Journal of Molecular Evolution (2022) 90:2–16
https://doi.org/10.1007/s00239-021-10037-x

ORIGINAL ARTICLE

Rapid Divergence of Key Spermatogenesis Genes in *nasuta*-Subgroup of *Drosophila*

Rajanikanth Chowdanayaka[1] · Ramachandra Nallur Basappa[1]

Received: 28 July 2021 / Accepted: 9 November 2021 / Published online: 22 November 2021
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

Abstract

The crosses between closely related *Drosophila* species usually produce sterile hybrid males with spermatogenesis disrupted at post-meiotic phase, especially in sperm individualization stage than the pre-meiotic stage. This is possibly due to the rapid interspecies divergence of male sex and reproduction-related genes. Here we annotated 11 key spermatogenesis genes in 35 strains of species belonging to *nasuta*-subgroup of *Drosophila*, where many interspecies crosses produce sterile males. We characterized the divergence and polymorphism in the protein coding regions by employing gene-wide, codon-wide, and lineage-specific selection analysis to test the mode and strength of selection acting on these genes. Our analysis showed signature of positive selection at *bag of marbles (bam)* and *benign gonial cell neoplasma (bgcn)* despite the selection constrains and the absence of endosymbiont infection which could potentially drive rapid divergence due to an arms race while *roughex* (*rux)* showed lineage-specific rapid divergence in frontal sheen complex of *nasuta*-subgroup. *cookie monster* (*comr*) showed rapid divergence consistent with the possibility of meiotic arrest observed in sterile hybrids of *Drosophila* species. Rapid divergence observed at *don juan* (*dj)* and *Mst98Ca-like* was consistent with fused sperm-tail abnormality observed in the hybrids of *Drosophila nasuta* and *Drosophila albomicans*. These findings highlight the potential role of rapid nucleotide divergence in bringing about hybrid incompatibility in the form of male sterility; however, additional genetic manipulation studies can widen our understanding of hybrid incompatibilities. Furthermore, our study emphasizes the importance of young species belonging to *nasuta*-subgroup of *Drosophila* in studying post-zygotic reproductive isolation mechanisms.

Keywords Rapid divergence · Spermatogenesis genes · Post-zygotic isolation · Hybrid male sterility · Positive selection · *nasuta*-subgroup of *Drosophila*

Introduction

Speciation occurs through the evolution of reproductive isolation (Dobzhansky 1937; Coyne and Orr 2004). Reproductive isolation is said to be achieved when there are barriers that prevents two species from producing fit hybrid offspring. Among sympatric *Drosophila* species, prezygotic reproductive isolation (sexual isolation) evolves faster than post-zygotic isolation (hybrid incompatibility), whereas among allopatric *Drosophila* species, prezygotic reproductive isolation and intrinsic post-zygotic isolation evolves roughly at the same rate (Coyne and Orr 1989, 1997). Between recently evolved species, intrinsic post-zygotic isolation manifests in the form of hybrid male sterility (HMS), hybrid female sterility (HFS) and hybrid inviability (HI). Bateson, Dobzhansky and Muller independently proposed a model to explain the evolution of hybrid incompatibility that result in intrinsic post-zygotic isolation (Bateson 1909; Dobzhansky 1937; Muller 1942). Hybrid incompatibility involves a negative epistatic interaction between genes from two different species. When two species diverge from one another, they accumulate genetic substitutions that function normally within their genomic background but, can cause disruption of gametogenesis or development when brought together in a hybrid (Coyne and Orr 2004, Dobzhansky 1937).

Genes expressed in male reproductive tract are known to evolve rapidly among closely related species. In *Drosophila* a pattern of faster evolution of male-specific genes relative

Handling editor: **Willie Swanson**.

✉ Ramachandra Nallur Basappa
genomics.fly@gmail.com

[1] Department of Studies in Genetics and Genomics, University of Mysore, Manasagangotri, Mysuru 570006, India

to female and non-reproductive genes has been shown both in terms of coding sequence divergence and loss and gain of genes among distantly related species (Haerty et al. 2007). *Drosophila* spermatogenesis is a multistep process where a single progenitor germline stem cell undergoes a sequential division and morphological changes to become a mature motile sperm. Spermatogenesis in *Drosophila* can be broadly divided into four stages: germline establishment, mitotic proliferation of germline cells, spermatid formation through meiotic division and spermatid differentiation or spermiogenesis (Fuller 1998; Wakimoto et al. 2004; White-Cooper and Bausek 2010). Given the rapid divergence of male-specific genes, it is not surprising that the common outcome of crosses between closely related species is hybrid male sterility (Haldane 1922).

Cytological studies of many *Drosophila* interspecies sterile hybrid males showed the disruption of spermatogenesis at the stage of spermatid differentiation. Spermatogenesis proceeds normally until the meiotic division and spermatid formation and encounter problems in the post-meiotic stages. These problems include the lack of synchrony in spermatid development and failure of spermatid individualization wherein the interconnected spermatid bundles fail to differentiate and mature into a motile sperm (Dobzhansky 1937). In the sterile male hybrids of *D. simulans* clade (*D. simulans, D. mauritiana*, and *D. sechellia*), it was observed that spermatogenesis arrest occurs both before and after the onset of meiosis, depending on the species pair used in the interspecies cross (Kulathinal and Singh 1998; Lachaise et al. 1986).

Several genes have been identified that play major role in spermatogenesis of *Drosophila*. *Bag of marbles* (*bam*), *benign gonial cell neoplasma* (*bgcn*), *roughex* (*rux*) are involved in regulating the early stage of spermatogenesis, *bam* and *bgcn* limit the number of mitotic divisions facilitating the mitosis to meiosis transition. *Always early* (*aly*), *achintya* (*achi*), *cookie monster* (*comr*), and *spermatocyte arrest* (*sa*) which are known as the meiotic arrest genes are involved in regulating the progression into meiosis and the initiation of spermiogenesis (Fuller 1998; Jiang and White-Cooper 2003; White-Cooper 2010). Finally, *don juan* (*dj*), *JYalpha*, *Mst84Dc* and *Mst98Ca* are involved in the maturation of round spermatid into elongated spermatid during spermiogenesis (Michalak and Noor 2004; Moehring et al. 2007; Santel et al. 1997).

The *Drosophila nasuta*-subgroup of *immigrans* species offers an excellent case to study the role of rapid genetic divergence in bringing about hybrid incompatibility. It is an young cluster of species with the history of multiple speciation events in a short period resulting in morphologically very similar species with varying degrees of post-zygotic isolation (Wilson et al. 1969) and extensive chromosomal polymorphism (Hatsumi et al. 1988; Suzuki et al. 1990).The *nasuta*-subgroup consist of a dozen closely related species or subspecies that are widely distributed across South-East Asia (Kitagawa et al. 1982; Wilson et al. 1969), and the crosses between many species of this subgroup produce sterile and sometimes fertile offspring. Females of *nasuta*-subgroup of *Drosophila* are morphologically indistinguishable, whereas males can be classified into three groups based on the markings on the frons and thorax (Kitagawa et al. 1982; Wilson et al. 1969). The first category includes *D. nasuta, D. albomicans, D. kepulauana* and *D. kohkoa*, which show a continuous silvery patch on their frons and dark band on their thorax. Second category comprises of *D. pulaua, D. sulfurigaster sulfurigaster, D. sulfurigaster bilimbata, D. sulfurigaster albostrigata,* and *D. sulfurigaster neonasuta*, which have whitish patch along the edges of their compound eyes. The third category includes *D. pallidifrons, Taxon-F, I and J*, which has reduced white patch. *D. niveifrons* is an exception with an X-shaped silvery patch on their forehead. The earliest known member of the subgroup *D. niveifrons* emerged about 3.5 million years ago (Mya), later the *D. pulaua, D. s. sulfurigaster*, *D. kohkoa* and *Taxon- F* diverged from *D. niveifrons* about 2.5 Mya. Rest of the species emerged between 0.7 and 1 Mya (Yu et al. 1999).

The study of pattern of rapid divergence of genes involved in hybrid incompatibilities is potentially an important way of understanding the mechanisms of speciation. Most of the studies are conducted in species of *D. melanogaster* and other sibling species. Investigations in recently diverged species has great potential in identifying underlying evolutionary forces that drive the process of speciation in the early stages of speciation. The *nasuta*-subgroup of *Drosophila* which comprises of many young species provides an excellent model to understand the role of these evolutionary processes in the process of speciation. The *nasuta*-subgroup comprises of many species which show symptoms of post-zygotic reproductive isolation such as hybrid inviability, hybrid male and female sterility (Kitagawa et al. 1982; Wilson et al. 1969; Nirmala and Krishnamurthy 1973). The availability of genome sequences of these species provides an opportunity to understand the role of rapid divergence in bringing about hybrid incompatibilities in young species group of *Drosophila*.

The main goal of our study is to characterize the nucleotide divergence of key spermatogenesis genes likely to have been involved in the hybrid incompatibility interaction resulting in the sterility across the species of *nasuta*-subgroup of *Drosophila*. We compare the interspecies polymorphism and divergence between species that result in an inviable hybrid and also performed gene-wide, codon-wide and lineage-specific selection analysis in the phylogenetic framework.

Materials and Methods

Ortholog Search

We downloaded whole genome raw sequences of 35 strains of 11 *Drosophila* species of *nasuta*-subgroup (Table S1) from NCBI-SRA (Mai et al. 2020; Mohanty and Khanna, 2017). Genome assemblies were built using UniCycler (Wick et al. 2017), which assembled the sequences into scaffolds. We used both paired end and single data and selected normal bridging mode which allows moderate contig size and moderate misassembly rate. We excluded contigs that are shorter than 100 base pairs in the final assembly. The amino acid sequences of individual spermatogenesis genes of *Drosophila melanogaster* amino acid sequences were acquired from Flybase (Marygold et al. 2013). These sequences were used in a tBLASTn (Gerts et al. 2006) search against the *D. albomicans* assembled genomes with a liberal cut off of $E=0.1$ in order to ensure the detection of divergent orthologs. The best BLAST hit scaffold for each gene was taken and 3 kb upstream and downstream of the homologous region was extracted. We included 3 kb upstream and downstream sequences to ensure that the open reading frame is not missed during the gene prediction. The extracted DNA sequence was used for gene prediction using AUGUSTUS webserver (Stanke and Morgenstern 2005), a Generalized Hidden Markov Model (GHMM) based gene prediction tool which predicts the gene structure, coding sequences and amino acid sequences of the respective gene. We selected *D. melanogaster* as the organism of reference since that was the only species from *Drosophila* group available on AUGUSTUS. Gene orthology was confirmed by reciprocal best BLAST hit approach by blasting the predicted amino acid sequence against the annotated *Drosophila albomicans* protein database. We also checked for the presence of conserved protein domains present in the respective genes using NCBI-Common conserved Domain Database (CDD) (Last accessed:16/07/2021) to confirm the orthology. We identified the upstream and downstream genes that were annotated to further support the correct orthology. Rest of the genomes were annotated using *D. albomicans* amino acid sequences as query.

