DOI: 10.1534/genetics.106.069245

Evidence for *de Novo* Evolution of Testis-Expressed Genes in the *Drosophila yakuba/Drosophila erecta* Clade

David J. Begun,*,[1] Heather A. Lindfors,* Andrew D. Kern* and Corbin D. Jones[†]

**Section of Evolution and Ecology, University of California, Davis, California 95616 and* [†]*Department of Biology and Genome Sciences Center, University of North Carolina, Chapel Hill, North Carolina 27599*

Manuscript received December 4, 2006
Accepted for publication April 11, 2007

ABSTRACT

The mutational origin and subsequent evolution of *de novo* genes, which are hypothesized to be genes of recent origin that are not obviously related to ancestral coding sequence, are poorly understood. However, accumulating evidence suggests that such genes may often function in male reproduction. Here we use testis-derived expressed sequence tags (ESTs) from *Drosophila yakuba* to identify genes that have likely arisen either in *D. yakuba* or in the *D. yakuba/D. erecta* ancestor. We found several such genes, which show testis-biased expression and are often *X*-linked. Comparative data indicate that three of these genes have very short open reading frames, which suggests the possibility that a significant number of testis-biased *de novo* genes in the *D. yakuba/D. erecta* clade may be noncoding RNA genes. These data, along with previously published data from *D. melanogaster*, support the idea that many *de novo* Drosophila genes function in male reproduction and that a small region of the *X* chromosome in the *melanogaster* subgroup may be a hotspot for the evolution of novel testis-biased genes.

THE availability of genome sequences, particularly those from closely related species, allows for systematic investigation into the origin and evolution of novel genes. Although difficult to rigorously define, here we use the term to refer to recently evolved (*i.e.*, species-specific or clade-specific) genes, which likely have major modifications of or wholesale departures from ancestral function. Well-annotated genomes provide the best substrate for identifying novel genes, as detailed enumeration of *bona fide* functional elements provides a starting point for genomic identification of related elements with recent origins. A long history of genomic investigation supports the importance of novelties deriving from duplication of pre-existing genes or parts thereof. For example, exon duplication, gene duplication (including via retrotransposition), and gene fusions contribute to new genes in many lineages (Ohno 1970; Li 1997), including Drosophila (Long and Langley 1993; Nurminsky *et al.* 1998; Betran *et al.* 2002; Long *et al.* 2003; Wang *et al.* 2004; Jones and Begun 2005; Loppin *et al.* 2005).

While the origin of new genes by duplication of pre-existing coding sequence is clearly established as an important component of genome evolution, the question of novel genetic functions that do not clearly derive from closely related genes has received less attention. A recent whole genome analysis of annotated *Drosophila melanogaster* genes was specifically designed to identify empirically validated genes that have no clearly homologous gene-related sequences in *D. melanogaster* or its close relatives (Levine *et al.* 2006). We refer to this class of orphan genes as "*de novo*," to suggest the possibility that they may derive from ancestrally noncoding sequence. Such genes would likely have novel functions that had recently evolved under directional selection in *D. melanogaster*. Levine *et al.* (2006) proposed that there are a minimum of five such genes in *D. melanogaster* and/or *D. simulans* that are probably absent from *D. yakuba*, *D. erecta*, and *D. anannasae*. These *D. melanogaster/D. simulans* putative *de novo* genes are strongly testis biased in expression, which supports the hypothesis that male reproductive functions are under particularly strong selection for novel functions. Interestingly, four of the five genes reported in Levine *et al.* (2006) are *X*-linked. One of these *X*-linked genes is located near previously identified novel, testis-expressed genes in *D. melanogaster*, which suggests the possibility that larger-scale, chromosomal phenomena contribute to the origination patterns of such genes (Levine *et al.* 2006).

Despite the obvious appeal of investigating novel genes in the *D. melanogaster* model system, a comprehensive view of the evolution of novelty requires investigation of other lineages. Comparative genomic investigation of novelty enables stronger inferences regarding evolutionary patterns and processes than can be gleaned from a strictly *D. melanogaster*-centric viewpoint. However, an investigative cost is incurred with increasing divergence from *D. melanogaster* because the advantage of its high quality genome sequence and annotation

[1]*Corresponding author:* Section of Evolution and Ecology, University of California, Davis, CA 95616. E-mail: djbegun@ucdavis.edu

is compromised by reduced quality of genome sequence alignments. In light of these considerations, the *melanogaster* subgroup of Drosophila is a nearly ideal system for the investigation of novelty. The subgroup contains a number of species of varying phylogenetic distance from each other and from *D. melanogaster*. Furthermore, sequence divergence between *D. melanogaster* and these other species is sufficiently low to allow high quality alignments over much of the genome.

