Recent Origins of Sperm Genes in *Drosophila*

Steve Dorus, Zoë N. Freeman, Elizabeth R. Parker, Benjamin D. Heath, and Timothy L. Karr

Department of Biology and Biochemistry, University of Bath, Bath, United Kingdom

Newly created genes often acquire testis-specific or enhanced expression but neither the mechanisms responsible for this specificity nor the functional consequences of these evolutionary processes are well understood. Genomic analyses of the *Drosophila melanogaster* sperm proteome has identified 2 recently evolved gene families on the *melanogaster* lineage and 4 genes created by retrotransposition during the evolution of the *melanogaster* group that encode novel sperm components. The expanded *Mst35B* (*protamine*) and *tektin* gene families are the result of tandem duplication events with all family members displaying testis-specific expression. The *Mst35B* family encodes rapidly evolving protamines that display a robust signature of positive selection within the DNA-binding high-mobility group box consistent with functional diversification in genome repackaging during sperm nuclear remodeling. The *Mst35B* paralogs also reside in a significant regional cluster of testis-overexpressed genes. Tektins, known components of the axoneme, are encoded by 3 nearly identical X-linked genes, a finding consistent with very recent gene family expansion. In addition to localized duplication events, the evolution of the sperm proteome has also been driven by recent retrotransposition events resulting in *Cdlc2*, *CG13340*, *Vha36*, and *CG4706*. *Cdlc2*, *CG13340*, and *Vha36* all display high levels of overexpression in the testis, and *Cdlc2* and *CG13340* reside within testis-overexpressed gene clusters. Thus, gene creation is a dynamic force in the evolution of sperm composition and possibly function, which further suggests that acquisition of molecular functionality in sperm may be an influential pathway in the fixation of new genes.

Introduction

The fundamental role of gene duplication in the evolution of functional novelty and biological diversity has long been recognized and is believed to be important to the evolution of species-specific traits (Ohno 1970). One primary mechanism of gene duplication involves segmental duplication of chromosomal regions resulting in tandem copies of genes that initially share features such as intron/exon boundaries and potentially *cis*-acting regulatory regions (Samonte and Eichler 2002). Alternatively, genes can be duplicated through retrotransposition, a process involving the insertion of a retrotranscribed mRNA molecular within a different genomic region (Long et al. 2003). This process results in a new gene lacking introns that is generally decoupled from *cis*-regulatory regions associated with the parental loci (Brosius 1999; Long et al. 2003).

Although the possible fates of newly duplicated genes have been explored from an evolutionary perspective (Force et al. 1999; Lynch and Conery 2000; Lynch and Force 2000) and, to a much lesser extent, through genetic (Loppin et al. 2005) and functional studies (Brown et al. 1998; Maston and Ruvolo 2002; Kalamegham et al. 2007), questions remain concerning the early evolution of new genes. Of particular interest are the mechanisms by which new genes acquire spatiotemporal expression patterns and how this influences the acquisition of their ultimate cellular function. For example, recent studies have demonstrated that a preponderance of new retrogenes acquire testis-specific expression both in *Drosophila* (Betran et al. 2002) and primates (Dorus et al. 2003; Emerson et al. 2004, Marques et al. 2005; Vinckenbosch et al. 2006). Despite being a widespread evolutionary process, little is known about the specific ramifications of new gene creation on testes function, spermatogenesis, or sperm composition. There are, however, 3 notable exceptions to this. The first involves the evolutionary history of *K81*, a strict paternal effect gene created through retrotransposition prior to the divergence of the *melanogaster* subgroup (Loppin et al. 2005). As a paternal effect lethal mutation, *K81* flies are viable, produce motile sperm, and have no adult phenotype. Instead, the phenotype is manifest during fertilization following sperm entrance into the egg (Yasuda et al. 1995). In wild-type eggs fertilized by sperm from *K81* homozygous males, paternal chromosomes systematically fail to properly separate sister chromatids during the first zygotic division leading to lethality early in embryogenesis (Loppin et al. 2005). This unique phenotype could result from a defect of sperm chromatin remodeling or paternal DNA replication during male pronuclear formation. Interestingly, the *K81* gene product is not an integral sperm protein but is expressed in primary spermatocytes where it presumably regulates aspects of spermatogenesis required for proper sperm function in the egg.

The second is the well-studied case of *Sdic*, a newly created gene encoding a protein present in the sperm tail (Nurminsky et al. 1998). *Sdic* is an unusual case of an X-linked chimeric gene specific to *Drosophila melanogaster* that was created through the duplication of annexinX (*AnnX*) and subsequent fusion with *Cdic*, a cytoplasmic dynein. This chimeric gene has also undergone a series of tandem gene duplication events resulting in a 4 gene cluster (Ponce and Hartl 2006). Although the precise function of *sdic* in sperm is not known, further analysis of this novel gene and other new testis-enriched genes should provide important insights into the processes involved in the acquisition of testes expression and integration of novel proteins as components of mature sperm.

The third example is the *Drosophila* gene, *mojoless* (*mjl*), which was created approximately 50 MYA through retrotransposition (Kalamegham et al. 2007). Evolutionary analysis indicate that this gene represents a retrotransposed copy of *shaggy* (*sgg*), a glycogen synthase kinase-3 encoding gene. By unknown mechanisms, *mjl* acquired male germ line expression and testis function as RNA interference (RNAi) resulted in male infertility. Interestingly, *mjl* partially rescues the *sgg* mutant phenotype indicating that it maintains some ancestral biochemical function despite its newly acquired role in male fertility.

Key words: sperm, gene duplication, retrotransposition, testis, protamines, proteomics.

E-mail: s.dorus@bath.ac.uk.

Mol. Biol. Evol. 25(10):2157–2166. 2008
doi:10.1093/molbev/msn162
Advance Access publication July 24, 2008

Genomic analyses of the *D. melanogaster* sperm proteome (DmSP) (Dorus et al. 2006) have allowed us to evaluate the importance of gene creation in the evolution of *Drosophila* sperm. These analyses determined that the *Mst35B* (*protamine*) and *tektin* gene families have been recently expanded through tandem gene duplication during *D. melanogaster* evolution. Among these gene families, positive selection driving the functional divergence of the *Mst35B* gene family was also observed. Remarkably, the X-linked *tektin* gene family is comprised of 3 closely related paralogs, with 2 of the loci sharing complete nucleotide identity. This finding is indicative of an extremely recent set of duplication events during *D. melanogaster* evolution. A survey of retrotransposed genes also identified 4 newly created genes, *CG13340*, *Vha36*, *CG4706*, and *Cdlc2*, which encode integral sperm components. Finally, *Mst35B* genes, the *tektins*, and *Cdlc2* are demonstrated to have testis-specific expression patterns. Interestingly, the *Mst35B* genes, *Cdlc2* and *CG13340* also localize to genomic regions with significant clustering of testis-overexpressed genes.