Sequence Alignment and Phylogenetic Inference

Codon-based phylogenetic analysis require the accurate alignment of ortholog sequences. We used the predicted coding sequences to build alignments using MUSCLE (Edgar 2004) and CLUSTAL Omega (Sievers et al. 2011). We translated the coding sequences into aminoacid sequences using Seaview v5.0.4 (Gouy et al. 2010) aminoacid translator and built the alignments. The aligned amino acid sequences were reverse translated and the resulting codon alignments were used in phylogenetic and selection inference.

Since many selection analyses we performed require the phylogenetic trees of the ortholog sequences, we built maximum-likelihood and Bayesian phylogenetic trees using IQtree (Trifinopoulos et al. 2016) (Last accessed: 16/07/2021) and MrBayes (Ronquist et al. 2012), respectively. Maximum-likelihood trees were built by selecting GTR nucleotide substitution model and node support was evaluated with 1000 bootstrapping replicates. Bayesian tree was constructed by running two independent runs, each with four chains (one cold and three heated). The analysis was run for 2 million generation saving every 1000th tree. The runs were terminated when the split frequency reached the value less than 0.001. 25% of the trees were discarded as burn-in while summarizing the trees. The summarized trees were edited and rendered using Figtree v1.4.4 (Rambaut 2010).

Divergence Analysis

To detect the nature and strength of selection acting on the spermatogenesis genes, we employed a combination of evolutionary analysis. These analyses estimate the ratio of non-synonymous to synonymous substitution rate ($dN/dS=\omega$) across the genes, codons and lineages. When there is no selection acting on a given gene/codon/lineage, both non-synonymous and synonymous substitutions are expected to become fixed with the same probability ($\omega=1$). In the presence of selection, selection advantage can increase the fixation probability of non-synonymous substitution ($\omega>1$, positive selection), or decrease it due to selection constrains ($\omega<1$, negative or purifying selection).

Several methods have been developed for detecting signature of positive selection based on the ratio of *dN/dS* at different levels such as whole alignment (gene-wide), branch specific, codon based and a combination of these. Gene-wide selection analysis were performed to detect the signature of selection using the alignments of all the species, without making any assumption about foreground branches. First, we employed *codeml* implemented in *Phylogenetic Analysis by Maximum Likelihood* v4 package (*PAML*) (Yang 2007). We compared a null model (*M7*) in which ω is assumed to be beta-distributed among sites and a selection model (*M8*), in which codons are allowed to have an extra category of positively selected sites with $\omega>1$. The significance of this test was validated using likelihood ratio test. A set of gene-wide selection tests were also employed using HyPhy (Kosakovsky Pond et al. 2005) webserver (Last accessed: 16/07/2021). We used *BUSTED* (Branch-site Unrestricted Statistical Test for Episodic Diversification) (Murrell et al.

2015), which specifically tests whether a gene has experienced positive selection in at least one site or one of the branches of a given phylogeny. *BSR* (Branch-site Random effects likelihood test) (Kosakovsky Pond et al. 2011) was used to test for episodic diversifying selection. Finally, *aBSREL* test (adaptive Branch-Site Random Effects Likelihood) which is a improved version of branch-site model, was used to test if positive selection has occurred in a proportion of branches.

Codon-based selection analysis was performed by employing maximum-likelihood methods implemented in *CODEML* of *PAML* v.4. *CODEML* estimates the ratio of non-synonymous to synonymous substitutions (ω) under various models allowing ω to vary among sites (site models) and branches (branch models) and a combination of both (branch-site models). A likelihood ratio test was performed in all the tests by comparing the null model against an alternative model. The test statistic $2\Delta l = 2(l_1 - l_2)$ where l_1 and l_2 are the likelihood values of null and alternative models, respectively, was calculated. The twice the difference between two likelihood values was compared with the chi-square distribution with degree of freedom to be the difference between number parameters. The Bayesian Empirical Bayes (BEB) approach was employed to identify positively selected sited by calculating the posterior probabilities of a particular site belongs to the class of sites under positive selection where sites with greater posterior probability ($\geq 95\%$) were considered to be under strong positive selection. Additionally, we performed mixed effects model of evolution (*MEME*) (Murrell et al. 2012) and Fast Unconstrained Bayesian Approximation for inferring selection (*FUBAR*) (Murrell et al. 2013) tests available in HyPhy webserver (Last accessed: 16/07/2021). *MEME* employs a mixed-effects maximum likelihood approach to test the hypothesis that individual sites have been subject to episodic positive selection or diversifying selection. *FUBAR* uses a Bayesian approach to infer synonymous (*dN*) and synonymous (*dS*) substitution rates on a per site basis for a given alignment and phylogeny.

We applied branch-site models for both frontal sheen complex and orbital sheen complex. The hybridization among species of frontal and orbital sheen complex often produces fertile hybrids, whereas hybridization between species from different complex often produce only sterile males and some combination of species produce both sterile males and females. Upon marking the branch of interest (foreground branch), the alternative hypothesis assigns some sites in the foreground branch to be under positive selection, whereas null hypothesis does not. The likelihood ratios of each model were compared and the significance and sites under positive selection were identified as stated above. Additionally, we also performed BSR and *aBSREL* tests described above to detect lineage-specific positive selection.

Polymorphism Analysis

We employed McDonals-Kreitman test (MK test) (McDonald and Kreitman 1991) using DnaSP v.6 (Librado and Rozas 2009) to detect the signature of recent selection by comparing the ω ratios within species with those between species. This test takes advantage of the intraspecific variation, where the ω ratios within species are expected to be equal to the ratios between species under neutral scenarios. We compared the combinations of species that produce fertile hybrids, sterile males, and both sterile males and females and compared their divergence rates. FDR correction was performed to account for multiple comparisons across genes for individual tests (Benjamini and Hochberg 1995).

Protein Domain Identification and Protein Modelling/Functional Assessment

We identified protein domains using Common Conserved domain Database (CDD) and Pfam 33.1 (El-Gebali et al. 2019), (Last accessed: 16/07/2021). The protein models *bam, bgcn*, *aly, comr* and *dj* were built using *D. melanogaster* protein structures as reference. We mapped all the sites with significant positive selection on to the three-dimensional protein structure using PyMOL (Schrodinger, LLC, 2015). We looked for the presence of human orthologs of all the genes analysed using DIOPT v8.0 integrated in Flybase (Marygold et al. 2013) (Last accessed: 16/07/2021). Human (Accession number: NP_001332905.1) and mice (Accession number: NP_001156485.1) ortholog sequences of *bgcn* were extracted from NCBI-Nucleotide database. Aminoacid sequences of *bgcn* from *D. melanogaster, D. albomicans, Mus musculus* and *Homo sapiens* were aligned using MUSCLE and sites under positive selection were mapped on to the conserved domains between the orthologs.

Results

Identification of Orthologous Spermatogenesis Genes in *nasuta*-Subgroup of *Drosophila*

We assembled a total of 38 genomes of species belonging to *nasuta*-subgroup of *Drosophila*. Amino acid sequences of 10 spermatogenesis genes of *D. melenogaster* were extracted from Flybase. These amino acid sequences were used as query in tBLASTn search against 6 *D. albomicans* genomes we assembled since it is the only species which has complete annotated genome available. We predicted the CDS and respective amino acid sequences of each gene from *D. albomicans* (See methods). We performed reciprocal blast search using NCBI-Blast program to reassure the right orthology. Further *D. albomicans* sequences were used

as query to annotate rest of the orthologs. We were able to extract a total of 331 orthologs from 35 genomes (Supplementary Table 1). A homolog of *Mst98Ca* was found upon the Blast search, we included the homolog in the analysis and named it *Mst98Ca-like* homolog. We excluded *D. nevifrons* and *D. immigrans* orthologs due to the high divergence of nucleotide sequences to avoid the problems associated with saturation of synonymous sites when comparing the diverged species. All the sequences are available in at figshare (https://figshare.com/s/ad586ca0f83d86871a45).

Phylogenetic Inference

The Bayesian phylogenetic trees (Figs. 1 and 2) constructed for individual genes were consistent with the species tree constructed by (Mai et al. 2020). *D. nasuta, D. albomicans* and *D. kepulauana* formed a single clade (*nasuta* subclade/frontal sheen complex). *D. pulaua, D. s. sulfurigaster, D. s. bilimbata, D. s. albostrigata, D. s. neonasuta* formed a seperate clade (*sulfurigaster* subclade/orbital sheen complex), whereas *Taxon-F*, the only species formed clade branching from the root of the tree. Node support values for all the major nodes was significant.

Rapid Divergence of Spermatogenesis Genes and Weaker Selective Constrain

We employed free ratio (*M0*) model of PAML to estimate the global ω (*dN/dS*) for all the spermatogenesis genes. The global ω estimates were similar for both the alignment methods used but varied significantly for each gene analysed (Supplementary Fig. 1). *bam* showed the highest ω (0.62) followed by *aly* (0.37), *dj* (0.34) and *comr* (0.30). 8 out of 11 spermatogenesis genes analysed showed higher ω than the reported median ω for spermatogenesis genes (0.10) in *D. melanogaster* subgroup (Haerty et al. 2007). *Mst98Ca* (0.009) had the least ω estimate followed by *sa* (0.066) and *JYalpha* (0.042). We employed the Fast, Unconstrained Bayesian Approximation (FUBAR) analysis, which detects sites evolving through purifying and diversifying selection. The strongest constrain was observed for *Mst98Ca* with 43.55% (Supplementary Fig. 2) of the codons evolving under negative purifying selection. The selective constrains was the weakest for early-stage spermatogenesis genes such as *Bam* (2.3%), *Bgcn* (4.19%) and *Rux* (2.47%).