The identification of putative *de novo* genes in species that are closely related to *D. melanogaster* requires empirical investigation into the gene complement of these other species. A comprehensive description of transcriptomes from other Drosophila species, which would greatly facilitate investigation of these issues, is currently unavailable. However, description of male reproductive tissue (*e.g.*, accessory gland and testis) transcriptomes may be a relatively efficient strategy for identifying putative *de novo* genes. For example, investigation of *D. yakuba* and *D. erecta* accessory gland cDNA libraries revealed a number of potential *de novo* genes in these species (Begun *et al.* 2006). It should be noted that these approaches bias one toward discovery of novel genes functioning in reproduction. Our goal here was to extend these earlier studies (Begun *et al.* 2006; Levine *et al.* 2006) by identifying potential *de novo* testis-biased genes in *D. yakuba* and/or *D. erecta* through analysis of a *D. yakuba* testis-derived cDNA library, which was sequenced as part of the *D. yakuba* genome project (http://www.dpgp.org and http://genome.wustl.edu).

MATERIALS AND METHODS

Construction of *D. yakuba* testis cDNA library and generation of ESTs: Testes from 100 *D. yakuba* males (line Tai18E2) were dissected in RNA-Later (Ambion). Total RNA was isolated using the Ambion mirVana miRNA kit and RNAsed (Ambion DNA-free kit). RACE-ready cDNA was synthesized from 2 μg of each prep [Invitrogen GeneRacer kit; the SSIII module and oligo(dT) primer were used for the RT step]. The resulting cDNA was amplified for five cycles using the Roche Expand High Fidelity PCR system. Amplified libraries were purified (QIAGEN QIAquick PCR purification kit), incubated in Promega Taq polymerase, and ligated into PCR4 TOPO vector (Invitrogen). The ligation was shipped to the Washington University Medical School Genome Sequencing Center (St. Louis), where clones were subjected to single-pass sequencing; the resulting EST data were then deposited into GenBank.

Identification of candidate novel genes from testis ESTs: We used procedures similar to those described in Wagstaff and Begun (2005), Begun *et al.* (2006), and Levine *et al.* (2006) to identity putative *de novo* genes. These procedures included an initial computational analysis of *D. yakuba* ESTs *vs.* genomic data from *D. yakuba*, *D. melanogaster*, *D. erecta*, and *D. ananassae* to identify a set of candidates. Genome assemblies of orthologous regions were then compared to generate microsyntenic alignments between *D. yakuba* and the other species. Candidate genes were characterized by RACE in *D. yakuba*. In many cases alignments, transcript data, and computational analysis of orthologous regions very strongly suggested that a given *D. yakuba* gene was absent from another species. In several cases, however, other species had highly diverged, yet apparently homologous sequence. For these cases, RT–PCR or reverse Northern analysis was used to investigate transcription in other species.

BLAST analysis of ESTs: We used BLAST analysis of *D. yakuba* genome assembly v2.0. (http://genome.wustl.edu) to identify the genomic region corresponding to each of 8772 ESTs. ESTs were also compared to SNAP (Korf 2004) gene predictions in *D. yakuba*. SNAP predictions that overlapped an EST were subsequently used in downstream analysis to evaluate regions of interest. Otherwise, ESTs alone were used in BLAST analyses. *D. yakuba* genomic regions corresponding to *D. yakuba* ESTs or EST/SNAP sequences were sequentially BLASTed *vs.* local BLAST databases of the *D. melanogaster* genome and to the NCBI trace archive for *D. ananassae* and *D. erecta*. We also compared these regions to all known transposable element sequences. Only the best BLAST hit from each database was retained for further analysis. For each *D. yakuba* region, we generated a summary score, which was the product of the *e*-values from non-*D. yakuba* BLAST comparisons (Levine *et al.* 2006). Regions with the highest values (*i.e.*, worst matches across all data sets) became our candidates for *D. yakuba*-specific *de novo* genes, which ultimately derive from the testis EST collection. Following the same basic protocol, we created an analogous list of *D. yakuba*/*D.erecta*-specific orphans by limiting the candidate list to those with genes present in both *D. yakuba* and *D.erecta*, but absent in the other genomic data sets.

