graphed using Kaleidagraph (version 4.1.3; Synergy). Mann–Whitney *U* test and student two-sample *t* tests were used to determine *P* values.

Expression, purification, and structural analysis

Cloning of Gdrd. We used genomic DNA from the Canton-S wild-type strain of *Dmel* for PCR to amplify the Gdrd sequence (FlyBase CG13477; for primer see Supplementary Table S1). The forward primer contains a BamHI and the reverse primer a HindIII cleavage site. As a stop codon, we used TAA. We digested the PCR product with both restriction enzymes (FastDigest, Thermo Scientific) for 3 h at 37 °C. As vector, we used the pHAT2 vector from the EMBL vector database, Heidelberg. This vector contains an N-terminal 6x-His-tag and the restriction sites mentioned above. We used the same procedure for digestion of the vector (1 h, 37 °C) and purified the cleaved vector from agarose gel. After purification of both vector and insert with the purification kit from Zymo Research, we ligated both with an insert:vector ratio of 4:1 using T4 ligase (Thermo Scientific, 1 h, 22 °C). The ligation mix was purified again (Zymo Research), and 2 µL of the purified reaction mix was added to 50 µL of chemical competent *E. coli* TOP10 cells. The cells were incubated for 30 min on ice, followed by a 90 s heat-shock at 42 °C. In total, 500 µL of LB medium (5 g of yeast extract, 6 g of tryptone, 5 g of NaCl) was added to the bacterial cell suspension and it was incubated for 1 h at 37 °C. After incubation, the cells were spread on an agar plate containing 50 µg/mL ampicillin (AMP) and incubated at 37 °C overnight.

Eight clones were picked from the plate and investigated through a colony PCR to check for the correct insert. Clones bearing the insert were incubated overnight in 5 mL of LB+AMP at 37 °C. The DNA was purified from the cells using the MiniPrep-Kit from Thermo Scientific and verified by sequencing (Microsynth, Seqlab, Germany). Finally, the correct DNA sequence was cloned into different BL21 strains (BL21(DE3), BL21 Star(DE3), T7 Express, and BL21(DE)pLysS) with the protocol mentioned above for expression.

In addition, a TEV cleavage site was cloned into the pHAT2 vector between the 6x-His-tag and target protein to remove the expression tag before NMR. Cloning and preparation of Gdrd were done as mentioned above.

Test expression and purification of Gdrd protein. To identify in which BL21 strain the protein gets expressed, we first performed text expression. We inoculated 10 mL of LB+AMP from glycerol stocks of all four BL21 cell types and let them grow until the solution got turbid (6–8 h, 37 °C). We then aliquoted the solutions into 3 × 3 mL and incubated for 30 min at different temperatures (37 °C, 28 °C, and 20 °C) before adding IPTG for a final concentration of 0.5 mM and expressing overnight.

In total, 500 µL of each cell culture was centrifuged (15,000 rpm, 2 min). Pellets were resuspended and lysed in 50 µL of BugBuster/Lysonase mix (Merck) through vortexing for 10 min. After centrifugation, the supernatant was mixed with the same volume of SDS-loading buffer. The pellet was resuspended in 5× diluted BugBuster, centrifuged and resuspended in 50 µL of SDS-loading buffer. In all, 10 µL of each fraction was loaded on an SDS-gel (200 V, 45 min) and dyed using InstantBlue. Strain and temperature showing the best results were used for large expression and purification. For pHAT2-Gdrd, this was BL21 (DE3) Star at 28 °C and for pHAT2-TEV-Gdrd, BL21 Star(DE3) at 20 °C.

We also tried to use MBP, Strep, Strep-Fh8, and C-terminal 6x-His-tags. None of these variants lead to soluble protein. Either the protein was not expressed at all or packed directly into inclusion bodies, from which refolding was also not successful.