Evidence of Gene-Wide and Codon-Based Positive Selection

All the genes except *Mst98Ca* showed signature of positive selection in at least one of the gene-wide selection analyses (Table 1). Likelihood ratio test (LRT) of *codeml* favoured

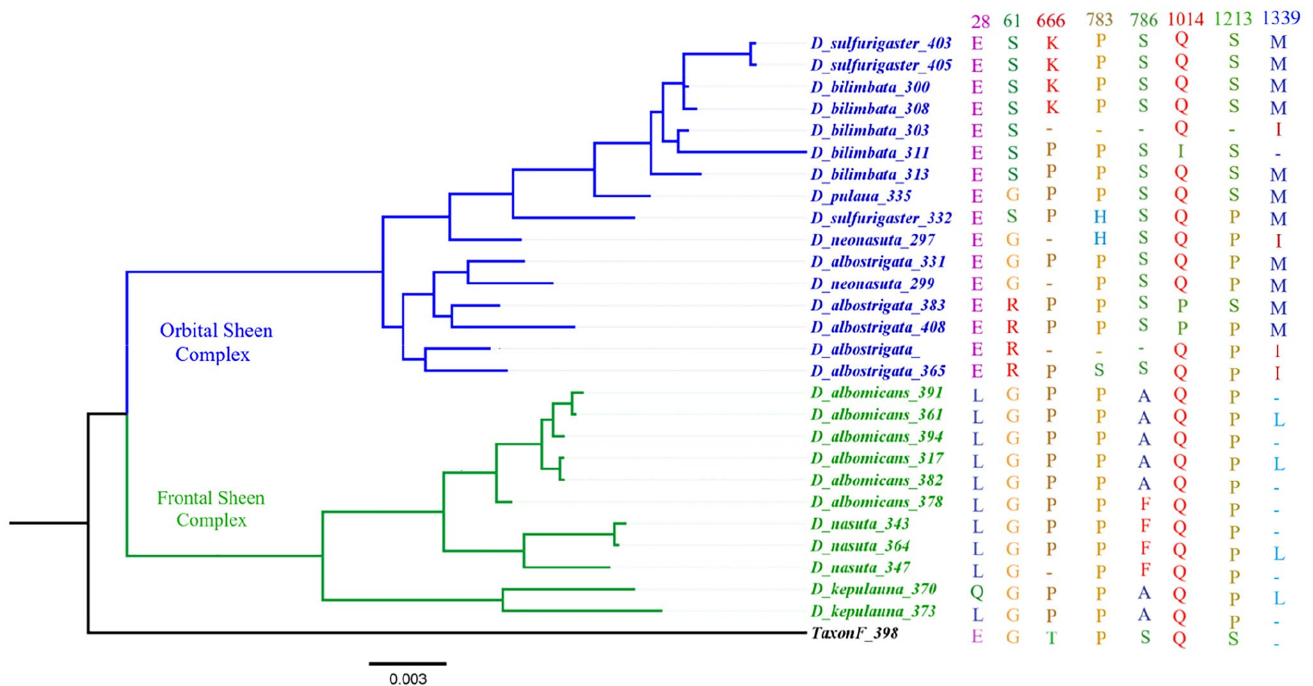

Fig. 1 Bayesian phylogeny of *bgcn* infered using nucleotide sequences of *nasuta*-subgroup species. Node support for each major clade is indicated. Position of amino acid sites under positive selection are shown next to the individual species

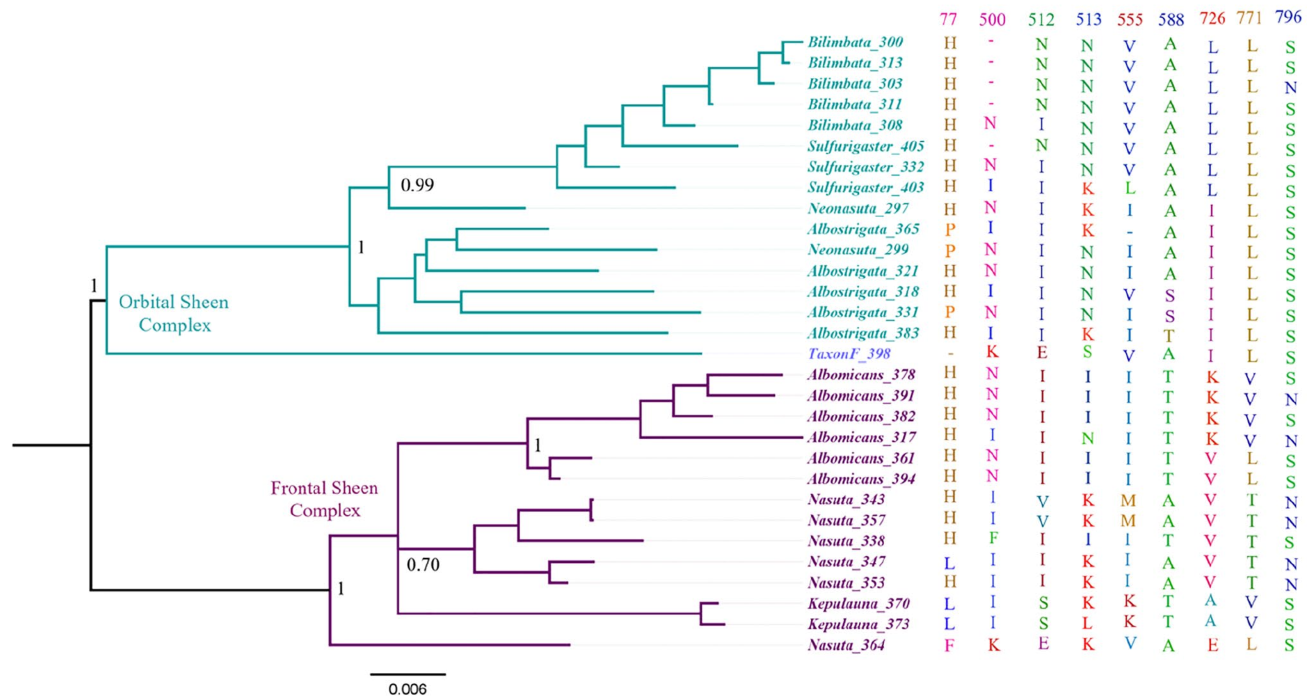

Fig. 2 Bayesian phylogeny of *comr* infered using nucleotide sequences of *nasuta*-subgroup species. Node support for each major clade is indicated. Position of amino acid sites under positive selection are shown next to the individual species

Table 1 Gene-wide tests for positive selection and McDonald-Kreitman test

Gene	*Codeml (M8)*	BUSTED	BSR	*aBSREL*	MK-test
bam	**0.0100**	0.500	> 0.05	< 0.05	> 0.05
bgcn	**0.0010**	**0.000**	**< 0.0001**	**0.0000**	> 0.05
rux	0.9950	**0.000**	**< 0.0001**	**0.0000**	> 0.05
achi	0.9950	**0.005**	**< 0.0001**	**0.0000**	**< 0.05**
aly	**0.0010**	0.105	**< 0.0001**	**0.0054**	**< 0.05**
comr	**0.0010**	**0.036**	**0.0000**	**0.0013**	**< 0.05**
sa	0.9000	**0.001**	**0.003**	**0.0091**	> 0.05
dj	**0.0010**	**0.000**	**< 0.0001**	**0.0000**	> 0.05
Jy-alpha	**0.0010**	**0.010**	**< 0.0001**	**0.0000**	> 0.05
mst98Ca	0.9000	0.064	> 0.05	< 0.05	> 0.05
mst-98Ca-Like	**0.0010**	0.500	**< 0.0001**	**0.0000**	**< 0.05**

A summary of FDR-corrected (5%) *P*-values (Benjamini and Hochberg 1995) obtained by individual test for each gene is shown

Statistically significant *P*-values are in bold

the M8 selection model for all the genes except *rux*, *achi*, *Mst98Ca* and *sa*. *BUSTED* detected positive selection at all the genes except for *aly*, *bam*, *Mst98Ca* and *Mst98Ca-like*. Both *BSR* and *aBSREL* showed positive selection for at least one of the lineages in the gene tree of all the genes except *bam* and *Mst98Ca*.

We investigated the nature of natural selection influencing the spermatogenesis genes in the codon level by employing maximum-likelihood models implemented in *PAML* (see methods). We employed two pairs of models *M1a* vs *M2a* and *M7* vs *M8* and only considered the sites with significant positive selection (posterior probability ≥ 90%) inferred by *M7* vs *M8* comparison. The Byes Empirical Bayes (BEB) implemented in *M8* identified 8 out of 11 genes with significant positively selected sites (Table 2). Early spermatogenesis genes such as *bam* and *bgcn* showed significant positive selection (PP ≥ 90%) at 3 and 8 sites, respectively. *rux* did not show any sites under positive selection. Spermatocyte arrest class genes such as *aly* and *comr* showed significant positive selection at 7 and 9 sites, respectively (Table 2). BEB identified one site with positive selection in *sa* but it was insignificant with PP less than 90%, whereas in *achi*, there were no sites under positive selection. *dj* and *JYalpha* showed 4 and 1 sites under positive selection (Table 2) among the genes involved in late spermatogenesis. *Mst98Ca* and *Mst98Ca-like* did not show any positive selection acting on any of the codons.

Additionally, we analysed the codon alignments for signature of positive selection using *MEME* and *FUBAR* (Supplementary Table 2). Among early spermatogenesis genes, MEME identified 7 and 2 sites under significant ($P \leq 0.05$) diversifying selection for *bgcn* and *rux,* respectively. *bam* did not show signature of diversifying selection at any sites.