Syntenic alignment and *de novo* status: BLAT comparisons of *D. yakuba* testis ESTs/cDNAs to the *D. yakuba* genome were used to identify *D. yakuba* genomic regions of variable size (generally several kilobases), which were then compared to the *D. melanogaster*, *D. erecta*, and *D. ananassae* genome assemblies (BLAT via UCSC Genome Browser (Kent *et al.* 2002, http://genome.ucsc.edu). Each putative orthologous gene region was then investigated in detail by pairwise alignments among species using the Martinez/Needleman-Wunsch algorithm as implemented in DNASTAR. For some *D. yakuba* gene regions, alignments to other species revealed evidence of homologous sequence in *D. melanogaster*, *D. erecta*, or *D. ananassae*, but no obvious evidence of an open reading frame (ORF) that was orthologous to the *D. yakuba* gene. In such cases, we computationally investigated the genomic sequence in the orthologous region for ORFs to determine protein-coding capacity in other (*i.e.*, non-*D. yakuba*) species and whether any predicted proteins associated with these ORFs showed sequence similarity or similar protein lengths relative to the candidate. For *D. yakuba* genes for which these analyses left the status of an orthologous ORF in doubt in another species, we used RT–PCR on RNA isolated from whole males and females of the appropriate species to investigate whether there was any evidence of transcription of the orthologous region (Table 1). These experiments were designed to detect highly diverged orthologous genes that would not typically be identified on the basis of sequence similarity. We also investigated whether putative *D. yakuba*/*D. erecta de novo* genes corresponded to *D. melanogaster* ESTs in the orthologous genomic region. We found no evidence of *D. melanogaster* ESTs (or genes) for any of the putative *de novo* genes presented in Table 1.

For some genomic regions, placement of RT–PCR primers for investigating transcription was problematic because of difficulty in identifying putative orthologous DNA. We investigated transcription for several of these regions using reverse Northern analysis, as previously described (Wagstaff and Begun 2005). Briefly, the region in question is amplified from the target species by PCR, run on an agarose gel, and blotted to a nylon membrane. The membrane was probed by P^{32}-labeled cDNA from adult males and females of the target species.

TABLE 1

Phylogenetic distribution and organization of genes

	Putative recently evolved genes in *D. yakuba* clade					
Gene (arm)	yak	teis	ere	mel	ana	No. of codons
zaa52b02y1 (*X*)	+	+	+[a]	−	−[c]	25, 27
zaa29a04x1 (*X*)	+	+[b]	−	−	−	224
zaa29a04x1-rel. (*X*)	+	+	−	−	−	194
zaa22f01y1 (*X*)	+	+	?	−	−	128
zaa93a11y1 (*2R*)	+	−	−	−	−	126
zaa57b07y1 (*3R*)	+	+	−[c]	−[c]	?	61 (23 for shared yak/teis CDS)
zaa20f10x1 (*4*)	+	+[b]	−[a]	−[c]	?	29
			Other genes			
zaa77f10y1 (*X*)	+	+	+[a]	+[a]	?	96
zaa36f12y1 (*X*)	+	+[b]	−[a]	+[a]	−[c]	79
zaa43d06x1 (*X*)	+	ND	+[a]	+[a]	−	104
zaa90g02y1 (*3L*)	+	+	−	+[a]	?	199

?, Identification of putative orthologous region is problematic; ND, no data.
[a] Investigated by RT–PCR.
[b] *D. teissieri* genomic sequence reveals an apparently orthologous region but no conserved ORF.
[c] Investigated by reverse Northern analysis.

Hybridization to the PCR product provides evidence of transcription. Unlike RT–PCR, this method has the advantage of requiring no *a priori* hypotheses regarding which bases in a genomic region are transcribed; a disadvantage is that it has low sensitivity compared to RT–PCR. Thus, conclusions regarding absence of transcription based on reverse Northerns may be less robust than conclusions from RT–PCR.

Characterization of transcripts and predicted proteins from candidate genes: Each candidate gene was subjected to 5′ and 3′ RACE on *D. yakuba* line Tai18E2, the genome sequence stock, followed by sequence analysis of RACE products. Sequence discrepancies (*e.g.*, indels or SNPs) between RACE products (or RT–PCR products, where applicable) and genomic sequence were resolved in favor of the *D. yakuba* genome sequence. We determined whether *D. yakuba* candidate genes were testis biased in expression by using RT–PCR on RNA isolated from dissected male testis, accessory gland and carcass, as well as RNA isolated from whole females, all from line Tai18E2. BLASTp was used to compare predicted proteins derived from RACE and EST analyses to protein databases.

Splicing of transcripts for non-*D. yakuba* species was inferred from sizes of genomic regions *vs.* RT–PCR products or by direct sequencing of RT–PCR products. In some instances, the longest ORF in *D. yakuba* corresponded to a *D. teissieri* sequence that did not share the putative *D. yakuba* initiation codon. In such cases we used the next in-frame, shared initiation codon as the putative start of the protein-coding region. Cases in which there was no shared initiation codon corresponding to a long ORF (see RESULTS) could be taken as support for the hypothesis that a gene produces a noncoding RNA.