For an expression of a larger amount of Gdrd a pre-culture of 5 mL 2xYT (10 g of yeast extract, 12 g of tryptone, 5 g of NaCl) + AMP was inoculated from glycerol stock and incubated at 37 °C overnight. This culture was added to 1 L of 2xYT + AMP and incubated at 37 °C until an OD$_{600}$ of 0.4–0.6 was reached. The culture was incubated for an additional hour at the appropriate temperature (20 or 28 °C, respectively) before IPTG was added to a final concentration of 0.5 mM and expression was done overnight. Cells were harvested via centrifugation (6000 rpm, 15 min, 4 °C) and resuspended in buffer A (20 mM phosphate, pH 7.2, 150 mM NaCl, 15 mM imidazole) and EDTA-free Protease Inhibitor cocktail was added (Roche). Cells were lysed using ultrasound (20 s 60% burst, 1 min pause, six cycles) and centrifuged for 30 min at 12,000 rpm at 4 °C to separate from cell debris. The supernatant was filtered through a 0.45-µm syringe filter, and a sample was taken for SDS-gel (S). The cell-free extract was loaded onto a HiPrep Ni^{2+} column (GE Healthcare) using an ÄKTA start. The flow-through was collected, and a sample was taken for SDS-Gel (FT). The column was washed with five column volumes of buffer A (sample W for SDS-Gel). The protein was eluted from the column using a 50% gradient of buffer B (buffer A + 250 mM imidazole). The eluted protein was fractionalized and analyzed by SDS-PAGE. The appropriate protein bands were combined, and the protein identity was verified by mass spectrometry (see online material Zenodo https://doi.org/10.5281/zenodo.4476357, and Supplementary Data 1 and 2.

pHAT2-TEV-Gdrd was purified the same way and the proteins were pooled. TEV protease was added to the protein (final concentration 1 mg/mL) and cleavage took place overnight at room temperature. After cleavage, the buffer was exchanged with buffer C (20 mM phosphate, pH 7.2, 150 mM NaCl) to remove imidazole using 3 kDa Amicon centricons (Merck). The protein was then re-loaded onto the Ni column to remove the His-tagged TEV protease. However, we encountered some stability issues during the cleavage process. While incubating the protein with TEV protease Gdrd tended to precipitate quite quickly after losing the His-tag. We tried the cleavage process under different conditions (4 °C, 20 °C, low salt, high salt, shorter cleavage time, different TEV concentrations, etc.). None of them helped to prevent the protein from precipitation. As consequence, we lost around 60–80% of protein during this step.

Since the protein contains no cysteine residues, we purified Gdrd without using reducing conditions. We recovered soluble protein (8 mg/mL for tagged Gdrd). Confirmation of the correct protein mass was determined by ESI-MS and trypsin MALDI-TOF-MS.

CD measurements. The protein sample was transferred into buffer D (20 mM phosphate, pH 7.2, 150 mM NaF, chloride free) using Amicon centricons. CD spectra were conducted using a Jasco J-815 spectrometer with a Jasco PTC-348WI Peltier type temperature control system (Jasco Corp, Hachioji, Japan) at constant nitrogen flow. Far-UV CD spectra were measured with a 1-mm path length quartz cuvette. Gdrd was recorded at a concentration of 50 µM, from 190 to 260 nm with a resolution of 1 nm (50 nm/min). The final spectrum was corrected by subtracting the corresponding baseline spectrum.

To estimate the protein secondary structure from the measured CD spectra, we used the K2D3 webserver (http://cbdm-01.zdv.uni-mainz.de/~andrade/k2d3)[44].

NMR measurements. Overnight culture of 25 mL of pHAT2-Gdrd and pHAT2-TEV-Gdrd in BL21 (DE3) Star cells in 2xYT were centrifuged (6000 rpm, 5 min) and resuspended in 10 mL M9 medium (standard protocol) containing 1 mg of 15NH$_4$Cl and used to innoculate a 1 L culture of M9 medium. The purification protocol was the same as mentioned above. The protein was measured on a Bruker-Biospin 600 MHz NMR by Phil Selenko, FMP Berlin, now Weizmann Institute, Israel. The NMR spectrum collected was processed using NMRViewJ (One-MoonScientific) and peak integration was performed using TopSpin v3.5 (Bruker) using a Lorentzian lineshape fit.