Table 2 Likelihood ratio test statistic for site models (M7 vs. M8)

Gene	-2ΔlnL[c] ([a]M7 vs. [b]M8)	[d]Df	*P*-value (after FDR correction)[e]	[f]Sites identified by BEB (PP ≥ 90%)	Sites identified by MEME (*P* ≤ 0.05)	Sites identified by FUBAR(PP ≥ 90%)
achi	0.51	2	0.9950	–		
bam	8.41	2	**0.0100**	45 G, 293 T**, 331 H*	–	45, 64, 217, 293, 331, 334, 384
bgcn	18.21	2	**0.0010**	28 Q, 61 G*, 666 P**, 783 P, 786 A*, 1014 Q, 1213 P, 1339 L**	429, 432, 666, 786, 1012, 1014, 1016	61, 63, 66, 428, 570, 666, 783, 786, 810, 1014, 1202, 1213, 1308
rux	0.0001	2	0.9750	–	44, 161	116, 161, 256
aly	59.63	2	**0.0010**	482 D**, 483 N*, 484 L**, 486 E**, 487 I, 488 L**, 489 P**	277, 488	48, 183, 232, 287, 339, 419
comr	28.43	2	**0.0010**	77 L**, 500 I*, 512 S, 513 K, 555 K**, 588 T, 726 A*, 771 V*, 796 S	76, 468, 482, 513, 746, 771, 878	77, 411, 490, 500, 502, 513, 555, 576, 588, 726, 771
sa	2.81	2	0.5000	328 T	328	328
dj	12.17	2	**0.0010**	10 R*, 14 V**, 117 S*, 259 E	10, 14	14, 10, 76, 169, 259
Jy-alpha	29.95	2	**0.0010**	820 V*	820, 828	820, 831
Mst98Ca	6.60	2	**0.0010**	–	0	1
Mst98Ca-like	0.15	2	0.9000	–	0	0

Statistically significant *P*-values are in bold

[a]M7 is a null model that assumes that $0 < \omega < 1$ is beta distributed among sites; [b]M8 (positive selection model) is the same as M7, but also includes an extra category of sites with $\omega > 1$. 2ΔlnL[c]: is twice the difference of the natural logs of the maximum likelihood of the models being compared. [d]Df is the degree of freedom = 2. [e]Positions of the sites identified by BEB are relative to the *D. albomicans* sequence (*PP ≥ 90%, **PP ≥ 95%). [f]*P* values obtained by likelihood ratio test after FDR correction (5%)

Among spermatocyte arrest class genes *aly*, *comr* and *sa* each showed 2,7 and one site under significant diversifying selection, whereas *achi* did not show diversifying selection. Among late spermatogenesis genes, only *Dj* and *JYalpha* both showed 2 sites each under significant diversifying selection. *FUBAR* identified 7 sites for *bam*, 13 sites for *bgcn*, and 3 sites for *rux* under positive selection with posterior probability ≥ 90 which is considered significant. *aly*, *comr* and sa showed 6, 11 and one sites under significant positive selection. Finally, *dj*, *JYalpha* and *Mst98Ca* showed 5, 2 and one site each under significant positive selection.

Test for Lineage-Specific Positive Selection

To investigate whether the signature of positive selection observed in gene-wide and codon-based selection analysis is due to the effect of single lineage, we applied branch-site and branch-specific models to infer positive selection. The phylogeny of *nasuta*-subgroup species splits into frontal sheen complex (FSC) and orbital sheen complex (OSC), we performed branch-site tests considering one of the lineages as foreground and the other as background branch (described in methods). Upon performing likelihood ratio test, we found that the signature of lineage-specific positive selection was insignificant for all the genes analysed (Table 3). Although insignificant, BEB identified sites under positive selection for 6 genes we analysed. *bgcn* showed positive selection in the branch leading to FSC with 10 sites identified by BEB (Table 3). Interestingly, *rux* and *achi* which did not show any sites positive selection in the site models of *codeml* showed although insignificant, some sites under positive selection in the branch-site test. BEB picked one and two sites, respectively, for FSC and OBS for *rux* and 3 sites in FSC for *achi* (Table 3). *aly* showed one site in the branch leading to FSC and *comr* and *dj* showed one and two sites, respectively, in the branch leading to OSC (Table 3).

Additionally, we employed *aBSREL* test to detect selection acting on a proportion of sites in individual lineages. All the genes except *bam* and *Mst98Ca* showed signature of positive selection in at least one of the branches in the phylogeny (Supplementary Fig. 3).

Selection Inference Using Pattern of Polymorphism and Divergence

Four (*achi*, *aly*, *comr* and *Mst98Ca*-like) of the eleven genes analysed showed the significant departure from neutrality in at least one of the hybridizing pair compared in MK test (Table 1). The departure from the neutrality observed at these four genes was due to both the excess of

Table 3 Likelihood ratio test statistic for branch-site tetst

Gene	Foreground branch (MA and MA1)[a]	-2ΔlnL[b]	*P*-value[c]	Sites identified by BEB[d]
bam	FSC	0.1482	0.5000	–
	OSC	0.0506	0.9000	–
bgcn	FSC	0.2398	0.5000	29 Q, 747 S, 767 G, 847 N, 872 C, 956 L, 1049 C, 1072 S, 1093 S, 1177 L
	OSC	0	0.9950	–
rux	FSC	0.6161	0.5000	210 H
	OSC	0.0604	0.9000	180 T, 245 I
achi	FSC	0	0.9950	321 F, 371 V, 434 A
	OSC	0	0.9950	–
aly	FSC	0	0.9950	51 V
	OSC	0	0.9950	–
comr	FSC	0	0.9950	–
	OSC	0.0076	0.9000	467 E
sa	FSC	0	0.9950	–
	OSC	0	0.9950	–
dj	FSC	0	0.9950	–
	OSC	0	0.9950	269 E, 291 D

[a]MA and MA1are the branch-site models employed. MA allows a proportion of codons with $dN/dS \geq 1$ on the foreground branches (*FSC* frontal sheen complex, *OSC* orbital sheen complex), whereas the MA1 model does not. 2ΔlnL[b] is twice the difference of the natural logs of the maximum likelihood of the models being compared. [c]Degrees of freedom = 1. [d]Positions of the sites identified by BEB are relative to the *D. albomicans* sequence

non-synonymous differences between species and excess of synonymous polymorphisms. Except for *achi*, there other genes (*aly*, *comr* and *Mst98Ca*-like) showed departure only in the comparison between species pair that result in sterile hybrids. There was pattern of increased synonymous and non-synonymous polymorphism in the comparison between the species producing fertile hybrids, whereas among the species that produce sterile hybrids, there was excess of between species divergence (Supplementary Table 14).

Discussion

Drosophila has long been used as a model to understand the mechanisms of speciation such as pattern of genetic diversification and identifying the genes involved in hybrid incompatibilities (Orr 1993). Most of these studies have been conducted in *D. melanogaster* subgroup (Bayes and Malik 2009; Brideau et al. 2006; Phadnis and Orr 2009; Presgraves 2003) where molecular mechanism of hybrid incompatibilities is understood in crosses between many sibling species. *D. melanogaster* and its sibling species have accumulated many such hybrid incompatibilities (Masly and Presgraves 2007; Presgraves 2003). Investigating a much younger subgroup potentially helps in understanding molecular mechanisms and evolutionary forces acting at the early stage of speciation process.

The *nasuta*-subgroup which diverged only about 3.5 MYA, with its pronounced difference in pre and post-zygotic reproductive isolation provides an excellent model to understand the process of speciation. Many species in the subgroup can produce viable, fertile and sterile offspring upon crossing between other members of the species complex (Kitagawa et al. 1982; Spieth et al. 1969; Wilson et al. 1969). Rapid divergence and has been established as one of the evolutionary forces capable of bringing about such incompatible interactions between closely related species. Our analysis of key spermatogenesis genes provide evidence for possible role of rapid divergence in bringing about hybrid incompatibilities.

We annotated a total of 331 orthologs of key spermatogenesis genes which are involved in the key stages of early, mid and late spermatogenesis process. We employed robust selection analysis to infer the mode and strength of Darwinian selection acting on these genes. Our study shows a pattern of high sequences divergence for five of eleven genes analysed between closely related hybridizing species of *nasuta*-subgroup of *Drosophila*. Such pattern of rapid divergence is expected for sex and reproduction-related genes (Haerty et al. 2007), but it is inconsistent considering the selection constraints on germline stem

cell regulatory genes such as *bam* and *bgcn*. However, evidence from previous population genetic studies of spermatogenesis genes with major role in germline stem cell (GSC) regulation (Bauer DuMont et al. 2007; Choi and Aquadro 2014; Civetta et al. 2006) suggest that many genes with role in stem cell regulation evolve adaptively.

We analysed three genes *bam*, *bgcn* and *rux*, with key role in early stage of spermatogenesis. *bam* and *bgcn* are two genetically interacting genes which are regulators of gametogenesis in both the sexes (Lavoie et al. 1999). In females, the proper functioning of *bam* and *bgcn* is essential for the initiation of cytoblast differentiation. In addition, *bam* and *bgcn* are also involved in the assembly of endoplasmic reticulum-like fusome. In males, *bam* and *bgcn* are required for the switch from spermatogonial program of mitotic divisions to the spermatocyte differentiation (Fuller 1998; Schulz et al. 2004). *rux* is an essential cell cycle regulator in *Drosophila*, which has been shown to down-regulate CyclinA-dependent activity during G1 phase and is also responsible for temporary G1 arrest. Considering the role of *bam*, *bgcn* and *rux* in regulating the developmental witches during gamatogenesis, one might expect them to evolve under high selective constrains. However, evidence for rapid amino acid evolution of *bam* and *bgcn* has been documented in *D. melanogaster* and *D. simulans* clade (Bauer DuMont et al. 2007; Civetta et al. 2006) and in *melanogaster* subgroup for *rux* (Avedisov et al. 2001; Llopart and Comeron 2008).