All analysis and inference based on EST data, which were used to generate our list of candidates, were ultimately rechecked after the *D. yakuba/D. erecta* genes of interest were characterized by 5′ and 3′ RACE. Thus, any biases associated with analysis of ESTs did not affect the ultimate inference of *de novo*/orphan status. Gene names used throughout this article derive from the definition line of the GenBank entry of the associated EST, which can be recovered from GenBank. For example, gene 22f01y1 (Table 1) is named on the basis of the definition line for the EST in GenBank accession CV790171, which can be recovered from GenBank by searching on zaa22f01.y1. ESTs corresponding to the genes in Table 1 can be found this way, using the zaa prefix. Throughout this article we use abbreviated gene names. New sequences for this report can be found in GenBank accessions EF508208–EF508269 and EF525530–EF525535. The phylogenetic relationships of the *melanogaster* group species investigated here (Table 1) can be found at http://flybase.bio.indiana.edu/blast/.

Polymorphism and divergence: Isofemale or inbred lines of *D. yakuba* (obtained from P. Andolfatto, University of California, San Diego) and *D. teisierri* (obtained from M. Long, University of Chicago) were used for the population genetic or divergence analysis. Direct sequencing of PCR products was used for inbred lines. For each isofemale line, high-fidelity PCR was followed by cloning, colony PCR, and, finally, sequencing of a single insert. Summary statistics were calculated in DnaSP (ROZAS *et al.* 2003).

RESULTS

We used sequence analysis of *D. yakuba* testis ESTs, along with genome assemblies and trace data from *D. yakuba*, *D. erecta*, *D. melanogaster*, and *D. ananassae*, to identify 11 putative *D. yakuba/D. erecta* novel genes (Table 1), 1 of which, 29a04-rel., was identified by BLAT analysis as a potential paralog of a gene corresponding to *D. yakuba* testis EST zaa29a04.x1. The existence of this paralog was later validated by RT–PCR and RACE analysis. Six of the 11 genes overlap a *D. yakuba* GLEAN gene model (supplemental Table 1 at http://www.genetics.org/supplemental/). Further investigation of the 11 genes by RT–PCR (or investigation of *D. melanogaster* ESTs) provided evidence that 3, 90g02, 36f12, and 77f10, are transcribed and spliced in *D. melanogaster*, despite our initial skepticism (due to poor sequence alignments)

that they were present in this species. Two of these genes, 36f12 and 90g02, may be testis-expressed genes that were present in the ancestor of the *melanogaster* subgroup and subsequently lost from *D. erecta* (Table 1). RT–PCR analysis of a fourth gene, 43d06 in *D. melanogaster*, showed evidence of transcription in this species. However, sequence analysis of the RT–PCR product showed no evidence of splicing, as was observed for the putative ortholog in *D. yakuba* and *D. erecta*. Given that our RT–PCR experiments provided no evidence of genomic DNA contamination of our cDNAs, transcription of an unspliced message in *D. melanogaster* is a possible explanation; this supports recent evolution of intron–exon organization for this gene, which, for the purposes of this article, we do not consider an orphan/*de novo* gene. Sequences from RT–PCR products in *D. melanogaster* (and *D. erecta*) are provided in supplemental Table 2 (http://www.genetics.org/supplemental/). It is worth noting that RT–PCR experiments used to investigate the possibility of transcription of putative orphans in *D. melanogaster* or *D. erecta* were often based on alignments of homologous sequences that would not typically be interpreted as containing orthologous ORFs due to their high divergence. Indeed, sequence divergence between *D. yakuba* and *D. melanogaster* for these four genes is unusually high or cannot be estimated at all because alignment is fundamentally ambiguous. The evolution of such genes is an interesting topic that will primarily be addressed elsewhere.

These data left us with seven candidate, recently evolved genes. For some genes, the inference that they are recently evolved was weakened by suspect alignments of microsyntenic regions, which made the design of RT–PCR experiments problematic. In these cases, reverse Northern analysis (Wagstaff and Begun 2005) provided no evidence of transcription of the orthologous regions (Table 1 and supplemental Figure 1 at http://www.genetics.org/supplemental/), which supports (but does not prove) our inference of recent evolution in *D. yakuba*/*D. erecta*. One gene, 93a11, appears to be absent from the orthologous region of *D. teissieri* as determined by sequence data from multiple *D. teissieri* isofemale lines, which suggests a recent origin in *D. yakuba*, subsequent to *D. yakuba*/*D. teissieri* speciation. None of the predicted proteins from the 11 genes that were investigated corresponds to known proteins or harbors known functional domains as determined by BLASTp analysis (Marchler-Bauer *et al.* 2003).