The use of whole-sperm proteomics provides an alternative experimental approach for the study of new gene creation as it focuses specifically on proteins that are developmentally incorporated as integral sperm components. This also eliminates a reliance on gene expression in the assessment of putative functionality of novel genes. Our analysis of the DmSP thus provides a new perspective on the dynamic influence of gene creation in spermatogenesis, and specifically on sperm evolution, and suggests that the acquisition of sperm functionality may be a common evolutionary conduit for the fixation and functional evolution of new genes.

Materials and Methods

Bioinformatic Identification of Novel Sperm Genes in *D. melanogaster*

Nucleotide coding sequences were downloaded from FlyBase (http://flybase.bio.indiana.edu/) for all sperm components characterized using whole-sperm mass spectrometry (MS) (Dorus et al. 2006). Local batch BlastN was used to compare these sequences against all coding gene sequences in the *D. melanogaster* genome (R5.7). Any comparison with significant nucleotide identity (>70% identity with >50% coding sequence coverage) was downloaded for further analysis. "In-frame," pairwise ClustalX alignments of the nucleotide sequences were conducted for these genes, and pairwise *d*S estimates were calculated using the method of Nei and Gojobori (1986). Candidate gene duplication events specific to *D. melanogaster* were tentatively identified by choosing paralogs with *d*S values lower than 0.175, a value slightly greater than the average synonymous divergence between *D. melanogaster* and *Drosophila simulans* orthologs (Dorus et al. 2006). Genes identified by these criteria were subject to further comparative genomic analyses (see below). A complete list of pairwise comparisons with dS less than 0.175 is provided in supplementary table 1 (Supplementary Material online).

Comparative Genomic and Evolutionary Analyses

The evolutionary history of putative duplication events was inferred from evolutionary divergence estimates complemented by comparisons of syntentic genomic regions in the *melanogaster* group and *Drosophila pseudoobscura* (Clark et al. 2007). Duplication events determined to be specific to the *D. melanogaster* lineage were subject to additional analyses. Available orthologous *Drosophila* coding sequences were aligned in-frame using ClustalX. Phylogenies were determined by exhaustive parsimony analysis as implemented by PAUP (Swofford 2003), and sequence analysis and ancestral node reconstruction were conducted using maximum likelihood methods implemented by the codeml program in the PAMLv3.14 package (Yang 1996). Branch-specific ω values were obtained using the free-ratio model and branch-specific rate acceleration assessed using 2-ratio models (Yang 1998). Pairwise *d*N and *d*S values were estimated using the method of Nei and Gojobori (1986). Model A of codeml was used to detect positive selection on codons along specific foreground branches (Zhang et al. 2005). To test the significance of the results from Model A, we used the more stringent comparison of Model A to itself with ω constrained to 1.0 (Test 2; Model A fix) and also the more conservative χ^2 to calculate *P* (Zhang et al. 2005). Parameter estimates and log-likelihood values under different models are provided in supplementary table 2 (Supplementary Material online). Likelihood ratio statistics ($2\Delta\ell$) for alternative hypothesis testing and positively selected sites by Bayes empirical methods are provided in supplementary table 3 (Supplementary Material online). Sliding-window analyses were conducted using the Swaap v. 1.0.2 software package with a window size of 60 bp and a sliding increment of 3 bp (Pride and Blaser 2002).

Reverse Transcriptase–Polymerase Chain Reaction

Ten male carcasses (minus gonads) and testes from 25 male equivalents were dissected, washed (3×) in phosphate-buffered saline, as described by Snook and Markow (2002), and flash frozen in liquid nitrogen. Total RNA was then isolated using manufacturer's protocols (Ambion RNAqueous Kit), and RNA was resuspended in TE and quantified. Reverse transcription was carried out by standard protocols (Promega Improm-II Reverse Transcription System) using 75-ng input RNA (final concentration 3.75 ng/μl) from whole males (minus gonads) and testis. Gene-specific polymerase chain reaction (PCR) primers were designed for all newly created genes with the exception of the *tektin* gene family members (primer sequences are provided in supplementary table 4, Supplementary Material online). *Tektin* gene family mRNA sequences are nearly identical (>99%) at the nucleotide level thus precluding gene-specific PCR amplification. PCR was conducted using equal quantities of complementary DNA from the whole fly (minus gonads) and testis, and equal volumes of each PCR were loaded onto the same 1% agarose gel. Optical density of the relevant bands were quantified using AIDA software (RayTest, Pforzheim, Germany).