Our codon-based analysis revealed that *bam* and *bgcn* are evolving under a strong positive selection, whereas *rux* only showed such signature in individual branches (FSC and OSC). One of the three positively selected site in *bam* is situated in the predicted nuclease domain (SI Fig.) and other two (Proline and Serine) on the PEST domain which is rich in proline (P), glutamic acid (E), serine (S) and threonine (T). PEST motif has been associated protein that are unstable and rapidly degraded by proteases (Rogers et al. 1986). The cytoplasmic form of Bam transiently expressed and it starts to accumulate at the cytoblast differentiation and disappears after completion of four rounds of mitosis (McKearin and Ohlstein 1995; Szakmary et al. 2005). Despite the significant positive selection detected in the *M7* vs *M8* comparison, branch-site models and polymorphism analysis failed to identify any sites under positive selection. This could be due to the fact that rapid divergence is common among genes that transiently expressed (Cutter and Ward 2005).

The predicted *bgcn* protein domain architecture of *D. albomicans* consists of 1325 amino acids (Fig. 3). *bgcn* is predicted to have helicase core module, an ankyrin repeat domain (ARD) inserted between the two helicase core domains and containing a pair of ankyrin repeats domains and two C-terminal extensions such as helicase-associated 2 (HA2) and oligonucleotide binding (OB) domains. 3 of 12 sites under significant positive selection are found in HA2 and OB domains, respectively. MEIOC and YTHDC2 are proposed to be the mammalian homolog of *bam* and *bgcn* known to play a role in the stem cell transition from mitotic to meiotic division. Ketu (keen to exit meiosis leaving testes under-populated) is a non-synonymous mutation in ythdc2 (Morohashi et al. 2011; Stoilov et al. 2002) and the for ketu mutation homozygotes are both male and female sterile in mice. Most insects' lineages have YTHDC2 orthologs with full architecture including YTH domain. However, the orthologs in *Drosophila* lack the YTH domain (Supplementary Fig. 5) suggesting the loss of YTH domain in Last Common Ancestor (LCA). Multiple sequence alignment of Human, mice and three *Drosophila* species (Supplementary

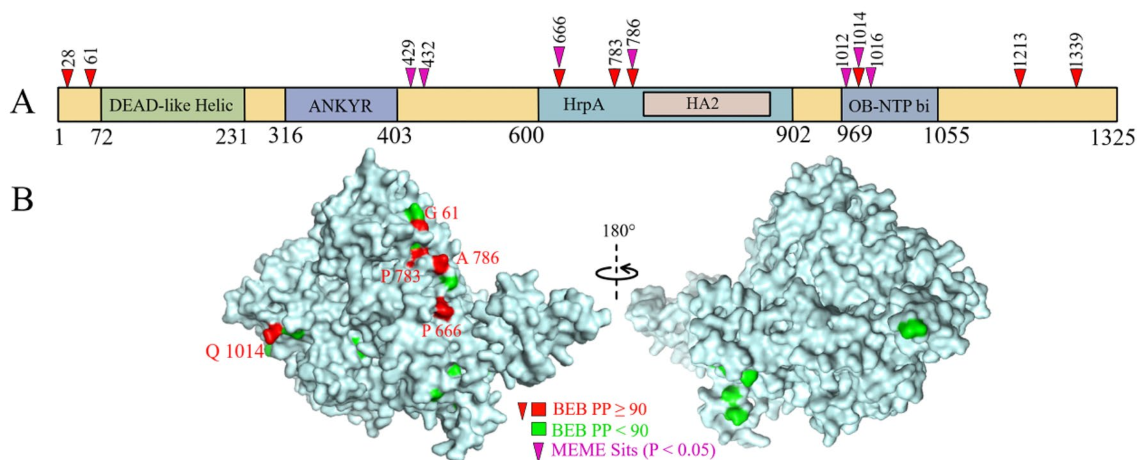

Fig. 3 Nucleotide divergence in early spermatogenesis gene *bgcn* among species of *nasuta*-subgroup of *Drosophila*. **A** Representation of *bgcn* protein showing predicted domains. Sites with significant signature of positive selection are shown in red and magentha. **B** Predicted three-dimensional model of *bgcn* protein (PDB of the template: 6up4.1.A). Amino acid sites under positive selection are highlighted (BEB posterior probability ≥ 90: Red, BEB posterior probability < 90: Green) (Color figure online)

Fig. 5) showed many sites under positive selection are distributed among highly conserved sites.

rux is a dose-dependent regulator of second meiotic division during spermatogenesis, in the absence of *rux* function, germ cells execute meiosis I and II, but then undergo and additional division as haploid cells. High expression of *rux* has been shown to result in failure to execute meiosis II (Lifschytz and Meyer 1977). Although, *M7* vs *M8* comparison of did not show any significance for positive selection MEME and FUBAR identified 2 and 3 sites, respectively. Branch-site model inferred weak positive selection on one and two sites in frontal and orbital sheen complex, respectively. Rux has a nuclear localization signal (NLS) domain which spans between 253 and 276 amino acids. The sites 256 falls in the NLS domain of the *rux* protein. NLS domain of *rux* shows high divergence in *melanogaster* subgroup. Mutants of *rux* show sterility in males but not in females, hence it has been proposed to be male-biased gene with role in spermatogenesis shows rapid divergence potentially driven by post-copulatory sexual selection/sexual conflict (Ellegren and Parsch 2007).

There is an existing hypothesis that germline genes coevolve with pathogens infecting the germline can result in elevated non-synonymous substation rate in *bam* and *bgcn* (Bauer DuMont et al. 2007). *Wolbachia* and *Spiroplasma* are two maternally inherited bacterial endosymbionts known to infect some *Drosophila* species (Mateos et al. 2006; Watts et al. 2009). Extensive divergence of *bam* due to *Wolbachia* infection between *D. melanogaster* and *D. simulans* affects *bam* function in females but has no apparent effect in males (Flores et al. 2015). *Wolbachia* infection can have both beneficial and deleterious effect on the fitness of *Drosophila* by increasing resistance viral infection and reducing the fecundity and life-span of the infected individuals, respectively (Chrostek et al. 2013). Maintaining the balance between both the beneficial and deleterious effects could potentially contribute to an 'arms race' between GSC regulatory genes and endosymbionts (Bauer DuMont et al. 2007). *D. ananassae* has been infected with *Wolbachia* for longer than *D. melanogaster* and despite which *bam* and *bgcn* did not show any signature of positive selection (Choi and Aquadro 2014). Considering that *D. nasuta* and *D. albomicans* are free from *Wolbachia* infection (Ravikumar et al. 2011), we can rule out the possibility of divergence of GSC genes to be driven by endosymbiont infection.

In *Drosophila* spermatogenesis, most transcription ceases during the entry into meiotic divisions. Therefore, the genes encoding proteins required for spermatid differentiation are transcribed in primary spermatocytes but translationally repressed until the appropriate time later in gamete development (Fuller 1998; White-Cooper and Bausek 2010). More than 2000 testis-specific transcripts are synthesized in primary spermatocyte (Doggett et al. 2011; White-Cooper 2010). Transcription in primary spermatocyte depends on a group of genes together named ''meiotic arrest'' genes (Ayyar et al. 2003; Jiang and White-Cooper 2003; Wang and Mann 2003; White-Cooper 2000, 1998). Broadly there are two meiotic arrest genes: *aly*-class (*aly*, *comr*, *tomb*, *topi* and *achi/vis*) and *can*-class (*can*, *mia*, *nht*, *rye* and *sa*). Among the aly-class genes, aly encodes the *Drosophila* homologue of *C. elegans synMuvB* gene *lin*-9 (Beitel et al. 2000; White-Cooper 2000). comr encodes a novel protein of unknown function (Jiang and White-Cooper 2003). achintya/vismay (achi/vis) and matotopetli (*topi*) encode sequence-specific DNA-binding proteins (Ayyar et al. 2003; Perezgasga et al. 2004; Wang and Mann 2003). The *can*-class genes encode the testis-specific TBP-associated factors (tTAFs), suggesting that their products form a testis-specific TFIID complex in primary spermatocytes (Hiller et al. 2001, 2004). We found evidence of rapid divergence at two of the four meiotic arrest genes analysed in the current study. *aly* and *comr* had seven and nine sites under positive selection with PP ≥ 90% in *M7* vs *M8* comparison of *codeml* (Table 2). The predicted *D. albomicans comr* protein has 891 amino acids and eight of the nine positively selected sites identified by BEB are mapped onto a single domain with unknown function (Supplementary Fig. 4). *comr* and *achi* also showed deviation from neutrality in polymorphism analysis. *D. melanogaster* comr predicted protein has an acidic domain in the C terminus of the protein (amino acids 518–570), and a predicted nuclear localisation sequence (NLS) (amino acids 583–589). In addition, a region that may represent a very divergent PB1 domain (amino acids 348–431). PB1 domains have been shown to mediate protein–protein interactions (Ito, 2001; Ponting et al. 2002).

After four rounds of mitotic divisions, the *Drosophila* germ cells enter meiotic prophase. After rapid meiotic divisions sperm morphogenesis takes places. During morphogenesis the chromatin undergoes condensation and the nuclei acquire needle-like shape. During this stage, the two mitochondrial derivatives elongate along the entire length of the axoneme to form the flagellum simultaneously in all 64 spermatids of one cyst. Throughout this process, the germ cells remain interconnected via cytoplasmic bridges. Finally, spermatids become individualized and stored as motile sperm. In *Drosophila* spermatogenesis, transcriptional activity ceases after the meiotic divisions while translation proceeds. Hence, many mRNAs are translationally repressed during meiotic prophase and translationally activated during sperm morphogenesis making translational control is a crucial feature of spermatogenesis (Schafer et al. 1990). The genes such as *don juan* (Santel et al. 1997) and *Drosophila* gene family *Mst(3)CGP* (Gigliotti et al. 1997; Kuhn et al. 1988; Schafer et al. 1990) are known to express during spermatogenesis and encode translationally repressed mRNA. *dj* encodes a protein of 29 kDa with structural similarities

to histone H1 and it is localized in haploid nuclei during chromatin condensation and nuclear shaping. It can also be detected in the mitochondrial derivatives of the flagellum (Santel et al. 1998). Of four spermatid differentiation genes analysed in the current study, *dj* showed signature of rapid divergence in all the tests employed (Table 1). One and two positively selected sites are present on the two predicted domains of *Dj* (Fig. 4). *JYAlpha* encodes the alpha subunit of Na^{+} and K^{+} adenosine triphosphatase (Na + /K + ATPase), a transmembrane protein involved in ion exchange (Blanco and Mercer, 1998). One of the four mammalian isoforms of the Na + /K + ATPase alpha subunit, a4, is expressed exclusively in testes and is essential for sperm motility (Woo et al. 2000). *JYAlpha* is located on the fourth chromosome of *D. melanogaster* but on the third chromosome of *D. simulans*. Because of this transposition event of *JYAlpha,* a fraction of hybrids completely lacks *JYAlpha* and are sterile. The coding region of *JYAlpha* shows no signs of divergence by positive natural selection between *D. melanogaster* and *D. simulans* making a special case of reproductive isolation without sequence evolution. Contrast to this, our analysis showed signature of rapid divergence and positive selection (3 sites) in both gene-wide and codon-based analysis performed (Tables 1 and 2).