Gene organization and natural variation: The *D. yakuba* genes described above are predicted to code for small proteins (Table 1). All transcripts are spliced (Figure 1), have canonical splice junctions, and are polyadenylated, which strongly supports the proposition that the genes are real rather than the result of experimental artifacts. Five of the genes have at least one untranslated exon. One gene has two 5′ untranslated exons; another has two 3′ untranslated exons. The putative ORFs for genes 52b02, 57b07, and 20f10 are <90 bp long. Thus, they either code for small peptides or do not produce proteins (Tupy *et al.* 2005). Support for the possibility that some of the genes described here are not protein coding comes from the fact that for 52b02 and 20f10, alignment of the orthologous region from *D. tesisieri* and *D. yakuba* suggests that *D. teissieri* does not have an orthologous start codon. The idea that several novel reproduction-related genes in the *D. yakuba*/*D. erecta* lineage may be RNA genes (Begun *et al.* 2006) is worth further consideration, but will require much additional work. Sequences corresponding to transcripts of the genes in Table 1 can be found in supplemental Table 1 (http://www.genetics.org/supplemental/).

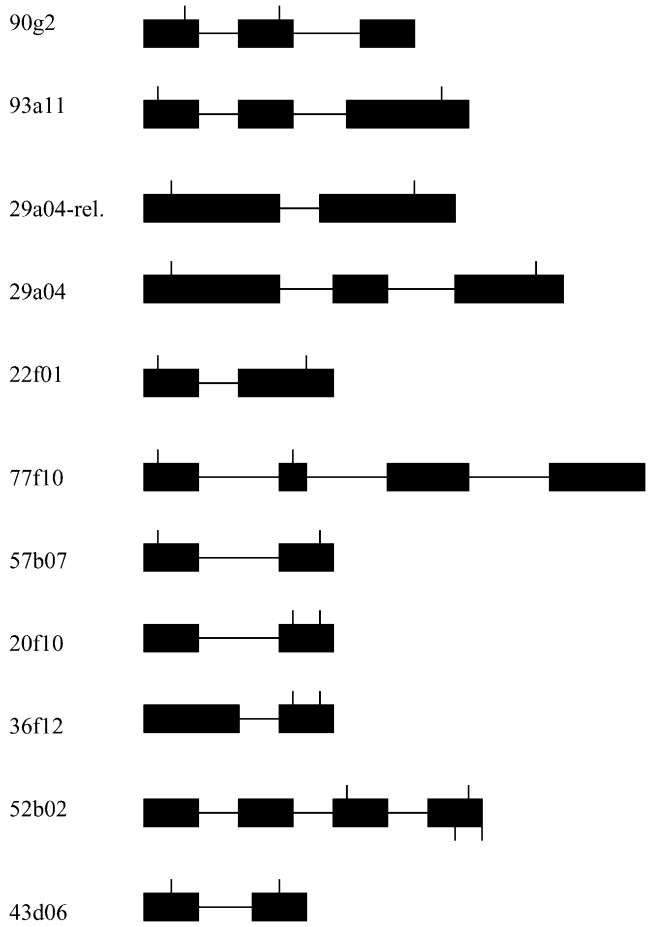

Figure 1.—Gene organization. Exons and introns are indicated by black boxes and horizontal lines, respectively. Putative start and stop codons are indicated by vertical tick marks on exons. Two potential start and stop codons are indicated for gene 52b02.

Levels of heterozygosity and divergence in putative coding regions were estimated for several genes (Table 2). As expected on the basis of analysis of recently evolved genes (Long *et al.* 2003; Jones and Begun 2005) and genes that are enriched for testis expression (Pröschel *et al.* 2006), the d_N/d_S ratios for these genes (Table 2) are high for Drosophila. For example, 90g02, which has 199 codons, has three ~25-codon windows with an average $d_N/d_S > 4$. Gene 90g02 is also the only gene (of three)

TABLE 2

Polymorphism and divergence between *D. yakuba* and *D. teissieri* for a sample of putative novel or rapidly evolving protein-coding genes

	D. yakuba π		*D. teissieri* π			
Gene	Silent	Replacement	Silent	Replacement	d_N	d_S
29a04-rel.	0.008	0.011 (7)	0.004	0.013 (6)	0.204	0.172
22f01	0.000	0.001 (7)	0.021	0.008 (7)	0.053	0.059
57b07					0.103	0.062
90g02	0.021	0.006 (7)	0.023	0.016 (8)	0.265	0.181
77f10	0.011	0.006 (6)	0.032	0.015 (2)	0.098	0.168

d_N and d_S were estimated using Jukes–Cantor correction in DnaSP (v.4.0). The number of alleles sampled is in parentheses.