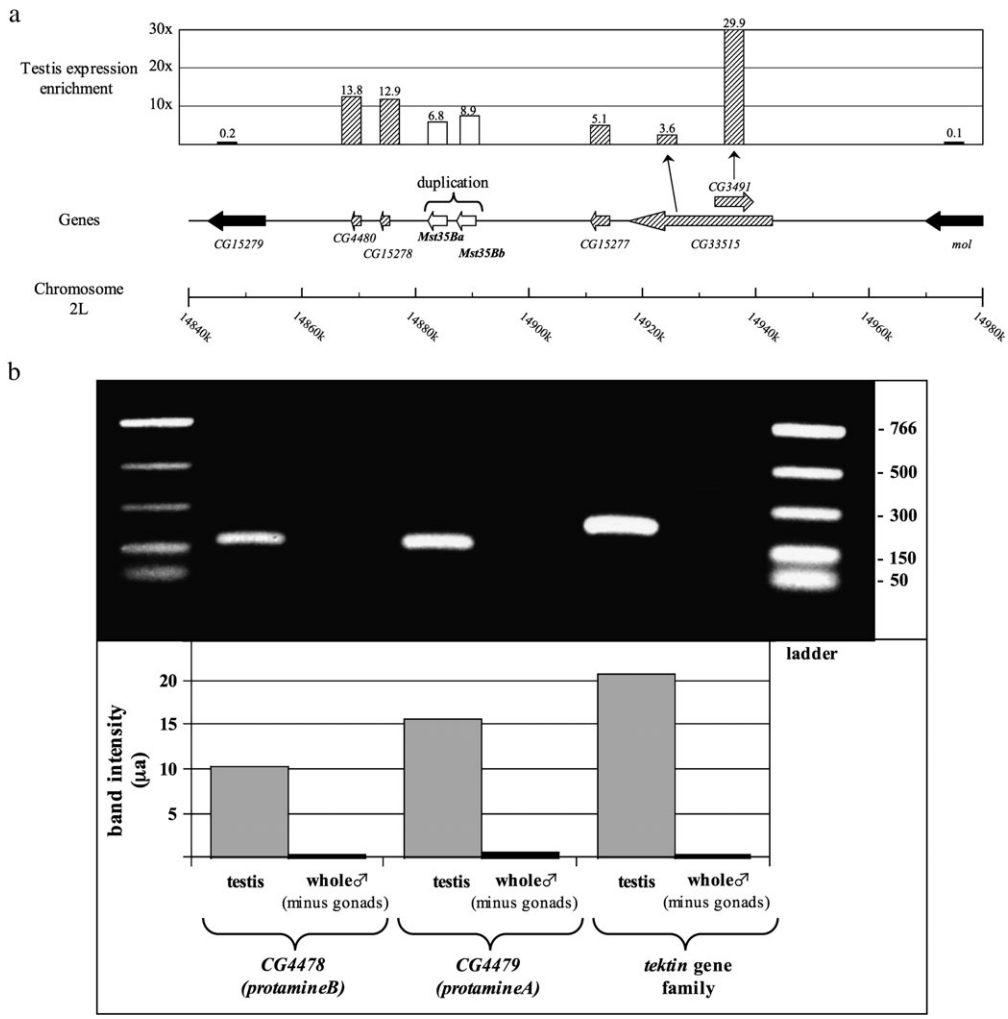

FIG. 1.—(*a*) Testis-overexpressed gene cluster surrounding the recently created *Mst35Ba* and *Mst35Bb* paralogs encoding protamines. Upper panel: levels of testis expressional enrichment in comparison to expression in the whole fly from Chintapalli et al. (2007). (*b*) Analysis of *Mst35Ba*, *Mst35Bb*, and *tektin* gene expression. Testis-specific expression detected by RT-PCR (upper panel). Optical density measurement comparing relative gene expression in the testis versus gonadectomized males (lower panel).

Due to the extremely high level of nucleotide identity between *tektin* paralogs, a detailed inspection of expressed sequence tag (EST) sequences was required to assess gene-specific expression patterns. We, therefore, examined the *D. melanogaster* EST database (http://www.ncbi.nlm.nih.gov/) using BlastN and manually determined the gene from which each EST originated when possible. Surveys of expression data were also conducted from previous microarray studies (Parisi et al. 2004; Chintapalli et al. 2007) to confirm testis overexpression compared with the whole fly and to analyze the expression profiles of neighboring genes in regions containing newly created sperm genes (see below).

Genomic Analyses of Gene Clustering Based on Testis Overexpression

Drosophila melanogaster microarray gene expression data were obtained from Chintapalli et al. (2007) (http://www.flyatlas.org/). An average was taken for the expression quantification data for all genes assayed in duplicate within this data set. The identification of testis-overexpressed gene clusters was performed using 2 alternative approaches based upon a stringent adjacent gene model (Li et al. 2005). The first approach, based on our empirical observations (see figs. 1 and 4*a*), determined the genome-wide probability of finding 6 genes with >5.0-fold enrichment in the testis (compared with the whole fly) in a group of 7 adjacent genes. The second approach calculated the probability that the average fold enhancement of testis expression (compared with the whole fly) for 7 adjacent genes was equal to or greater than that observed for the genomic regions containing *Mst35B* (10.62×) and *Cdlc2* (9.21×). The probability of retrotransposed genes relocating next to or within previously existing testis-overexpressed gene clusters was calculated assuming a stochastic model of gene insertion. Determination of all possible insertion locations was based on the total number of genes in the *melanogaster* genome (R5.7) and significant testis-overexpressed gene clusters were

Table 1
MS Identification and Testis Enrichment of Novel Sperm Genes in *Drosophila melanogaster*

Gene Name	Peptide Fragments Identified by MS	Testis Enrichment [a]	EST Data[b] (no. of hits)
1. *Drosophila melanogaster* protamines			
Mst35Ba (dProtamineA)	QGPVTNNAYLNFVR	6.78×	testis: 6; all other: 0
Mst35Bb (dProtamineB)	Same as above	8.65×	testis: 8; all other: 3
2. *Drosophila melanogaster tektin* gene family			
CG17450 (*tektin*)	WPTADMNEYNERKDHR YTSNEWYNNNMTK HMPEQPVTNQLTK ITDLTFWR LIAEMSDINELQR ALLVEINNLR VNALFIDR VQQELFDMEK NELNAELEK HLFLLQK LGGLHEK	14.37×	Unique ESTs—testis: 2; all other: 1 Nonunique ESTs—testis: 33; all other: 5
CG32819 (*tektin*) *CG32820* (*tektin*)	Same as above	Not analyzed by microarray	Unique ESTs—testis: 3; all other: 1 Nonunique ESTs—testis: 33; all other: 5

[a] Microarray data from Parisi et al. (2004) and Chintapalli et al. (2007) are consistent with the RT-PCR results presented in figure 1*b*.
[b] High levels of nucleotide identity among *tektin* gene family precluded unique assignment of most EST sequences to a particular loci. Those that were unique confirm testis-biased expression of all 3 loci.

defined as any adjacent set of 7 genes that include 6 or more genes with >5.0-fold expressional enrichment in the testis.

Results

Mst35B Protamine Gene Duplication in *D. melanogaster*

MS of purified sperm identified a tryptic peptide encoded by *Mst35Ba* and *Mst35Bb*, genes encoding the integral sperm proteins, dProtamineA and dProtamineB (table 1). Both of these protamines have also been previously demonstrated to localize to sperm (Raja and Renkawitz-Pohl 2005). *Mst35Ba* and *Mst35Bb* represent a gene duplication specific to *D. melanogaster* as levels of synonymous divergence (*d*S) were significantly lower between *Mst35B* paralogs than orthologous comparisons between *D. melanogaster* and the single copy present in *D. simulans*. In concert with this result, divergence at these loci was also significantly lower than the average *d*S between *D. melanogaster* and *D. simulans* (supplementary table 5, Supplementary Material online). Comparative genomic analysis within the *melanogaster* subgroup identified a single orthologous *Mst35B* gene with syntenic correspondence. Finally, close genomic proximity of *Mst35Ba* and *Mst35Bb* (449 bp) and the conservation of gene structure (intron/exon boundaries) between these paralogs suggest that they were created through a tandem duplication event (fig. 1*a*).

Mst35B Paralogs Are Testis-Specific and Reside in a Testis-Overexpressed Gene Cluster

As expected, given their presumed function in genome repackaging during spermatogenesis, both *Mst35B* paralogs are significantly overexpressed in the testis in comparison to the whole fly (minus gonads) as determined by gene-specific reverse transcriptase–polymerase chain reaction (RT-PCR; fig. 1*b*). This finding is consistent with 2 previous microarray studies (Parisi et al. 2004; Chintapalli et al. 2007) and our own analysis of *D. melanogaster* EST databases (table 1). The *Mst35B cis*-acting promoters have not been characterized, so it cannot be determined if promoter regions have been duplicated along with the protein-coding exons. However, *Mst35B* loci are embedded within an ~70-kb region containing 5 additional genes overexpressed in the testis (fig. 1*a*). Two analyses of the genome-wide clustering of testis-overexpressed genes confirm significant clustering within this region. First, an analysis of genome regions containing at least 6 out of 7 genes with >5-fold enrichment in the testis demonstrates that they are exceedingly rare with a frequency of 0.00093 within the genome. Second, the 10.62-fold average testis enrichment in this gene cluster is extremely high and also found at an extremely low frequency in the remainder of the genome (0.0013). Despite the concentration of testis-overexpressed genes in this genomic region, only the *Mst35B* paralogs encode empirically identified sperm proteins (Dorus et al. 2006).