Thermal shift assay. The melting point of Gdrd was determined by a thermal shift assay (TSA). Either 50 µM or 20 µM purified Gdrd protein (in buffer C) was mixed with SYPRO orange 200× diluted in dimethyl sulfoxide (DMSO) rendering a final concentration of 5% DMSO and 10× SYPRO orange. As blank buffer C was mixed with 200× SYPRO orange in DMSO. Denaturation curves were measured in 96-well plates in a Roche LightCycler 480 II using wavelengths of 465 and 580 nm for excitation and emission. A linear slope corrected sigmoid was fitted to the data and used to determine the melting temperature. For each Gdrd concentration, the measurements were performed six times (12 measurements in total) and the thermal transition temperature was calculated as the mean of all 12 measurements after blank subtraction (blank measure in triplicate) ±1 standard deviation.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

MD simulation start file, figures, and additional data (fasta files for Gdrd and all Gdrd ancestors, and mass spectrometry data for Gdrd) that support the findings of this study are available under the Zenodo https://doi.org/10.5281/zenodo.4476357. Also, a list of accession codes for publicly available datasets has been deposited on Zenodo as well as mentioned in Supplementary Table 2. Figures that have associated raw data: Figs. 1b,c, 2c, 4c–e, 5c–e, Supplementary Figs. 2f, 6, 7, and 8. Source data are provided with this paper.

Code availability

Scripts for ASR used in this study are available under the Zenodo https://doi.org/10.5281/zenodo.4476357.

Received: 9 January 2020; Accepted: 3 February 2021;
Published online: 12 March 2021