on which we could carry out a McDonald–Kreitman test (McDonald and Kreitman 1991) that showed a significant deviation from the neutral model (fixed synonymous = 20, polymorphic synonymous = 16, fixed nonsynonymous = 89, polymorphic nonsynonymous = 26, *G*-test, $P = 0.01$), supporting adaptive protein divergence. Remarkably, syntenic alignments strongly suggest that this gene was recently lost in the *D. erecta* lineage (Table 1), in spite of the fact that it has a history of directional selection in the sibling species. This finding, which mirrors results from *melanogaster* subgroup *Acps* (Begun and Lindfors 2005), supports the idea that functional roles of reproduction-related genes and the modes of selection impinging on them may change over relatively short time scales.

Gene expression: The putative novel genes discussed here were discovered by investigation of the *D. yakuba* testis EST collection, which suggests that they are likely to be testis biased in expression (Parisi *et al.* 2003). Our RT–PCR experiments (Figure 2) show that they are in fact strongly testis biased in expression, which supports the idea that these genes function in male reproduction. Only putative orphan 20f10, which is located on the dot chromosome, showed evidence of expression in males and females (data not shown).

Chromosomal location: Four of the seven putative *de novo* genes are on the *X* chromosome. Under the hypothesis that these genes should appear as *X*-linked *vs.* autosomal in proportions based on the size of the *D. melanogaster X vs.* autosomes, we expect ~20% of *de novo* genes to be *X*-linked. The binomial probability of observing four or more *X*-linked *de novo* genes is 0.03. However, if the putative duplicated *de novo* genes zaa29a04x1 and zaa29a04x1-related derive from a single origination event, there are six originations, three of which are *X*-linked. This interpretation provides no statistical support for *X*-linked enrichment in the *D. yakuba/D. erecta* clade.

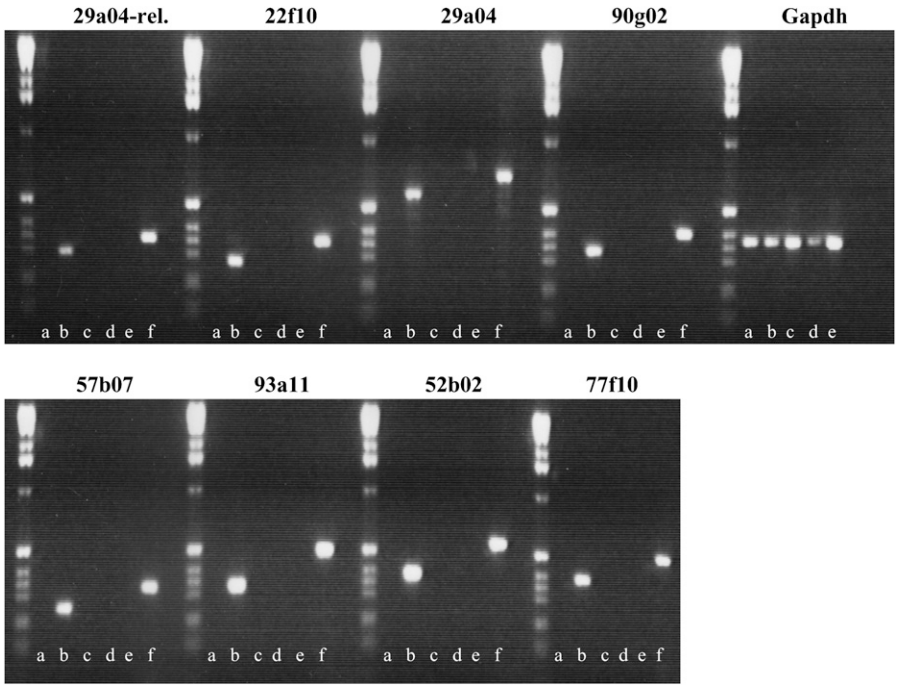

Figure 2.—RT–PCR results from various tissues of *D. yakuba* line Tai18E2. a, accessory gland; b, testis; c, reproductive tract remainder; d, male carcass; e, whole female; f, genomic DNA. Ladder, 1 kb ladder.