Positive Selection on *Mst35B* during *D. melanogaster* Evolution

Maximum likelihood analyses provided evidence of positive selection in the recent evolutionary history of these paralogs. First, a significant evolutionary rate increase was observed for the *Mst35Ba* lineage following the duplication event ($P = 0.012$) in comparison to the remainder of the phylogeny (fig. 1*b*; supplementary tables 2 and 3, Supplementary Material online). Although branch-specific maximum likelihood analyses did not identify the *Mst35Ba* lineage ($\omega = 2.9$) as significantly greater than 1.0, branch-site models (Model A) provided significant evidence for the influence of positive selection on a class of sites on this lineage ($P = 0.029$). Additionally, a significant asymmetry in evolutionary rates is observed between the *Mst35Ba* and *Mst35Bb* lineages (fig. 2). Although asymmetry is common (Conant and Wagner 2003; Fares et al. 2006), significant rate asymmetry is less likely among local "DNA-based" duplicates (Cusack and Wolfe 2007).

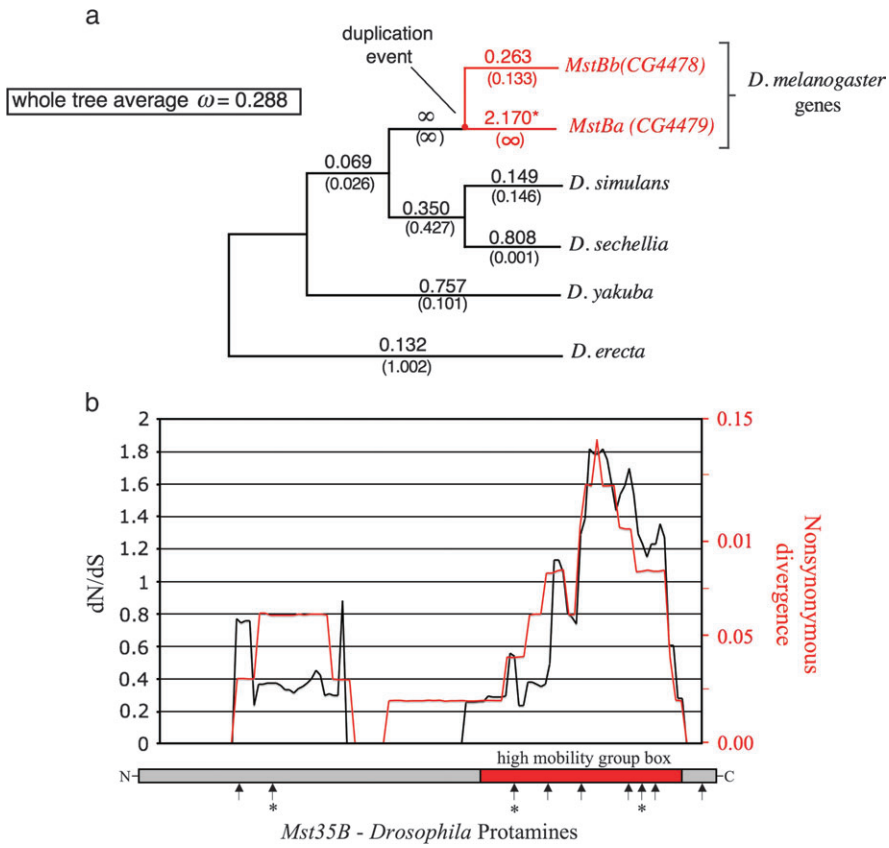

FIG. 2.—Maximum likelihood analysis of *Mst35B* gene family in *Drosophila*. Values above lines indicate branch-specific ω values (*d*N/*d*S) for the entire coding sequence, and values in parenthesis under individual branches indicate ω values for the HMG box domain. Red lines represent branches of the phylogenetic tree that depict gene evolution following the gene duplication event, and the asterisk marks the *Mst35Ba* lineage where significant evidence for positive selection was observed. Phylogenetic tree is not drawn to scale. (*b*) Sliding-window analysis comparing *Mst35Ba* and *Mst35Bb*. The HMG box domain corresponds precisely with the primary region of the gene undergoing accelerated evolution. Arrows indicate sites in the ω_2 category on the *Mst35Ba* foreground lineage (asterisks indicate sites with posterior probabilities greater than 0.90). Six of the 9 sites are within the HMG domain, which represents a significant concentration in comparison to the remainder of the gene.

Protamines contain a DNA-binding high-mobility group (HMG) box domain that has been implicated in both gene regulation (Travers 2002) and chromatin remodeling (Ragab et al. 2005). A sliding-window analysis of the *Mst35B* paralogs distinctly identified the HMG box domain as the rapidly evolving region within the gene (fig. 2*b*). Consistent with this observation, an $\omega > 1.0$ was observed on the *MstBa* lineage for this domain (fig. 2*a*) that contained 6 sites in the ω_2 ($\omega > 1.0$) category using branch-site maximum likelihood analysis (fig. 2*b*). Finally, the concentration of these codon sites within the 44 amino acid HMG box domain is significantly higher than the 3 sites identified in the remainder of the gene ($P = 0.032$; 2-tailed Fisher's exact test). This finding is suggestive of functional diversification of the HMG box and consistent with the rapid evolution of DNA-/RNA-binding sperm proteins in *Drosophila* (Dorus et al. 2006) and protamines in other taxa (Wyckoff et al. 2000).

Extremely Recent Expansion of an X-linked Sperm Structural Gene Family

Our previous analysis of the DmSP also identified 11 unique protein fragments specific to the proteins encoded by the X-linked *tektin* gene family, which includes *CG17450* and its paralogs, *CG32820* and *CG32819* (table 1). Due to the very high level of amino acid identity, MS was unable to determine if all, or only a subset, of the tektins are present in mature sperm. However, tektins are known sperm axonemal microtubule–binding proteins (Cao et al. 2006) and given their testis-specific expression (see below), it is reasonable to assume that all 3 loci encode sperm structural proteins.