References

1. Schlötterer, C. Genes from scratch—the evolutionary fate of de novo genes. *Trends Genet.* **31**, 215–219 (2015).
2. McLysaght, A. & Hurst, L. D. Open questions in the study of de novo genes: what, how and why. *Nat. Rev. Genet.* **17**, 567–578 (2016).
3. Schmitz, J. F. & Bornberg-Bauer, E. Fact or fiction: Updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Research* **6**, 57 (2017).
4. Van Oss, S. B. V. & Carvunis, A.-R. De novo gene birth. *PLoS Genet.* **15**, e1008160 (2019).
5. Liberles, D. A., Kolesov, G. & Dittmar, K. Understanding gene duplication through biochemistry and population genetics. in *Evolution after Gene Duplication*, (eds Dittmar, K. & Liberles, D.) 1–21 (John Wiley & Sons, Ltd, 2011).
6. Bornberg-Bauer, E. & Albà, M. M. Dynamics and adaptive benefits of modular protein evolution. *Curr. Opin. Struct. Biol.* **23**, 459–466 (2013).
7. Begun, D. J., Lindfors, H. A., Thompson, M. E. & Holloway, A. K. Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. *Genetics* **172**, 1675–1681 (2006).
8. Cai, J., Zhao, R., Jiang, H. & Wang, W. De novo origination of a new protein-coding gene in *Saccharomyces cerevisiae*. *Genetics* **179**, 487–496 (2008).
9. Neme, R. & Tautz, D. Phylogenetic patterns of emergence of new genes support a model of frequent de novo evolution. *BMC Genomics* **14**, 117 (2013).
10. McLysaght, A. & Guerzoni, D. New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos. Trans. R. Soc. B Biol. Sci.* **370**, 20140332 (2015).
11. Schmitz, J. F., Ullrich, K. K. & Bornberg-Bauer, E. Incipient de novo genes can evolve from frozen accidents that escaped rapid transcript turnover. *Nat. Ecol. Evol.* **2**, 1626–1632 (2018).
12. Vakirlis, N. et al. A molecular portrait of de novo genes in yeasts. *Mol. Biol. Evol.* **35**, 631–645 (2018).
13. Prabh, N. & Rödelsperger, C. De novo, divergence, and mixed origin contribute to the emergence of orphan genes in pristionchus nematodes. *G3 Genes Genomes Genet.* **9**, 2277–2286 (2019).
14. Zhang, L. et al. Rapid evolution of protein diversity by de novo origination in Oryza. *Nat. Ecol. Evol.* **3**, 679 (2019).
15. Zhou, Q. et al. On the origin of new genes in Drosophila. *Genome Res.* **18**, 1446–1455 (2008).
16. Carvunis, A.-R. et al. Proto-genes and de novo gene birth. *Nature* **487**, 370–374 (2012).
17. Klasberg, S., Bitard-Feildel, T., Callebaut, I. & Bornberg-Bauer, E. Origins and structural properties of novel and de novo protein domains during insect evolution. *FEBS J.* **285**, 2605–2625 (2018).
18. Zhao, L., Saelao, P., Jones, C. D. & Begun, D. J. Origin and spread of de novo genes in *Drosophila melanogaster* populations. *Science* **343**, 769–772 (2014).
19. Ruiz-Orera, J., Messeguer, X., Subirana, J. A. & Albà, M. M. Long non-coding RNAs as a source of new peptides. *eLife* **3**, e03523 (2014).
20. Palmieri, N., Kosiol, C. & Schlötterer, C. The life cycle of Drosophila orphan genes. *eLife* **3**, e01311 (2014).
21. Levy, A. How evolution builds genes from scratch. *Nature* **574**, 314–316 (2019).
22. Khalturin, K., Hemmrich, G., Fraune, S., Augustin, R. & Bosch, T. C. More than just orphans: are taxonomically-restricted genes important in evolution? *Trends Genet.* **25**, 404–413 (2009).
23. Baalsrud, H. T. et al. De novo gene evolution of antifreeze glycoproteins in codfishes revealed by whole genome sequence data. *Mol. Biol. Evol.* **35**, 593–606 (2018).
24. Zhuang, X., Yang, C., Murphy, K. R. & Cheng, C.-H. C. Molecular mechanism and history of non-sense to sense evolution of antifreeze glycoprotein gene in northern gadids. *Proc. Natl Acad. Sci. USA* **116**, 4400–4405 (2019).
25. Chen, L., DeVries, A. L. & Cheng, C.-H. C. Convergent evolution of antifreeze glycoproteins in Antarctic notothenioid fish and Arctic cod. *PNAS* **94**, 3817–3822 (1997).