Hybrid incompatibilities such as inviability or sterility result from failed interactions between the genomes of parental species in F1 hybrids. Sterility of heterogametic sex is one of the most frequent result of crosses between closely related species (Haldane 1922). In Drosophila genus, the males being the heterogametic sex, the males show the sterility phenotypes. Several recent studies have suggested that disruptions in gene expression may be one source for sterility phenotypes (Hoekstra and Coyne 2007; Ortiz-Barrientos et al. 2006; Ranz and Machado 2006). The fact that sperm development is disrupted in *Drosophila* interspecies sterile hybrids, combined with the knowledge of spermatogenesis gene function in *Drosophila melanogaster*, has recently led to a series of studies comparing patterns of spermatogenesis gene expression in fertile parental species and sterile hybrids. The studies suggest that, more post-meiotic (spermiogenesis) than meiotic and pre-meiotic genes have been found to be significantly under expressed in sterile hybrids compared to parental species (Catron and Noor 2008; Michalak and Noor 2003, 2004; Moehring et al. 2007). Genome-wide miss-expression comparisons of *D. simulans*, *D. mauritiana* and their sterile male progeny found *don juan, Mst84Dc* and *Mst98Ca*, the three spermatid differentiation genes to be consistently down regulated in sterile hybrids (Michalak and Noor 2003; Moehring et al. 2007). Consistent with our analysis proving rapid divergence of *don juan*, and homolog of *Mst98Ca*, abnormalities such as fused sperm tails have been observed in crosses between some strains of *D. nasuta* and *D. albomicans* (Zhang et al. 2015). The same study showed *Mst98Ca* mapping on to on of the one of the hybrid male sterility QTL.

A typical speciation genetics study starts with studying the divergent reproductive traits between two species. Numerous such studies have identified genes that are rapidly diverging between closely related species, but these genes cannot be qualified as ‘speciation genes’ considering the possibility of genetic divergence after the speciation event. Nevertheless, two common pattern that have emerged from so far speciation genetic. The first is the ‘faster male’ evolution where HMS evolves at a rate an order of magnitude higher than HFS and HI (Tao et al. 2003; Tao and Hartl 2003). Second is the “large X” evolution in which HMS genes are enriched on the X chromosomes (Masly and Presgraves, 2007; Tao and Hartl 2003; White et al. 2012).

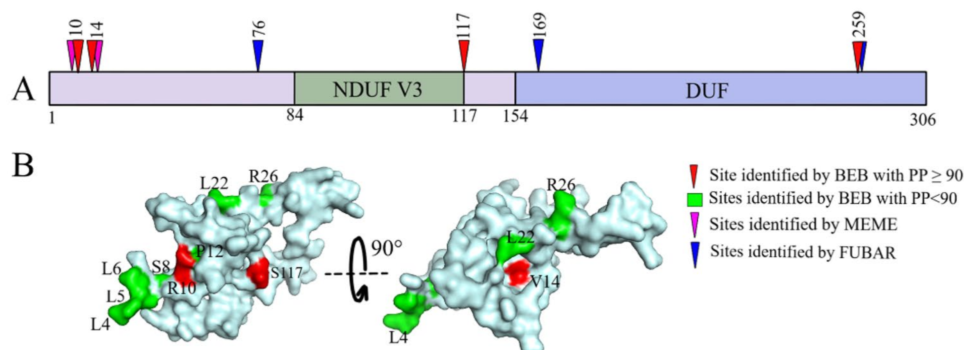

Fig. 4 Nucleotide divergence in late spermatogenesis gene *dj* among species of *nasuta*-subgroup of *Drosophila*. **A** Representation of domain architecture of *dj* protein. *dj* has a NDUF V3 domain and a large domain with multiple subdomains of unknown function. Sites with significant signature of positive selection inferred from BEB are shown in red, sites identified by MEME and FUBAR are shown in magenta and blue, respectively. **B** Three-dimensional model of *dj* protein subunit covering the NDUF V3 domain (PDB of the template: 6a70.1.B). Amino acid sites under positive selection are highlighted (BEB posterior probability ≥ 90: Red, BEB posterior probability < 90: Green) (Color figure online)

The above two patterns are better explained by the "conflict theory" where genomic divergence is driven by selfish genes, prominently by sex ratio distortion (SRD), also called sex chromosome meiotic drive (Frank 1991; Hurst and Pomiankowski 1991; Meiklejohn and Tao 2010). Meiotic drive is generally harmful to a genome since it breaches Mendelian ratio by gaining more than 50% transmission while quenching its homolog's share in the gene pool of next generation. Thus, suppressors to silence the distorter are under strong selection to evolve and make the meiotic drive cryptic (Hartl 1975). When an SRD arises on the X chromosome, counter evolution on the Y and the autosomes is anticipated, hence, SRD operates as a perpetual dynamo for genome evolution and bouts of this distortion-suppression process eventually lead to speciation (Meiklejohn and Tao 2010). *D. albomicans* has been shown to have a SRD in a hybridization between *D. albomicans* (Okinawa) females and *D. nasuta* (India) males. The F_1 males from this cross produce female-biased offspring. The driver was found to be located on the neo-X chromosome of *D. albomicans*, along with a drive suppressor, while *D. nasuta* was found to be suppressor-free (Yang et al. 2004). The same study also reported sterility in hybrid F_1 and F_2 males probably due to an interaction between the 3rd and Y chromosomes of *D. nasuta* and the autosomes of *D. albomicans.*

Combining the observed patterns such as 'faster-sex', 'faster-male', 'large-X effect' and 'conflict theory', our study proposes that the rapid evolution of spermatogenesis genes involved at the key stages of the process is by-product of the combination of these forces acting together in the whole of *nasuta*-subgroup. Despite the evolutionary constrains and no history of endosymbiont infection, GSC genes such as *bam* and *bgcn* showed higher divergence mediated by Darwinian positive selection. The hybrid male sterile phenotypes observed in the crossed between the species of *nasuta*-subgroup are consistent with the observed rapid divergence of late spermatogenesis genes such as *dj* and *Mst98Ca*. An extended investigation involving studying the specific stages of spermatogenesis arrest in the interspecies crosses would help in enhancing our understanding of intrinsic post-zygotic reproductive isolation in this subgroup. Comprehensive molecular population genetic analysis of more spermatogenesis loci would help in confirm the lineages or species-specific effect of positive selection and its role in hybrid male sterility phenotypes. Our study is the first attempt of understanding the genetic basis of post-zygotic reproductive isolation in *nasuta*-subgroup of *Drosophila* and lays a foundation for future exploration in the subgroup. Further detailed investigations using the genetic manipulation studies will enrich our understanding of the potential role of rapid divergence in bringing about hybrid male sterility.

Conclusions

In this study we have examined the molecular evolution of candidate genes with key role in various stages of spermatogenesis in species of *nasuta*-subgroup of *Drosophila*. We found evidence of rapid divergence at two early spermatogenesis genes, *bam* and *bgcn*. Another cell cycle regulator *rux* only showed lineage-specific positive selection in frontal sheen complex of the subgroup. We also observed signature of rapid divergence at *dj* and *Mst98Ca*, the key genes involved spermatid individualization. Our observations are consistent with the presence of *Mst98Ca* at one of the HMS QTL and of sperm-tail abnormality phenotype observed in the hybrids of *D. nasuta* and *D. albomicans*.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00239-021-10037-x.

Acknowledgements We thank the Department of Studies in Genetics and Genomics, University of Mysore for providing the infrastructure.

Author Contributions RC: Conceptualization, Methodology, Investigation, Writing—original draft, Writing—review & editing. RNB: Investigation, Resources, Writing—review & editing.

Funding This work was supported by Department of Biotechnology (DBT), Ministry of Science and Technology, Government of India to NBR (grant number BT/PR9871/BID/7/472/2013); We acknowledge Council of Scientific and Industrial Research (CSIR), Govt. of India for providing fellowship to RC.

Declarations

Conflict of interest The Author declares that they have no conflict of interest.