CG17450, *CG32820*, and *CG32819* were identified as recent duplicates in *D. melanogaster* based on the extremely low levels of synonymous divergence between them (supplementary table 5, Supplementary Material online). Remarkably, no nucleotide divergence (either synonymous or nonsynonymous) exists between *CG32819* and *CG32820*, a finding suggestive of a very recent evolutionary origin. Comparative genomic analyses also confirmed the recent timing of these duplication events: single regions with syntenic correspondence to *D. melanogaster CG17450* were identified throughout the *melanogaster* subgroup, whereas *CG32820* and *CG32819* are unique to *D. melanogaster*. Colocalization and conservation of itron/exon boundaries strongly suggest that expansion of the gene cluster occurred through local tandem duplications. The most parsimonious explanation is that duplication of the ancestral region (also including the ancestor of *CG33502*) created a second region containing either

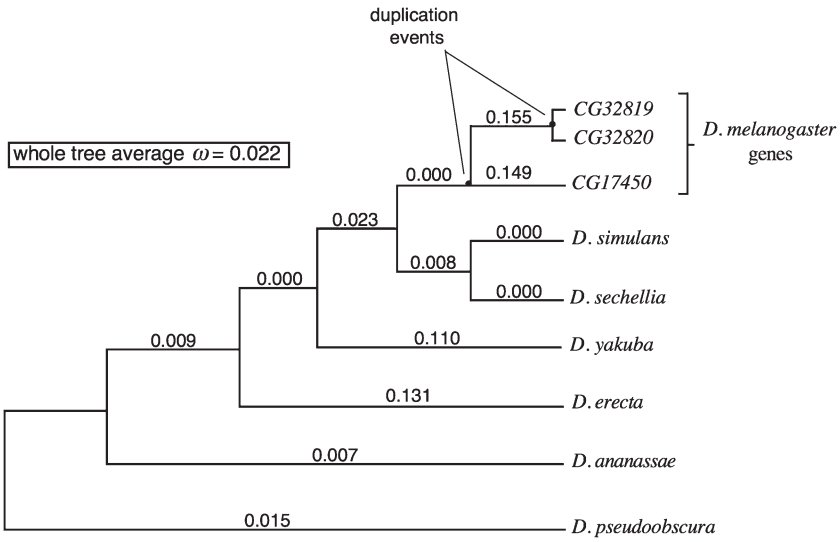

FIG. 3.—Maximum likelihood analysis of *tektin* gene family *CG17450* in *Drosophila*. Values indicate branch-specific ω values (*d*N/*d*S). Note that *CG32819* and *CG32820* share 100% nucleotide identity, and therefore, no ω values are presented. Phylogenetic tree is not drawn to scale.

CG32820/*CG32857* or *CG32819*/*CG33500* followed by another more recent 2 gene duplication event, resulting in the final gene pair (supplementary fig. 1, Supplementary Material online). Complete nucleotide identity exists between the ~6-kb region containing genes *CG32820*/*CG32857* and *CG32819*/*CG33500*.

RT-PCR assays of the *tektin* loci demonstrated specific expression in the testis (fig. 1*b*). Nucleotide identity among these loci precludes the design of gene-specific primers. However, at least 2 of the 3 loci are expressed and have enhanced expression in the testis based on detailed analysis of ESTs (table 1). Unlike the *Mst35B* gene family, no evidence exists for clustering of testis-overexpressed loci in genomic regions around these tektins (supplementary fig. 1, Supplementary Material online). Maximum likelihood analysis of the *tektin* gene family failed to identify the evidence of accelerated evolution or the signature of positive selection (fig. 3). In fact, *d*N/*d*S ratios are significantly below 1.0 on all lineages of the phylogeny, and only a modest increase is observed on the lineages following gene duplication events.

Sperm Genes Created through Retrotransposition

The DmSP contains 4 previously identified retrogenes created during the evolution of the *melanogaster* group (Betran et al. 2002) (table 2). However, only 2 of the parental loci (*CG32063* and *ctp*) were also found in the DmSP. The other 2 retroposed DmSP genes arose from parental loci not highly expressed in the testis and therefore acquired testis expression and sperm localization by other mechanisms following their creation. Although based on a limited number of events, the status of parental loci as a sperm gene does not appear to be a prerequisite for the future acquisition of sperm function by the retrogene. However, retrotransposition of sperm genes consistently results in the creation of novel sperm retrogenes. It is also noteworthy that 2 of these events involved the retrotransposition of X-linked genes to create autosomal sperm loci (table 2). Analysis of a complete suite of retroposition events in *Drosophila* will be needed to fully characterize the dynamics of retrogene inclusion in the DmSP.

Retrotransposition into Testis-Overexpressed Gene Clusters

Given that retrotransposition events do not usually transfer *cis*-acting regulatory sequences, we sought to determine if the establishment of testis-specific expression may be related to expressional characteristics of the genomic region within which the new gene has been relocated.

Table 2
Recent Retrotransposition Events Resulting in the Creation of Novel Sperm Components

Parental Gene[a]				New Gene[a]				
Gene	Unique Sperm MS ID[b]	Genome Location	Testis Enrichment[c]	Gene	Unique Sperm MS ID[b]	Genome Location	Testis Enrichment[c]	Gene Function
CG32063	Yes	3L: 67E	6.06×	*CG13340*	Yes	2R: 5 0C	3.13×	Leucyl aminopeptidase
ctp	Yes	X: 4C	No	*Cdlc2*	Yes	2L: 22A	8.67×	Dynein light chain
CG8310	No	X: 3A	No	*Vha36*	Yes	2R: 52A	No	Proton transporter
Acon	No	2L: 39B	No	*CG4706*	Yes	3R: 86D	14.48×	Aconitase

[a] Retrotransposed genes were previously characterized in Betran et al. (2002).
[b] Full list of peptide fragments identified by whole-sperm MS can be found in Dorus et al. (2006).
[c] Microarray data from Parisi et al. (2004) and Chintapalli et al. (2007) are consistent with RT-PCR data presented in figure 4*a*.