26. Brockhausen, I., Schachter, H. & Stanley, P. O-GalNAc Glycans. in *Essentials of Glycobiology,* second edn. (eds Varki, A. et al.) (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 2009).
27. Pan, X. et al. A DNA integrity network in the yeast *Saccharomyces cerevisiae*. *Cell* **124**, 1069–1081 (2006).
28. Bungard, D. et al. Foldability of a natural de novo evolved protein. *Structure* **25**, 1687–1696.e4 (2017).
29. Vakirlis, N. et al. De novo emergence of adaptive membrane proteins from thymine-rich genomic sequences. *Nat. Commun.* **11**, 781 (2020).
30. Gubala, A. M. et al. The goddard and saturn genes are essential for Drosophila male fertility and may have arisen de novo. *Mol. Biol. Evol.* **34**, 1066–1082 (2017).
31. Keeling, D. M., Garza, P., Nartey, C. M. & Carvunis, A.-R. The meanings of ’function’ in biology and the problematic case of de novo gene emergence. *eLife* **8**, e47014 (2019).
32. Lupas, A., Van Dyke, M. & Stock, J. Predicting coiled coils from protein sequences. *Science* **252**, 1162–1164 (1991).
33. Truebestein L, Leonard TA. Coiled-coils: The long and short of it. *Bioessays* **38**. 903–916 https://doi.org/10.1002/bies.201600062 (2016).
34. Xu, D. & Zhang, Y. Toward optimal fragment generations for ab initio protein structure assembly. *Proteins* **81**, 229–239 (2013).
35. Reva, B. A., Finkelstein, A. V. & Skolnick, J. What is the probability of a chance prediction of a protein structure with an RMSD of 6 A? *Fold Des.* **3**, 141–147 (1998).
36. Berendsen, H. J. C., van der Spoel, D. & van Drunen, R. GROMACS: a message-passing parallel molecular dynamics implementation. *Comput. Phys. Commu.* **91**, 43–56 (1995).
37. Pronk, S. et al. GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit. *Bioinformatics* **29**, 845–854 (2013).
38. Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* **1-2**, 19–25 (2015).
39. Yang, J.-M. & Tung, C.-H. Protein structure database search and evolutionary classification. *Nucleic Acids Res.* **34**, 3646–3659 (2006).
40. Tung, C.-H., Huang, J.-W. & Yang, J.-M. Kappa-alpha plot derived structural alphabet and BLOSUM-like substitution matrix for rapid search of protein structure database. *Genome Biol.* **8**, R31 (2007).
41. Dong, R., Pan, S., Peng, Z., Zhang, Y. & Yang, J. mTM-align: a server for fast protein structure database search and multiple protein structure alignment. *Nucleic Acids Res.* **46**, W380–W386 (2018).
42. Kelly, S. M., Jess, T. J. & Price, N. C. How to study proteins by circular dichroism. *Biochim. Biophys. Acta BBA - Proteins Proteomics* **1751**, 119–139 (2005).
43. Greenfield, N. J. Using circular dichroism spectra to estimate protein secondary structure. *Nat. Protoc.* **1**, 2876–2890 (2006).
44. Louis-Jeune, C., Andrade-Navarro, M. A. & Perez-Iratxeta, C. Prediction of protein secondary structure from circular dichroism using theoretically derived spectra. *Proteins* **80**, 374–381 (2012).
45. Jinek, M. et al. RNA-Programmed genome editing in human cells. *eLife* **2**, e00471 (2013).
46. Fabian, L. & Brill, J. A. Drosophila spermiogenesis. *Spermatogenesis* **2**, 197–212 (2012).
47. Basiri, M. L. et al. A migrating ciliary gate compartmentalizes the site of axoneme assembly in Drosophila spermatids. *Curr. Biol.* **24**, 2622–2631 (2014).
48. Soulavie, F. et al. Hemingway is required for sperm flagella assembly and ciliary motility in Drosophila. *MBoC* **25**, 1276–1286 (2014).
49. Santel, A., Winhauer, T., Blümer, N. & Renkawitz-Pohl, R. The Drosophila don juan (dj) gene encodes a novel sperm specific protein component characterized by an unusual domain of a repetitive amino acid motif. *Mech. Dev.* **64**, 19–30 (1997).