References

Avedisov SN, Rogozin IB, Koonin EV, Thomas BJ (2001) Rapid evolution of a cyclin A inhibitor gene, roughex, in Drosophila. Mol Biol Evol 18:2110–2118. https://doi.org/10.1093/oxfordjournals.molbev.a003752

Ayyar S, Jiang J, Collu A, White-Cooper H, White RAH (2003) Drosophila TGIF is essential for developmentally regulated transcription in spermatogenesis. Development 130:2841–2852. https://doi.org/10.1242/dev.00513

Bateson W (1909) Heredity and variation in modern lights. In: Darwin and Modern Science. Cambridge University Press, Cambridge, pp. 85–101

Bauer DuMont VL, Flores HA, Wright MH, Aquadro CF (2007a) Recurrent positive selection at Bgcn, a key determinant of germ line differentiation, does not appear to be driven by simple coevolution with its partner protein bam. Mol Biol Evol 24:182–191. https://doi.org/10.1093/molbev/msl141

Bayes JJ, Malik HS (2009) Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. Science 326:1538–1541. https://doi.org/10.1126/science.1181756

Beitel GJ, Lambie EJ, Horvitz HR (2000) The C. elegans gene lin-9, which acts in an Rb-related pathway, is required for gonadal sheath cell development and encodes a novel protein. Gene 254:253–263. https://doi.org/10.1016/S0378-1119(00)00296-1

Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B Methodol 57:289–300

Blanco G, Mercer RW (1998) Isozymes of the Na-K-ATPase: heterogeneity in structure, diversity in function. Am J Physiol Ren Physiol. https://doi.org/10.1152/ajprenal.1998.275.5.f633

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA (2006) Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. Science 314:1292–1295. https://doi.org/10.1126/science.1133953

Catron DJ, Noor MAF (2008) Gene expression disruptions of organism versus organ in Drosophila Species hybrids. PLoS ONE 3:e3009. https://doi.org/10.1371/journal.pone.0003009

Choi JY, Aquadro CF (2014) The coevolutionary period of Wolbachia pipientis infecting Drosophila ananassae and its impact on the evolution of the host germline stem cell regulating genes. Mol Biol Evol 31:2457–2471. https://doi.org/10.1093/molbev/msu204

Chrostek E, Marialva MSP, Esteves SS, Weinert LA, Martinez J, Jiggins FM, Teixeira L (2013) Wolbachia variants induce differential protection to viruses in Drosophila melanogaster: a phenotypic and phylogenomic analysis. PLoS Genet 9:1003896. https://doi.org/10.1371/journal.pgen.1003896

Civetta A, Rajakumar SA, Brouwers B, Bacik JP (2006) Rapid evolution and gene-specific patterns of selection for three genes of spermatogenesis in Drosophila. Mol Biol Evol 23:655–662. https://doi.org/10.1093/molbev/msj074

Coyne JA, Orr HA (1989) Patterns of speciation in Drosophila. Evolution 43:362–381. https://doi.org/10.1111/j.1558-5646.1989.tb04233.x

Coyne JA, Orr HA (1997) “Patterns of speciation in Drosophila” revisited. Evolution 51:295

Coyne JA, Orr HA (2004) Speciation. Sinauer Associates, Inc.

Cutter AD, Ward S (2005) Sexual and temporal dynamics of molecular evolution in C. elegans development. Mol Biol Evol 22:178–188. https://doi.org/10.1093/molbev/msh267

Dobzhansky T (1937) Genetics and the origin of species. Columbia University Press, New York

Doggett K, Jiang J, Aleti G, White-Cooper H (2011) Wake-up-call, a lin-52 paralogue, and always early, a lin-9 homologue physically interact, but have opposing functions in regulating testis-specific gene expression. Dev Biol 355:381–393. https://doi.org/10.1016/j.ydbio.2011.04.030

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–1797. https://doi.org/10.1093/nar/gkh340

El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD (2019) The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995

Ellegren H, Parsch J (2007) The evolution of sex-biased genes and sex-biased gene expression. Nat Rev Genet. https://doi.org/10.1038/nrg2167

Flores HA, Bubnell JE, Aquadro CF, Barbash DA (2015) The Drosophila bag of marbles gene interacts genetically with Wolbachia and shows female-specific effects of divergence. PLoS Genet 11(8):e1005453. https://doi.org/10.1371/journal.pgen.1005453

Frank SA (1991) Divergence of meiotic drive-suppression systems as an explanation for sex-biased hybrid sterility and inviability. Evolution 45:262–267. https://doi.org/10.1111/j.1558-5646.1991.tb04401.x

Fuller MT (1998) Genetic control of cell proliferation and differentiation in Drosophila spermatogenesis. Semin Cell Dev Biol 9:433–444. https://doi.org/10.1006/scdb.1998.0227

Gerts EM, Yu YK, Agarwala R, Schäffer AA, Altschul SF (2006) Composition-based statistics and translated nucleotide searches: improving the TBLASTN module of BLAST. BMC Biol 4:41. https://doi.org/10.1186/1741-7007-4-41

Gigliotti S, Balz V, Malva C, Schäfer MA (1997) Organisation of regulatory elements in two closely spaced Drosophila genes with common expression characteristics. Mech Dev 68:101–113. https://doi.org/10.1016/S0925-4773(97)00137-8

Gouy M, Guindon S, Gascuel O (2010) Sea view version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol 27:221–224. https://doi.org/10.1093/molbev/msp259

Haerty W, Jagadeeshan S, Kulathinal RJ, Wong A, Ram KR, Sirot LK, Levesque L, Artieri CG, Wolfner MF, Civetta A, Singh RS (2007) Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics 177:1321–1335. https://doi.org/10.1534/genetics.107.078865

Haldane JBS (1922) Sex ratio and unisexual sterility in hybrid animals. J Genet 12:101–109. https://doi.org/10.1007/BF02983075

Hartl DL (1975) Modifier theory and meiotic drive. Theor Popul Biol 7:168–174. https://doi.org/10.1016/0040-5809(75)90012-X

Hatsumi M, Morishige Y, Wakahama KI (1988) Metaphase chromosomes of four species of the Drosophila nasuta subgroup. Jpn J Genet 63:435–444. https://doi.org/10.1266/jjg.63.435

Hiller MA, Lin TY, Wood C, Fuller MT (2001) Developmental regulation of transcription by a tissue-specific TAF homolog. Genes Dev 15:1021–1030. https://doi.org/10.1101/gad.869101

Hiller M, Chen X, Pringle MJ, Suchorolski M, Sancak Y, Viswanathan S, Bolival B, Lin TY, Marino S, Fuller MT (2004) Testis-specific TAF homologs collaborate to control a tissue-specific transcription program. Development 131:5297–5308. https://doi.org/10.1242/dev.01314

Hoekstra HE, Coyne JA (2007) The locus of evolution: evo devo and the genetics of adaptation. Evolution. https://doi.org/10.1111/j.1558-5646.2007.00105.x

Hurst LD, Pomiankowski A (1991) Causes of sex ratio bias may account for unisexual sterility in hybrids: a new explanation of Haldane’s rule and related phenomena. Genetics 128:841–858

Ito T (2001) Novel modular domain PB1 recognizes PC motif to mediate functional protein-protein interactions. EMBO J 20:3938–3946. https://doi.org/10.1093/emboj/20.15.3938

Jiang J, White-Cooper H (2003) Transcriptional activation in Drosophila spermatogenesis involves the mutually dependent function of aly and a novel meiotic arrest gene cookie monster. Development. https://doi.org/10.1242/dev.00246

Kitagawa O, Wakahama K-I, Fuyama Y, Shimada Y, Takanashi E, Hatsumi M, Uwabo M, Mita Y (1982) Genetic studies of the Drosophila nasuta subgroup, with notes on distribution and morphology. Jpn J Genet 57:113

Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679. https://doi.org/10.1093/bioinformatics/bti079

Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K (2011) A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28:3033–3043. https://doi.org/10.1093/molbev/msr125

Kuhn R, Schäfer U, Schäfer M (1988) Cis-acting regions sufficient for spermatocyte-specific transcriptional and spermatid-specific translational control of the Drosophila melanogaster gene mst(3)gl-9. EMBO J 7:447–454. https://doi.org/10.1002/j.1460-2075.1988.tb02832.x

Kulathinal R, Singh RS (1998) Cytological characterization of premeiotic versus postmeiotic defects producing hybrid male sterility among sibling species of the Drosophila melanogaster complex. Evolution 52:1067–1079. https://doi.org/10.1111/j.1558-5646.1998.tb01834.x

Lachaise D, David JR, Lemeunier F, Tsacas L, Ashburner M (1986) The reproductive relationships of drosophila sechellia with d. mauritiana, *D. simulans*, and *D. melanogaster* from the afrotropical region. Evolution 40:262–271. https://doi.org/10.1111/j.1558-5646.1986.tb00468.x

Lavoie CA, Ohlstein B, McKearin DM (1999) Localization and function of bam protein require the benign gonial cell neoplasm gene product. Dev Biol 212:405–413. https://doi.org/10.1006/dbio.1999.9346

Librado P, Rozas J (2009) DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25:1451–1452

Lifschytz E, Meyer GF (1977) Characterisation of male meiotic-sterile mutations in Drosophila melanogaster—the genetic control of meiotic divisions and gametogenesis. Chromosoma 64:371–392. https://doi.org/10.1007/BF00294944

Llopart A, Comeron JM (2008) Recurrent events of positive selection in independent Drosophila lineages at the spermatogenesis gene roughex. Genetics 179:1009–1020. https://doi.org/10.1534/genetics.107.086231

Mai D, Nalley MJ, Bachtrog D (2020) Patterns of genomic differentiation in the *Drosophila nasuta* Species complex. Mol Biol Evol 37:208–220. https://doi.org/10.1093/molbev/msz215

Marygold SJ, Leyland PC, Seal RL, Goodman JL, Strelets VB, Wilson RJ, Ashburner M, Drysdale R, de Grey A, Gelbart W, Broll K, Crosby L, dos Santos G, Emmert D, Falls K, Gramates LS, Matthews B, Russo S, Schroeder A, StPierre S, Zhou P, Zytkovicz M, Brown NH, Adryan B, Attrill H, Costa M, Field H, Marigold S, McQuilton P, Millburn G, Ponting L, Osumi-Sutherland D, Stefancsik R, Tweedie S, Kaufman T, Matthews K, Goodmen J, Grumbling G, Strelts V, Thurmond J, Wong JD, Werner-Washburne M, Cripps R, Platero H (2013) FlyBase: Improvements to the bibliography. Nucleic Acids Res 41:D751–D757. https://doi.org/10.1093/nar/gks1024

Masly JP, Presgraves DC (2007) High-resolution genome-wide dissection of the two rules of speciation in Drosophila. PLoS Biol 5:e243. https://doi.org/10.1371/journal.pbio.0050243

Mateos M, Castrezana SJ, Nankivell BJ, Estes AM, Markow TA, Moran NA (2006) Heritable endosymbionts of Drosophila. Genetics 174:363–376. https://doi.org/10.1534/genetics.106.058818

McDonald JH, Kreitman M (1991) Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652–654

McKearin D, Ohlstein B (1995) A role for the Drosophila bag-of-marbles protein in the differentiation of cystoblasts from germline stem cells. Development 121:2937

Meiklejohn CD, Tao Y (2010) Genetic conflict and sex chromosome evolution. Trends Ecol Evol 25:215–223. https://doi.org/10.1016/j.tree.2009.10.005