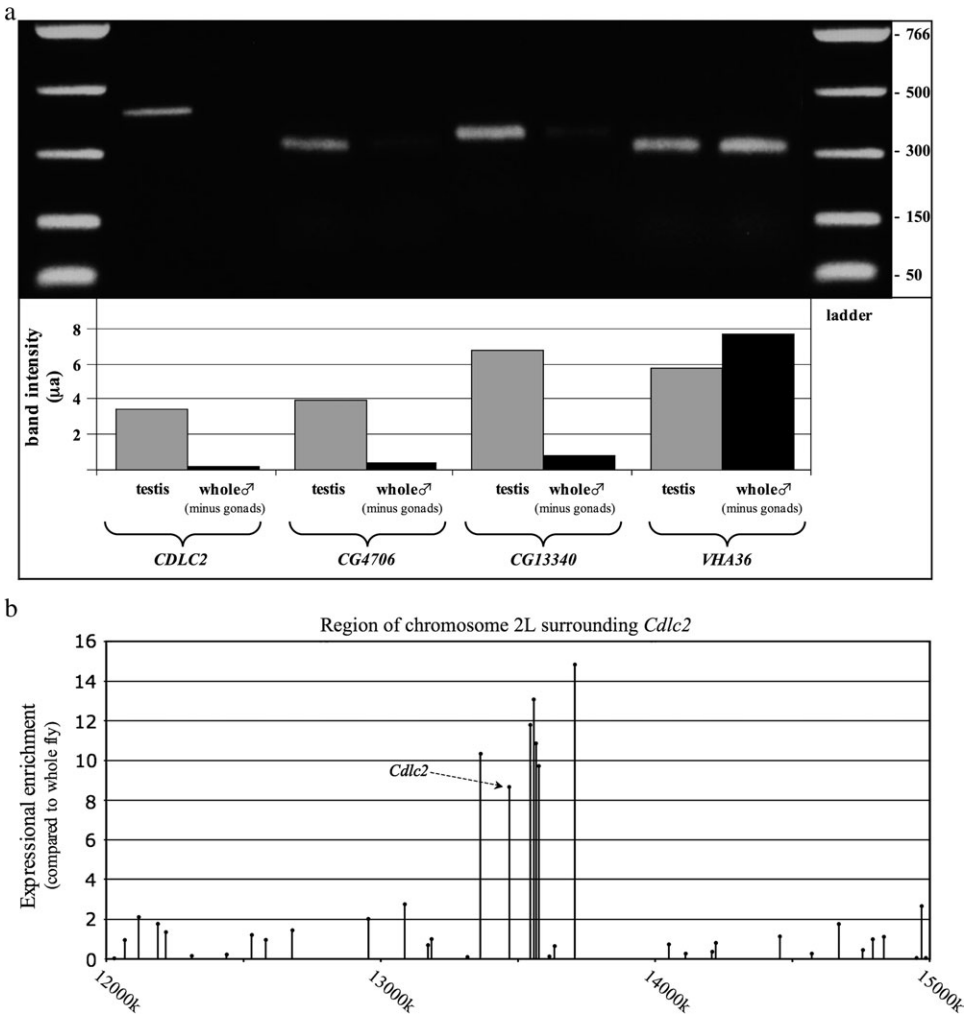

FIG. 4.—(*a*) RT-PCR of retrotransposed genes encoding sperm proteins (upper panel). Optical density measurement comparing relative gene expression in the testes versus gonadectomized males (lower panel). Testis-specific expression was observed for *Cdlc2*, whereas testis overexpression was observed for *CG4706* and *CG13340*. *Vha36* appears to be more widely expressed throughout the whole fly. (*b*) Testis-overexpressed gene cluster surrounding novel sperm retrogene, *Cdlc2*. Levels of enrichment in the testis in comparison to the whole fly were obtained from microarray analyses by Chintapalli et al. (2007). Enrichment levels are plotted corresponding to the transcriptional start site for each gene.

Our analysis identified one retrogene, *Cdlc2*, which resides in a ~15-kb genomic region with expressional characteristics consistent with this possibility. *Cdlc2* is testis-specific by RT-PCR (fig. 4*a*) and retrotransposed into a significant cluster of testis-overexpressed genes (9.21-fold average expressional enrichment in the testis, fig. 4*b*). The parental gene of *Cdlc2*, *ctp*, is present in the DmSP but not overexpressed in the testis. Clusters of testis-overexpressed genes with common properties of the *Cdlc2* region are exceedingly rare in the genome, comprising less than 0.01% of all possible adjacent groups of 7 genes. The probability of a retrogene inserting within such a cluster is also exceedingly small (0.7%) further suggesting a role for regional expression regulation in the establishment of testis-specific expression. It is also noteworthy that the testis-overexpressed retrogene, *CG13340*, has also been previously demonstrated, using a different analytical method, to reside in a much larger genomic region (~234 kb) enriched for testis-expressed genes (Parisi et al. 2004).

Discussion

Gene duplication, and the subsequent evolution of genetic novelty, is now recognized as a critical process in evolution (Ohno 1970; Long et al. 2003). Although understood as a widespread process, the specific molecular impact of gene duplication on the evolution of protein function is still being elucidated. Furthermore, the complexity of whole-cell systems makes it technically difficult to integrate the role of gene duplication and neofunctionalization into a framework of cellular function and evolution. We have approached this general problem by first identifying the principle protein components of whole *D. melanogaster* sperm, a cell type of central importance to sexual reproduction and Darwinian fitness (Dorus et al. 2006). Sperm proteome characterization forms a powerful basis for the study of gene creation and function as it allows for the conclusive identification of new genes encoding novel sperm components during spermatogenesis.

Our previous evolutionary analysis of the DmSP did not identify widespread rapid evolution of genes encoding sperm components (Dorus et al. 2006). The absence of rapid evolution across the DmSP is perhaps not surprising as many integral sperm proteins are involved in essential and oftentimes ubiquitous functions, such as metabolism and energetics. This finding highlights the central role of sperm in sexual reproduction and likely reflects the ancient evolutionary history of sperm (Baccetti 1982, 1986). In this study, we have utilized an alternative approach to specifically determine if gene creation is a mechanism driving the evolution of sperm form and function and demonstrate that *Drosophila* sperm has, in fact, been impacted by the recent origination and diversification of novel sperm genes.

Proteomic and comparative genomic approaches used in this study revealed the expansion of the *Mst35B* and *tektin* gene families during the evolution of *D. melanogaster*. In both cases, gene family expansion occurred through tandem duplication events and resulted in the creation of testis-specific sperm genes. Despite these similarities, the selective forces influencing the evolutionary histories of these expanded gene families are quite distinct. In the case of *Mst35B*, rapid evolution and a significant signature of positive selection was observed that is highly concentrated within the DNA-binding HMG box domain. This finding is consistent with the evolution of the DmSP where DNA-/RNA-binding proteins were found to evolve at an accelerated rate similar to that of male accessory gland genes (Dorus et al. 2006). It is true, however, that relaxation of functional constraint could also contribute to the observed evolutionary acceleration of these new genes. HMG box domains are found in a variety of critical proteins essential for sex determination or sex differentiation, such as the testis-determining *SRY* and *SOX9* genes in mammals (Harley et al. 2003) and the mating-type (*MAT*) locus of yeast (Butler et al. 2004). Protamines have also been shown to be influenced by positive selection in primates (Wyckoff et al. 2000), but the underlying selective forces are unclear (Rooney and Zhang 1999). It is also known that protamine mutants are not haploinsufficient in *Drosophila*, but detailed complementation analyses of *Mst35Ba* and *Mst35Bb* mutants have yet to be conducted (Raja and Renkawitz-Pohl 2005). Such analyses will be crucial in determining to what extent the rapid evolution of the HMG box domain has resulted in neofunctionalization that may be relevant to the process of genome condensation and packaging during spermatogenesis.

The *tektin* gene family, on the other hand, has originated much more recently and appears to be under strong selective constraint. This gene family encodes tektin-like cytoskeletal proteins that interact directly with microtubules in the sperm axoneme (Steffen and Linck 1988), and their more constrained evolutionary characteristics are consistent with the conservative evolution of sperm structural components (Dorus et al. 2006). Selective pressures, associated with expression level requirements during spermatogenesis, may have driven the rapid and recent increase in copy number. We note that 2 of the *tektin* genes, *CG32819* and *CG32820*, are identical at the nucleotide level suggesting an extremely recent duplication event. Our comparative analysis of all *D. melanogaster* genes reveals that the incidence of genes sharing identical exon and intron sequences is rare (<0.3% of all genes) and that this gene pair may therefore be among the youngest genes in the genome. Population genetic screens will be essential to determine if the most recent duplication event is fixed within worldwide *melanogaster* populations or whether it represents a copy number polymorphism.