The duplicates zaa29a04x1 and zaa29a04x1-related reside ~5 kb apart at the base of the *X* chromosome (*D. yakuba* assembly version 2.0). Remarkably, the *D. melanogaster*/*D. simulans* novel genes, *Sdic* (Nurminsky *et al.* 1998), *CG15323* (Levine *et al.* 2006), and *hydra* (H.-P. Yang, personal communication) are also testis biased in expression and located in the same genomic region. Thus, the ~465-kb *D. melanogaster* region spanned by *Sdic* and *hydra*, which is not rearranged between *D. melanogaster* and *D. yakuba* (J. Ranz, personal communication), contains five novel testis-biased genes in two lineages. These data strongly suggest that a conserved property of this chromosomal region supports increased origination rates of novel testis-specific genes in Drosophila. The fact that the putative *D. yakuba* testis-biased orphan, 93a11, is only ~7.5 kb from the *D. yakuba*/*D. erecta*-specific accessory gland protein gene, *Acp223* (Begun *et al.* 2006), further supports the notion that physical locations of male-biased novel genes are nonrandom.

DISCUSSION

Investigation of the male reproduction related transcriptome in the *melanogaster* subgroup of Drosophila is an effective strategy for identifying candidate *de novo* genes. We recently reported on putative *de novo* accessory gland protein genes from *D. yakuba* (Begun *et al.* 2006). The goal of the work presented here was to generate a set of candidate *de novo* testis-expressed genes in the *D. yakuba* lineage for further evolutionary and functional investigation.

D. melanogaster novel, testis-biased genes were significantly enriched for *X*-linkage (Levine *et al.* 2006). The same general pattern was observed in the *D. yakuba* lineage, although it was not significant. Nevertheless, appearance of the same chromosomal pattern in two Drosophila lineages (in which 7 of 11 proposed originations are *X*-linked, when only 2 are expected) suggests the possibility that it is a manifestation of a fundamental process relating to the origin and/or fixation of such genes. The trend toward *X*-linkage in putative *de novo* genes is an interesting counterpoint to the origin of gene duplications by retrotransposition in Drosophila (Betran *et al.* 2002) and to the observation that genes showing male-biased expression are under-represented on the *D. melanogaster X* chromosome (Parisi *et al.* 2003).

Retrotransposed genes are more likely to have parental copies on the *X* and descendant, autosomal copies with testis-specific expression (Betran *et al.* 2002; Betran and Long 2003; Bai *et al.* 2007), suggesting that novel, testis-expressed genes should tend to be autosomal. One possible explanation for the difference in chromosomal distribution for retrotransposed genes *vs.* the genes described here is that the genes described here are generally younger than the genes identified by Betran *et al.* (2002) and Bai *et al.* (2007). Alternatively, the mutation process underlying duplication by retrotransposition could be biased toward *X* ancestry and autosomal descent, while the mutation process for the genes described here could simply be biased toward the *X* chromosome. An interesting, if speculative, possibility for the preponderance of *X*-linkage among putative *de novo* testis-biased Drosophila genes is hypertranscription of the *X* chromosome in the male germline due to dosage compensation (Gupta *et al.* 2006). This model proposes that transcription of noncoding regions in the male germline would be more likely to occur for the *X* chromosome because this chromosome would be hypertranscribed relative to the autosomes. Hypertranscription of genes could be associated with spurious transcription through regional effects (*e.g.*, transcriptional domains) or through more local effects, such as readthrough transcription.

The issue of *X*-linkage among *de novo* genes is related to a remarkable feature of the data presented here and in previous work on novel *D. melanogaster* genes (Levine *et al.* 2006), namely, that the proximal region of the *X* appears to be a hotspot for fixation of novel testis-expressed genes. Two novel *D. yakuba* genes and three novel *D. melanogaster* genes are contained within a 500-kb region of the proximal *X* chromosome. This region may correspond to a testis expression domain in *D. melanogaster* (Boutanaev *et al.* 2002), which could lead to an increased origination rate of male reproduction-related genetic novelties by cooptation of noncoding DNA into coding function (Begun *et al.* 2006; Levine *et al.* 2006). This idea, if true, suggests that the extent and distribution of such chromosomal expression domains across lineages could have a substantial impact on the evolution of novelty in different species.

The path toward an understanding of the biological basis of selection for novelty in Drosophila male reproduction lies in revealing the functional biology of the genes reported here and elsewhere (*e.g.*, Long and Langley 1993, Betran *et al.* 2002, Nurminsky *et al.* 1998, Levine *et al.* 2006, Begun *et al.* 2006), through reverse genetic, cell biological, and biochemical analysis. Moreover, the existence of independently evolved, novel genes in closely related species provides an opportunity for comparative investigation of novel genes functioning in male reproduction. Such research could reveal whether similar biological processes are under selection for novelty in *D. melanogaster* and *D. yakuba* or, instead, whether each lineage has a unique constellation of selection pressures driving the evolution of novel male reproductive function.

We thank Melissa Eckert and Mia Levine for technical assistance and two anonymous reviewers and Larry Harshman for useful comments. This work was supported by NSF grant DEB-0327049 and NIH grant GM071926.