50. Oliva, R. & Dixon, G. H. Vertebrate protamine genes and the histone-to-protamine replacement reaction. in *Progress in Nucleic Acid Research and Molecular Biology*, Vol. 40 (eds Cohn, W. E. & Moldave, K.) 25–94 (Academic Press, 1991).
51. Jayaramaiah Raja, S. & Renkawitz-Pohl, R. Replacement by *Drosophila melanogaster* protamines and Mst77F of Histones during chromatin condensation in late spermatids and role of sesame in the removal of these proteins from the male pronucleus. *Mol. Cell Biol.* **25**, 6165–6177 (2005).
52. Tokuyasu, K. T. Dynamics of spermiogenesis in *Drosophila melanogaster*. 3. Relation between axoneme and mitochondrial derivatives. *Exp. Cell Res.* **84**, 239–250 (1974).
53. Begun, D. J., Lindfors, H. A., Kern, A. D. & Jones, C. D. Evidence for de novo evolution of testis-expressed genes in the *Drosophila yakuba*/*Drosophila erecta* clade. *Genetics* **176**, 1131–1137 (2007).
54. Moyers, B. A. & Zhang, J. Phylostratigraphic bias creates spurious patterns of genome evolution. *Mol. Biol. Evol.* **32**, 258–267 (2015).
55. Domazet-Lošo, T. et al. No evidence for phylostratigraphic bias impacting inferences on patterns of gene emergence and evolution. *Mol. Biol. Evol.* **34**, 843–856 (2017).
56. Weisman, C. M., Murray, A. W. & Eddy, S. R. Many, but not all, lineage-specific genes can be explained by homology detection failure. *PLoS Biol.* **18**, e3000862 (2020).
57. Babu, M. M. The contribution of intrinsically disordered regions to protein function, cellular complexity, and human disease. *Biochem. Soc. Trans.* **44**, 1185–1200 (2016).
58. Toll-Riera, M. & Albà, M. M. Emergence of novel domains in proteins. *BMC Evol. Biol.* **13**, 47 (2013).
59. Kleppe, A. S. & Bornberg-Bauer, E. Robustness by intrinsically disordered C-termini and translational readthrough. *Nucleic Acids Res.* **46**, 10184–10194 (2018).
60. Bitard-Feildel, T., Heberlein, M., Bornberg-Bauer, E. & Callebaut, I. Detection of orphan domains in Drosophila using "hydrophobic cluster analysis". *Biochimi* **119**, 244–253 (2015).
61. Tretyachenko, V. et al. Random protein sequences can form defined secondary structures and are well-tolerated in vivo. *Sci. Rep.* **7**, 15449 (2017).
62. Maia, T. M., Gogendeau, D., Pennetier, C., Janke, C. & Basto, R. Bug22 influences cilium morphology and the post-translational modification of ciliary microtubules. *Biol. Open* **3**, 138–151 (2014).
63. Tokuyasu, K. T., Peacock, W. J. & Hardy, R. W. Dynamics of spermiogenesis in *Drosophila melanogaster*. *Z. Zellforsch.* **124**, 479–506 (1972).
64. Arama, E., Agapite, J. & Steller, H. Caspase activity and a specific cytochrome c are required for sperm differentiation in drosophila. *Dev. Cell* **4**, 687–697 (2003).
65. Vieillard, J. et al. Transition zone assembly and its contribution to axoneme formation in Drosophila male germ cells. *J. Cell Biol.* **214**, 875–889 (2016).
66. Rogowski, K. et al. Evolutionary divergence of enzymatic mechanisms for posttranslational polyglycylation. *Cell* **137**, 1076–1087 (2009).
67. Joly, D. & Lachaise, D. Polymorphism in the sperm heteromorphic species of the Drosophila obscura group. *J. Insect Physiol.* **40**, 933–938 (1994).
68. Pitnick, S., Hosken, D. J. & Birkhead, T. R. Sperm morphological diversity. in *Sperm Biology*. (eds Birkhead, T. R. et al.) 69–149 (Academic Press, London, 2009).
69. de Almeida Rego, Ld. N. A., Alevi, K. C. C., de Azeredo-Oliveira, M. T. V. & Madi-Ravazzi, L. Ultrastructural features of spermatozoa and their phylogenetic application in Zaprionus (Diptera, Drosophilidae). *Fly* **10**, 47–52 (2016).
70. Ruiz-Orera, J. et al. Origins of de novo genes in human and chimpanzee. *PLoS Genet.* **11**, e1005721 (2015).
71. Sormanni, P. et al. Simultaneous quantification of protein order and disorder. *Nat. Chem. Biol.* **13**, 339–342 (2017).
72. Jones, D. T. Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.* **292**, 195–202 (1999).
73. Gruber, M., Söding, J. & Lupas, A. N. Comparative analysis of coiled-coil prediction methods. *J. Struct. Biol.* **155**, 140–145 (2006).
74. Alva, V., Nam, S.-Z., Söding, J. & Lupas, A. N. The MPI bioinformatics Toolkit as an integrative platform for advanced protein sequence and structure analysis. *Nucleic Acids Res.* **44**, W410–415 (2016).
75. Xu, D. & Zhang, Y. Ab initio protein structure assembly using continuous structure fragments and optimized knowledge-based force field. *Proteins* **80**, 1715–1735 (2012).
76. Braberg, H. et al. SALIGN: a web server for alignment of multiple protein sequences and structures. *Bioinformatics* **28**, 2072–2073 (2012).
77. Gasteiger, E. et al. Protein identification and analysis tools on the ExPASy server. in *The Proteomics Protocols Handbook*, (ed. Walker, J. M.) 571–607 (Humana Press, 2005).
78. Fernandez-Escamilla, A.-M., Rousseau, F., Schymkowitz, J. & Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **22**, 1302–1306 (2004).
79. Sormanni, P., Aprile, F. A. & Vendruscolo, M. The CamSol method of rational design of protein mutants with enhanced solubility. *J. Mol. Biol.* **427**, 478–490 (2015).
80. Lancaster, A. K., Nutter-Upham, A., Lindquist, S. & King, O. D. PLAAC: a web and command-line application to identify proteins with prion-like amino acid composition. *Bioinformatics* **30**, 2501–2502 (2014).
81. Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402 (1997).
82. Notredame, C., Higgins, D. G. & Heringa, J. T-Coffee: a novel method for fast and accurate multiple sequence alignment. *J. Mol. Biol.* **302**, 205–217 (2000).
83. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
84. Aadland, K., Pugh, C. & Kolaczkowski, B. High-throughput reconstruction of ancestral protein sequence, structure, and molecular function. in *Computational Methods in Protein Evolution*, *Methods in Molecular Biology*. (ed. Sikosek, T.) 63–81 (Springer New York, 2019).
85. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graphics* **14**, 33–38 (1996).
86. Manier, M. K. et al. Resolving mechanisms of competitive fertilization success in *Drosophila melanogaster*. *Science* **328**, 354–357 (2010).
87. Gratz, S. J. et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. *Genetics* **196**, 961–971 (2014).
88. Ge, D. T., Tipping, C., Brodsky, M. H. & Zamore, P. D. Rapid screening for CRISPR-directed editing of the Drosophila genome using white coconversion. *G3 Bethesda* **6**, 3197–3206 (2016).
89. Schindelin, J. et al. Fiji—an open source platform for biological image analysis. *Nat. Methods* **9**, 676–682 (2012).
90. Karlsson, E. et al. Coupled binding and helix formation monitored by synchrotron-radiation circular dichroism. *Biophys. J.* **117**, 729–742 (2019).
91. The PyMOL Molecular Graphics System, Version 1.8.4 Schrödinger, LLC. www.pymol.org.