Although evident from comparative genomic analyses that this duplication event occurred on the *melanogaster* lineage, it could be posited that gene conversion is responsible for complete sequence homogenization between *tektin* gene copies. Explicit tests for gene conversion are confounded by the fact that the gene cluster itself originated through unequal crossover events. However, we do not favor gene conversion as an explanation for several reasons. First, complete sequence identity exists between the regions containing *CG32819* and *CG32820*, whereas the region containing *CG17450* is divergent throughout its length. It seems improbable that pervasive gene conversion occurs between 2 of the 3 neighboring loci without also homogenizing the third loci to an appreciable and detectable level. Second, complete nucleotide identity exists between 2 neighboring ~6-kb regions (supplementary fig. 1, Supplementary Material online). Complete homogenization of such large regions is at odds with previous estimates of gene conversion tract length (~750 bp) in *Drosophila* (Hilliker et al. 1994) and more recent estimates in experimental mammalian systems (Ruksc et al. 2008). Lastly, population genetic studies indicate that gene conversion is less influential as a homogenizing force among X-linked paralogs than those located on autosomes (Thornton and Long 2005). We therefore favor the interpretation that the *tektin* gene cluster is the result of very recent gene duplication events, the last of which may not have yet fixed in worldwide *D. melanogaster* populations.

Our discovery of the *tektin* sperm gene cluster on the *Drosophila* X chromosome is somewhat unexpected for several reasons. First, it has been demonstrated that the *Drosophila* X chromosome contains a paucity of testis-overexpressed genes and testis-enriched gene clusters (Boutanaev et al. 2002; Parisi et al. 2003, 2004). Second, there is also a paucity of integral sperm genes on the *Drosophila* X (Dorus et al. 2006). Finally, it has been shown that genes with male-biased expression or functionality tend to move off of the X chromosome and onto the autosomes during evolution (Betran et al. 2002; Dai et al. 2006; Shiao et al. 2007). To our knowledge, only one other example of a newly created, testis-specific X-linked gene cluster has been reported in *D. melanogaster*, the *Sdic* gene cluster (Ponce and Hartl 2006). Therefore, the *tektin* and *Sdic* gene clusters represent an unusual and rare class of recently expanded X-linked sperm structural gene families.

It has now been demonstrated in both primates and *Drosophila* that genes created through retrotransposition often acquire testis expression (Betran et al. 2002; Dorus et al. 2003; Vinckenbosch et al. 2006; Bai et al. 2007; Shiao et al. 2007). Our results extend this observation by demonstrating that some retrotransposed genes in *Drosophila* are not only expressed during spermatogenesis but also encode protein components of sperm. Although the mechanisms responsible for the evolution of testis expression are not

well understood, it has been argued that large, chromatin-mediated regulatory domains within the genome may promote testis expression, a hypothesis supported by the existence of testis-enriched clusters of genes (Boutanaev et al. 2002; Parisi et al. 2004; Divina et al. 2005). Our identification of regions of significant clustering that contain new sperm genes also points to a possible link between regional gene regulation mechanisms, the establishment of testis expression, and subsequent functional acquisition. A more thorough analysis of the genomic distribution of testes-overexpressed genes will be necessary to determine the extent to which such a functional and mechanistic link exists. However, alternative mechanisms cannot be ruled out, including promoter sharing with neighboring genes (Loppin et al. 2005), the direct inheritance of parental promoters (Feral et al. 2001), the use of promoter elements from nearby transposable elements (van de Lagemaat et al. 2003), and the acquisition of new untranslated regions from neighboring sequences (Shiao et al. 2007).

Little is currently known about the mechanisms underlying how, when or why proteins become incorporated into a functional sperm during evolution. It is clear from our previous study that integral sperm genes are nonrandomly distributed throughout the genome suggesting higher order regulatory properties may be involved. However, the retrogenes described in this study do not localize to any of the previously characterized DmSP gene clusters. Ultimately, a variety of complementary experimental approaches will be needed to elucidate the cellular and developmental mechanisms responsible for the inclusion of testis-expressed genes as novel sperm components.

This study highlights the influence of new gene creation on the evolution of *Drosophila* sperm and provides strong evidence for the influence of positive selection on the functional divergence of the recently created *Mst35B protamine* paralogs. Analysis of newly retrotransposed sperm genes also suggests that localization in clusters of testis-overexpressed genes may be associated with the establishment of testis expression and acquisition of male-biased functions. The acquisition of functionality in sperm, and more generally in spermatogenesis, appears to be a common conduit in the fixation and early functional evolution new genes. Future comparative evolutionary genomics and sperm proteomics will provide the needed functional knowledge base for a more comprehensive understanding of the ultimate acquisition of new gene functionality and the process of cellular evolution through gene creation.

Supplementary Material

Supplementary figure 1 and tables 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank 2 anonymous reviewers for their insightful comments and Steve Russell for detailed discussions concerning the characteristics of high-mobility group box domains. This work was supported in part by a National Research Services Award from the National Institutes of Health (T.L.K./S.D.), a Research Council of the United Kingdom Fellowship (S.D.), a Wolfson Merit award from the Royal Society (T.L.K.), and the Biotechnology and Biological Sciences Research Council (T.L.K.).

Literature Cited

Baccetti B. 1982. The evolution of the sperm tail. Symp Soc Exp Biol. 35:521–532.

Baccetti B. 1986. Evolutionary trends in sperm structure. Comp Biochem Physiol A. 85:29–36.

Bai Y, Casola C, Feschotte C, Betran E. 2007. Comparative genomics reveals a constant rate of origination and convergent acquisition of functional retrogenes in *Drosophila*. Genome Biol. 8:R11.

Betran E, Thornton K, Long M. 2002. Retroposed new genes out of the X in *Drosophila*. Genome Res. 12:1854–1859.

Boutanaev AM, Kalmykova AI, Shevelyov YY, Nurminsky DI. 2002. Large clusters of co-expressed genes in the *Drosophila* genome. Nature. 420:666–669.

Brosius J. 1999. RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene. 238:115–134.

Brown CJ, Todd KM, Rosenzweig RF. 1998. Multiple duplications of yeast hexose transport genes in response to selection in a glucose-limited environment. Mol Biol Evol. 15:931–942.

Butler G, Kenny C, Fagan A, Kurischko C, Gaillardin C, Wolfe KH. 2004. Evolution of the MAT locus and its Ho endonuclease in yeast species. Proc Natl Acad Sci USA. 101:1632–1637.

Cao W, Gerton GL, Moss SB. 2006. Proteomic profiling of accessory structures from the mouse sperm flagellum. Mol Cell Proteomics. 5:801–810.

Chintapalli VR, Wang J, Dow JA. 2007. Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. Nat Genet. 39:715–720.

Clark AG, Eisen MB, Smith DR, Bergman CM, Oliver B, Markow TA. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. Nature. 450:203–218.

Conant GC, Wagner A. 2003. Asymmetric sequence divergence of duplicate genes. Genome Res. 13:2052–2058.