LITERATURE CITED

Bai, Y., C. Casola, C. Feschotte and E. Betran, 2007 Comparative genomics reveals a constant rate of origination and convergent

acquisition of functional retrogenes in Drosophila. Genome Biol. **8**(1): R11.

Begun, D. J., and H. A. Lindfors, 2005 Rapid evolution of genomic Acp complement in the melanogaster subgroup of Drosophila. Mol. Biol. Evol. **22:** 2010–2021.

Begun, D. J., H. A. Lindfors, M. E. Thompson and A. K. Holloway, 2006 Recently evolved genes identified from *Drosophila yakuba* and *D. erecta* accessory gland expressed sequence tags. Genetics **172:** 1675–1681.

Betran, E., and M. Long, 2003 Dntf-2r, a young Drosophila retroposed gene with male specific expression. Genetics **164:** 977–988.

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

Boutanaev, A. M., A. I. Kalmykova, Y. Y. Shevelyov and D. I. Nurminsky, 2002 Large clusters of co-expressed genes in the Drosophila genome. Nature **420**(6916): 666–669.

Gupta, V., M. Parisi, D. Sturgill, R. Nutall, M. Doctolero *et al.*, 2006 Global analysis of X-chromosome dosage compensation. J. Biol. **5**(1): 3.

Jones, C. D., and D. J. Begun, 2005 Parallel evolution of chimeric fusion genes. Proc. Natl. Acad. Sci. USA **102:** 11373–11378.

Kent, W. J., C. W. Sugnet, T. S. Furey, K. M. Roskin, T. H. Pringle *et al.*, 2002 The Human Genome Browser at UCSC. Genome Res. **12**(6): 996–1006.

Korf, I., 2004 Gene finding in novel genomes. BMC Bioinformatics **5:** 59.

Levine, M. T., C. D. Jones, A. D. Kern, H. A. Lindfors and D. J. Begun, 2006 Novel genes derived from non-coding DNA in *Drosophila melanogaster* are frequently X-linked and show testis-biased expression. Proc. Natl. Acad. Sci USA **103:** 9935–9939.

Li, W.-H., 1997 *Molecular Evolution*. Sinauer Associates, Sunderland, MA.

Long, M., and C. H. Langley, 1993 Natural selection and the origin of *jingwei*, a chimeric processed functional gene in Drosophila. Science **260:** 91–95.

Long, M., M. Deutsch, W. Wang, E. Betran, F. G. Brunet *et al.*, 2003 Origin of new genes: evidence from experimental and computational analysis. Genetica **118:** 171–182.

Loppin, B, D. Lepetit, S. Dorus, P. Couble and T. Karr, 2005 Origin and neofunctionalization of a Drosophila paternal effect gene essential for zygote viability. Curr. Biol. **15:** 87–93.

Marchler-Bauer, A., J. B. Anderson, C. DeWeese-Scott, N. D. Fedorova, L. Y. Geer *et al.*, 2003 CDD: a curated Entrez database of conserved domain alignments. Nucleic Acids Res. **31:** 383–387.

McDonald, J., and M. Kreitman, 1991 Adaptive protein evolution at the Adh locus in Drosophila. Nature **351:** 652–654.

Nurminsky, D. I., M. V. Nurminskaya, D. DeAguiar and D. L. Hartl, 1998 Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature **396:** 572–575.

Ohno, S., 1970 *Evolution by Gene Duplication*. Springer-Verlag, Berlin.

Parisi, M., R. Nuttall, D. Naiman, G. Bouffard, J. Malley *et al.*, 2003 Paucity of genes on the Drosophila X chromosome showing male-biased expression. Science **299:** 697–700.

Prŏschel, M., Z. Zhang and J. Parsch, 2006 Widespread adaptive evolution of Drosophila genes with sex-biased expression. Genetics **174:** 893–900.

Rozas, J., J. C. Sanchez-DelBarrio, X. Messegyer and R. Rozas, 2003 DnaSP, DNA polymorphism analysis by the coalescent and other methods. Bioinformatics **19:** 2496–2497.

Tupy, J. L., A. M. Bailey, G. Dailey, M. Evans-Holm, C. W. Siebel *et al.*, 2005 Identification of putative noncoding polyadenylated transcripts in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **102:** 5495–5500.

Wagstaff, B. J., and D. J. Begun, 2005 Comparative genomics of accessory gland protein genes in *Drosophila melanogaster* and *D. pseudoobscura*. Mol. Biol. Evol. **22:** 818–832.

Wang, W., H. Yu and M. Long, 2004 Duplication-degeneration as a mechanism of gene fission and the origin of new genes in Drosophila species. Nat. Genet. **36:** 523–527.

Communicating editor: L. Harshman