Acknowledgements

We thank Susan Hawat (Department of Plant Biochemistry and Biotechnology, WWU Muenster) and Prof. Simone König (Core Unit Proteomics) for analyzing our proteins via mass measurements, Prof. Frank M. Boeckler (Molecular Design & Pharm. Biophysics, University of Tuebingen), and his group for providing us with their TEV-plasmid, Prof. Phillip Selenko (FMP Berlin, now Weizmann Institute, Israel) for measuring the NMR sample and helping with the analysis of the HSMC. P.H.P. and G.D.F. were supported by NSF CAREER grant #1652013 to GDF. Fly stocks were obtained from the Bloomington Drosophila Stock Center (supported by NIH grant P40OD018537), and plasmids were provided by the Drosophila Genomics Resource Center (supported by NIH grant 2P40OD010949). B.H. received funding from the EU under the Horizon 2020 Research and Innovation Framework Programme No. 722610. CJJ thanks the ARC Centre of Excellence for Innovations in Peptide and Protein Science and the ARC Centre of Excellence in Synthetic Biology for support.

Author contributions

A.L., B.H., and E.B.-B. designed structural research; A.L. and B.H. performed structural predictions and ancestral reconstruction; A.L. performed cloning, expression, purification, CD, TSA, and NMR measurements; T.S. helped with CD measurement and interpreting data; A.L., B.H., A.M.D., and C.J.J. ran MD simulations and evaluated the results; A.M.D. helped to interpret NMR data; P.H.P. and G.D.F. designed *Drosophila* experiments. P.H.P. performed genetic experiments and cytological analyses. All authors wrote and approved the final version of the paper.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-021-21667-6.

Correspondence and requests for materials should be addressed to G.D.F. or E.B.-B.

Peer review information *Nature Communications* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.