Cusack BP, Wolfe KH. 2007. Not born equal: increased rate asymmetry in relocated and retrotransposed rodent gene duplicates. Mol Biol Evol. 24:679–686.

Dai H, Yoshmatsu TF, Long M. 2006. Retrogene movement within- and between-chromosomes in the evolution of *Drosophila* genomes. Gene. 385:96–102.

Divina P, Vlcek C, Strnad P, Paces V, Forejt J. 2005. Global transcriptome analysis of the C57BL/6J mouse testis by SAGE: evidence for nonrandom gene order. BMC Genomics. 6:29.

Dorus S, Busby SA, Gerike U, Shanbanowitz J, Hunt DF, Karr TL. 2006. Genomic and functional evolution of the sperm proteome. Nat Genet. 38:1440–1445.

Dorus S, Gilbert SL, Forster ML, Barndt RJ, Lahn BT. 2003. The CDY-related gene family: coordinated evolution in copy number, expression profile and protein sequence. Hum Mol Genet. 12:1643–1650.

Emerson JJ, Kaessmann H, Betran E, Long M. 2004. Extensive gene traffic on the mammalian X chromosome. Science. 303:537–540.

Fares MA, Byrne KP, Wolfe KH. 2006. Rate asymmetry after genome duplication causes substantial long-branch attraction artifacts in the phylogeny of *Saccharomyces* species. Mol Biol Evol. 23:245–253.

Feral C, Guellaen G, Pawlak A. 2001. Human testis expresses a specific poly(A)-binding protein. Nucleic Acids Res. 29:1872–1883.
Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. Genetics. 151:1531–1545.
Harley VR, Clarkson MJ, Argentaro A. 2003. The molecular action and regulation of the testis-determining factors, SRY (sex-determining region on the Y chromosome) and SOX9 [SRY-related high-mobility group (HMG) box 9]. Endocr Rev. 24:466–487.
Hilliker AJ, Harauz G, Reaume AG, Gray M, Clark SH, Chovnick A. 1994. Meiotic gene conversion tract length distribution within the rosy locus of *Drosophila melanogaster*. Genetics. 137:1019–1026.
Kalamegham R, Sturgill D, Siegfried E, Oliver B. 2007. *Drosophila mojoless*, a retroposed *GSK-3*, has functionally diverged to acquire an essential role in male fertility. Mol Biol Evol. 24:732–742.
Li Q, Lee BTK, Zhang L. 2005. Genome-scale analysis of positional clustering of mouse testis-specific genes. BMC Genomics. 6:7.
Long M, Betran E, Thornton K, Wang W. 2003. The origin of new genes: glimpses from the young and old. Nat Rev Genet. 4:865–875.
Loppin B, Lepetit D, Dorus S, Couble P, Karr TL. 2005. Origin and neofunctionalization of a *Drosophila* paternal effect gene essential for zygote viability. Curr Biol. 15:87–93.
Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. Science. 290:1151–1155.
Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics. 154:459–473.
Marques AC, Dupanloup I, Vinckenbosch N, Reymond A, Kaessmann H. 2005. Emergence of young human genes after a burst of retroposition in primates. PLoS Biol. 3:e357.
Maston GA, Ruvolo M. 2002. Chorionic gonadotropin has a recent origin within primates and an evolutionary history of selection. Mol Biol Evol. 19:320–335.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol. 3:418–426.
Nurminsky DI, Nurminskaya MV, De Aguiar D, Hartl DL. 1998. Selective sweep of a newly evolved sperm-specific gene in *Drosophila*. Nature. 396:572–575.
Ohno S. 1970. Evolution by gene duplication. Berlin (Germany): Springer.
Parisi M, Nuttall R, Edwards P, et al. (12 co-authors). 2004. A survey of ovary-, testis-, and soma-biased gene expression in *Drosophila melanogaster* adults. Genome Biol. 5:R40.
Parisi M, Nuttall R, Naiman D, Bouffard G, Malley J, Andrews J, Eastman S, Oliver B. 2003. Paucity of genes on the *Drosophila* X chromosome showing male-biased expression. Science. 299:697–700.
Ponce R, Hartl DL. 2006. The evolution of the novel *Sdic* gene cluster in *Drosophila melanogaster*. Gene. 376:174–183.
Pride DT, Blaser MJ. 2002. Concerted evolution between duplicated genetic elements in *Helicobacter pylori*. J Mol Biol. 316:629–642.
Ragab A, Thompson EC, Travers AA. 2005. HMGD and HMGZ interact genetically with the Brahma chromatin remodelling complex in *Drosophila*. Genetics. 172:1069–1078.
Raja SJ, Renkawitz-Pohl R. 2005. 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.
Rooney AP, Zhang J. 1999. Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? Mol Biol Evol. 16:706–710.
Ruksc A, Bell-Rogers PL, Smith JDL, Baker MD. 2008. Analysis of spontaneous gene conversion tracts within and between mammalian chromosomes. J Mol Biol. 377:337–351.
Samonte RV, Eichler EE. 2002. Segmental duplications and the evolution of the primate genome. Nat Rev Genet. 3:65–72.
Shiao M, Khil P, Camerini-Otero RD, Shiroishi T, Moriwaki K, Yu H, Long M. 2007. Origins of new male germ-line functions from X-derived autosomal retrogenes in the mouse. Mol Biol Evol. 24:2242–2253.
Snook RR, Markow TS. 2002. Efficiency of gamete usage in nature: sperm storage, fertilization and polyspermy. Proc R Soc Lond B BIol Sci. 269:467–473.
Steffen W, Linck RW. 1988. Evidence for tektins in centrioles and axonemal microtubules. Proc Natl Acad Sci USA. 85:2643–2647.
Swofford DL. 2003. PAUP*: phylogenetic analysis using parsimony (and other methods) 4.0. Sunderland (MA): Sinauer.
Thornton K, Long M. 2005. Excess of amino acid substitutions relative to polymorphism between X-linked duplications in *Drosophila melanogaster*. Mol Biol Evol. 22:273–284.
Travers AA. 2002. Priming the nucleosome: a role for HMGB proteins? EMBO Rep. 4:131–136.
van de Lagemaat LN, Landry J, Mager DL, Medstrand P. 2003. Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet. 19:530–536.
Vinckenbosch N, Dupanloup I, Kaessmann H. 2006. Evolutionary fate of retroposed gene copies in the human genome. Proc Natl Acad Sci USA. 103:3220–3225.
Wyckoff GJ, Wang W, Wu CI. 2000. Rapid evolution of male reproductive genes in the descent of man. Nature. 403:304–309.
Yang Z. 1996. Phylogenetic analysis using parsimony and likelihood methods. J Mol Evol. 42:294–307.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol. 15:568–573.
Yasuda GK, Schubiger G, Wakimoto BT. 1995. Genetic characterization of *ms(3)K81*, a paternal effect gene of *Drosophila melanogaster*. Genetics. 140:219–229.
Zhang JZ, Nielsen R, Yang ZH. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

Marta Wayne, Associate Editor

Accepted July 18, 2008