Michalak P, Noor MAF (2003) Genome-wide patterns of expression in Drosophila pure species and hybrid males. Mol Biol Evol 20:1070–1076. https://doi.org/10.1093/molbev/msg119

Michalak P, Noor MAF (2004) Association of misexpression with sterility in hybrids of *Drosophila simulans* and *D. mauritiana*. J Mol Evol 59:277–282. https://doi.org/10.1007/s00239-004-2622-y

Moehring AJ, Teeter KC, Noor MAF (2007) Genome-wide patterns of expression in Drosophila pure species and hybrid males. II. Examination of multiple-species hybridizations, platforms, and life cycle stages. Mol Biol Evol 24:137–145. https://doi.org/10.1093/molbev/msl142

Mohanty S, Khanna R (2017) Genome-wide comparative analysis of four Indian Drosophila species. Mol Genet Genomics 292:1197–1208. https://doi.org/10.1007/s00438-017-1339-8

Morohashi K, Sahara H, Watashi K, Iwabata K, Sunoki T, Kuramochi K, Takakusagi K, Miyashita H, Sato N, Tanabe A, Shimotohno K, Kobayashi S, Sakaguchi K, Sugawara F (2011) Cyclosporin A associated helicase-like protein facilitates the association of hepatitis C virus RNA polymerase with its cellular cyclophilin B. PLoS ONE 6:e18285. https://doi.org/10.1371/journal.pone.0018285

Muller H (1942) Isolating mechanisms, evolution and temperature. Biol Symp 71–125

Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL (2012) Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:1002764. https://doi.org/10.1371/journal.pgen.1002764

Murrell B, Moola S, Mabona A, Weighill T, Sheward D, Kosakovsky Pond SL, Scheffler K (2013) FUBAR: a fast, unconstrained bayesian AppRoximation for inferring selection. Mol Biol Evol 30:1196–1205. https://doi.org/10.1093/molbev/mst030

Murrell B, Weaver S, Smith MD, Wertheim JO, Murrell S, Aylward A, Eren K, Pollner T, Martin DP, Smith DM, Scheffler K, Kosakovsky Pond SL (2015) Gene-wide identification of episodic selection. Mol Biol Evol 32:1365–1371. https://doi.org/10.1093/molbev/msv035

Nirmala SS, Krishnamurthy NB (1973) Cytogenetic studies on *Drosophila neonasuta*-A member of the *nasuta* subgroup. J Mysore Univ 26:162–167

Orr HA (1993) Haldane's rule has multiple genetic causes. Nature 361:532–533. https://doi.org/10.1038/361532a0

Ortiz-Barrientos D, Chang AS, Noor MAF (2006) A recombinational portrait of the *Drosophila pseudoobscura* genome. Genet Res 87:23–31. https://doi.org/10.1017/S0016672306007932

Perezgasga L, Jiang JQ, Bolival B, Hiller M, Benson E, Fuller MT, White-Cooper H (2004) Regulation of transcription of meiotic cell cycle and terminal differentiation genes by the testis-specific Zn-finger protein matotopetli. Development 131:1691–1702. https://doi.org/10.1242/dev.01032

Phadnis N, Orr HA (2009) A single gene causes both male sterility and segregation distortion in Drosophila hybrids. Science 323:376–379. https://doi.org/10.1126/science.1163934

Ponting CP, Ito T, Moscat J, Diaz-Meco MT, Inagaki F, Sumimoto H (2002) OPR, PC and AID: all in the PB1 family. Trends Biochem Sci. https://doi.org/10.1016/S0968-0004(01)02006-0

Presgraves DC (2003) A fine-scale genetic analysis of hybrid incompatibilities in Drosophila. Genetics 163:955–972

Rambaut A (2010) FigTree v1.3.1. institute of evolutionary biology. University of Edinburgh, Edinburgh

Ranz JM, Machado CA (2006) Uncovering evolutionary patterns of gene expression using microarrays. Trends Ecol Evol. https://doi.org/10.1016/j.tree.2005.09.002

Ravikumar H, Prakash BM, Sampathkumar S, Puttaraju HP (2011) Molecular subgrouping of Wolbachia and bacteriophage WO infection among some Indian Drosophila species. J Genet 90:507

Rogers S, Wells R, Rechsteiner M (1986) Amino acid sequences common to rapidly degraded proteins: the PEST hypothesis. Science 234:364–368. https://doi.org/10.1126/science.2876518

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP (2012) Mrbayes 3.2: efficient bayesian phylogenetic inference and model choice across a large model space. Syst Biol 61:539–542. https://doi.org/10.1093/sysbio/sys029

Santel A, Winhauer T, Blümer N, Renkawitz-Pohl R (1997) 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. https://doi.org/10.1016/S0925-4773(97)00031-2

Santel A, Blümer N, Kämpfer M, Renkawitz-Pohl R (1998) Flagellar mitochondrial association of the male-specific Don Juan protein in *Drosophila spermatozoa*. J Cell Sci 111:3299–3309

Schafer M, Kuhn R, Bosse F, Schafer U (1990) A conserved element in the leader mediates post-meiotic translation as well as cytoplasmic polyadenylation of a Drosophila spermatocyte mRNA. EMBO J 9:4519–4525. https://doi.org/10.1002/j.1460-2075.1990.tb07903.x

Schulz C, Kiger AA, Tazuke SI, Yamashita YM, Pantalena-Filho LC, Jones DL, Wood CG, Fuller MT (2004) A misexpression screen reveals effects of bag-of-marbles and TGFβ class signaling on the Drosophila male germ-line stem cell lineage. Genetics 167:707–723. https://doi.org/10.1534/genetics.103.023184

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J, Thompson JD, Higgins DG (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. https://doi.org/10.1038/msb.2011.75

Spieth TH (1969) Courtship and mating behavior of the *Drosophila nasuta* subgroup of species. Univ Tex Publ 6918:255–270

Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. https://doi.org/10.1093/nar/gki458

Stoilov P, Rafalska I, Stamm S (2002) YTH: a new domain in nuclear proteins. Trends Biochem Sci. https://doi.org/10.1016/S0968-0004(02)02189-8

Suzuki YM, Kitagawa O, Wakahama KI (1990) Chromosomal analysis and phylogenetic relationships in the *Drosophila nasuta* subgroup I. Phylogenetic relationships within the *Drosophila sulfurigaster* species complex. Genetica 80:53–66. https://doi.org/10.1007/BF00120120

Szakmary A, Cox DN, Wang Z, Lin H (2005) Regulatory relationship among piwi, pumilio, and bag-of-marbles in Drosophila germline stem cell self-renewal and differentiation. Curr Biol 15:171–178. https://doi.org/10.1016/j.cub.2005.01.005

Tao Y, Hartl DL (2003) Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. Iii. Heterogeneous accumulation of hybrid incompatibilities, degree of dominance, and implications for haldane's rule. Evolution 57:2580. https://doi.org/10.1554/03-094

Tao Y, Chen S, Hartl DL, Laurie CC (2003) Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. Genetics 164:1383–1397

Trifinopoulos J, Nguyen LT, von Haeseler A, Minh BQ (2016) W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res 44:W232–W235. https://doi.org/10.1093/nar/gkw256

Wakimoto BT, Lindsley DL, Herrera C (2004) Toward a comprehensive genetic analysis of male fertility in *Drosophila melanogaster*. Genetics 167:207–216. https://doi.org/10.1534/genetics.167.1.207

Wang Z, Mann RS (2003) Requirement for two nearly identical TGIF-related homeobox genes in Drosophila spermatogenesis. Development 130:2853–2865. https://doi.org/10.1242/dev.00510

Watts T, Haselkorn TS, Moran NA, Markow TA (2009) Variable incidence of spiroplasma infections in natural populations of Drosophila species. PLoS ONE 4:e5703. https://doi.org/10.1371/journal.pone.0005703

White MA, Stubbings M, Dumont BL, Payseur BA (2012) Genetics and evolution of hybrid male sterility in house mice. Genetics 191:917–934. https://doi.org/10.1534/genetics.112.140251

White-Cooper H, Schafer MA, Alphey LS, Fuller MT (1998) Transcriptional and post-transcriptional control mechanisms coordinate the onset of spermatid differentiation with meiosis I in Drosophila. Development 125:125–134

White-Cooper H, Leroy D, MacQueen A, Fuller MT (2000) Transcription of meiotic cell cycle and terminal differentiation genes depends on a conserved chromatin associated protein, whose nuclear localisation is regulated. Development 127(24):5463–5473

White-Cooper H (2010) Molecular mechanisms of gene regulation during Drosophila spermatogenesis. Reproduction. https://doi.org/10.1530/REP-09-0083

White-Cooper H, Bausek N (2010) Evolution and spermatogenesis. Philos Trans R Soc B Biol Sci. https://doi.org/10.1098/rstb.2009.0323

Wick RR, Judd LM, Gorrie CL, Holt KE (2017) Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. PLoS Comput Biol 13:e1005595. https://doi.org/10.1371/journal.pcbi.1005595

Wilson DF (1969) Cytogenetic relations in the *Drosophila nasuta* subgroup of the immigrans group of species. Univ Texas Publ 6918:207–253

Woo AL, James PF, Lingrel JB (2000) Sperm motility is dependent on a unique isoform of the Na, K-ATPase. J Biol Chem 275:20693–20699. https://doi.org/10.1074/jbc.M002323200

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586–1591. https://doi.org/10.1093/molbev/msm088

Yang Y-Y, Lin F-J, Chang H (2004) Sex ratio distortion in hybrids of *Drosophila albomicans* and *D. nasuta*. Zool Stud Taipei 43:622

Yu H, Wang W, Fang S, Zhang YP, Lin FJ, Geng ZC (1999) Phylogeny and evolution of the Drosophila nasuta subgroup based on mitochondrial ND4 and ND4L gene sequences. Mol Phylogenet Evol 13(3):556–565. https://doi.org/10.1006/mpev.1999.0667

Zhang L, Sun T, Woldesellassie F, Xiao H, Tao Y (2015) Sex ratio meiotic drive as a plausible evolutionary mechanism for hybrid male sterility. PLoS Genet 11:e1005073. https://doi.org/10.1371/journal.pgen.1005073