Genomes gain and lose genes at surprisingly high rates in both unicellular (Jain et al. 1999; Ochman and Jones 2000; Lynch and Conery 2003) and multicellular organisms (Patthy 1999; Betran et al. 2002; Harrison et al. 2002; Borevitz et al. 2003; Holland 2003; Tian et al. 2003). Identifying the mutational and population genetic mechanisms involved in gene loss and gain is critical to understanding the forces shaping genome variation. The spread of gene duplications, by either drift or natural selection (Wagner 2001; Ohta 2003; reviewed in Wolfe and Li 2003), is one mechanism by which gene number increases (Haldane 1932; Ohno 1970). Gene duplications may, in many cases, evolve new functions or become subfunctionalized simply as a result of amino acid or expression evolution and not as a consequence of large-scale changes in gene organization (Ohno 1970; Lynch and Conery 2000; Lynch et al. 2001; Hughes 2002; Betran and Long 2003; Katju and Lynch 2003).

Occasionally, duplication events lead to radical reorganization of gene structures that likely lead to dramatic and immediate functional divergence. One type of radical reorganization is gene fusion, whereby two previously separate and independent genes are fused to form a single contiguous gene. Such chimeric fusion genes (CFGs) have been identified in several taxa. For example, in plants CFGs are implicated in cytoplasmic rates in both unicellular (Jain et al. 1999; Ochman and Jones 2000; Lynch and Conery 2003) and multicellular organisms (Patthy 1999; Betran et al. 2002; Harrison et al. 2002; Borevitz et al. 2003; Holland 2003; Tian et al. 2003). Identifying the mutational and population genetic mechanisms involved in gene loss and gain is critical to understanding the forces shaping genome variation. The spread of gene duplications, by either drift or natural selection (Wagner 2001; Ohta 2003; reviewed in Wolfe and Li 2003), is one mechanism by which gene number increases (Haldane 1932; Ohno 1970). Gene duplications may, in many cases, evolve new functions or become subfunctionalized simply as a result of amino acid or expression evolution and not as a consequence of large-scale changes in gene organization (Ohno 1970; Lynch and Conery 2000; Lynch et al. 2001; Hughes 2002; Betran and Long 2003; Katju and Lynch 2003). Occasionally, duplication events lead to radical reorganization of gene structures that likely lead to dramatic and immediate functional divergence. One type of radical reorganization is gene fusion, whereby two previously separate and independent genes are fused to form a single contiguous gene. Such chimeric fusion genes (CFGs) have been identified in several taxa. For example, in plants CFGs are implicated in cytoplasmic male sterility (He et al. 1996). A few CFGs have also been found in vertebrates (Finta and Zaphiropoulos 2000; Rogalla et al. 2000; Thomson et al. 2000; Courseaux and Nahon 2001). Finally, several novel CFGs have been described in Drosophila, such as jingwei in Drosophila tessieri and D. yakuba (Long and Langley 1993), Silic in D. melanogaster (Nurminsky et al. 1998), and Adh-finnegan (Sullivan et al. 1994; Begun 1997).

Two novel Drosophila genes, jingwei and Adh-finnegan, were previously thought to be Alcohol-dehydrogenase (Adh) pseudogenes (Fischer and Maniatis 1985; Jeffs and Ashburner 1991). In both cases, further analysis showed that these genes were functional genes that acquired protein-coding sequence 5’ of the Adh-derived region of gene (Long and Langley 1993; Begun 1997). jingwei is a fusion of the amino terminus of a gene known as yellow emperor and a retrotransposed Adh (Wang et al. 2000). The jingwei expression profile has diverged from its Adh ancestors. In D. tessieri expression is now limited to the testes (like ymp), although in D. yakuba it is expressed in other tissues as well (Long and Langley 1993). Adh-finnegan was created by the chromosomal duplication of Adh combined with the recruitment of a new 5’ exon of unknown origin (Begun 1997). Adh-finnegan appears to be expressed broadly in adult tissues (Sullivan et al. 1994). Although these two Adh-derived fusion genes arose via different mechanisms and show dramatically different expression patterns, directional selection appears to have driven rapid amino acid evolution in both genes (Long and Langley 1993; Begun 1997). The fact that two novel Drosophila genes are...
derived from Adh and share some common aspects of their evolution raises two important questions:

1. Is Adh overrepresented among novel fly genes? Or, does the discovery of jingwei and Adh-finnegan—both of which were discovered unintentionally—reflect the intense study of Adh in Drosophila?

2. If Adh frequently participates in radical reorganizations associated with novel function, what general principles of the evolution of novel function are revealed by these examples of repeated evolution?

Given the history of the discovery of jingwei and Adh-finnegan, a report of a third putative Adh pseudogene in the obscura group of Drosophila (Marfany and Gonzalez-Duarte 1992; Luque et al. 1997) attracted our attention. DNA sequencing showed that this putative pseudogene originated by retrotransposition. Results from polytene in situ hybridization showed that this Adh retrosequence had transposed to chromosome arm E from chromosome U, which is the expected location of Adh on the basis of the conservation of Muller elements (Ashburner 1989). This retrotransposed Adh was found in D. subobscura, D. guanche, and D. madeirensis, but not in D. ambigua (Visa et al. 1991; Marfany and Gonzalez-Duarte 1992; Luque et al. 1997), suggesting that the gene likely arose within the past 3 million years. Luque et al. (1997) sequenced six clones harboring this putative Adh retrosequence, two each from genomic libraries of D. subobscura, D. guanche, and D. madeirensis. Comparisons of putative Adh pseudogenes to Adh for these species revealed frameshift mutations or indels in 5′ and 3′ regions flanking the Adh coding regions. The codon homologous to the ATG initiation codon of the ancestral Adh was CTG in both D. guanche clones. Premature stop codons were evident in one of the two D. guanche clones and one of the two D. subobscura clones, but none of the D. madeirensis clones. These observations suggested that at least some of these Adh sequences were no longer functional. The fact that the duplicate Adh was a retrosequence was also interpreted as support for the pseudogene hypothesis, as retrotransposed sequences potentially lack regulatory elements necessary for proper expression. All three species showed elevated amino acid substitution rates (dS) relative to Adh. None of the putative retropseudogenes, however, showed a nonsynonymous to synonymous substitution rate (dN/dS) close to one, the expectation for a neutrally evolving pseudogene. Moreover, codon bias increased in the putative Adh retropseudogenes, an unexpected result for a nonfunctional gene. Overall, the data presented a conflicting picture of the Adh retrosequence. Some aspects of the data supported the pseudogene hypothesis, yet others were strangely inconsistent with the hypothesis and were similar to the situation previously observed in the repleta group Adh duplication (Sullivan et al. 1994; Begun 1997). We present here our analysis of this retrotransposed Adh. We show that this putative Adh retropseudogene is actually part of a new chimeric fusion gene that is the result of an Adh mRNA inserting into the Gapdh-like gene, CG9010. This fusion gene is actively and widely transcribed. While the 5′ promoters and transcription patterns of this gene are similar to those of CG9010, the protein-coding region has diverged for both the CG9010 and the Adh-like regions. Population genetic and phylogenetic analyses suggest that this amino acid evolution resulted from directional selection shortly after the chimeric fusion gene was formed.

**Materials and Methods**

**Stocks:** D. subobscura stocks were obtained from the Species Stock Center and from A. Davis. D. guanche, D. hydei, D. pseudoobscura, D. melanogaster, and D. yakuba stocks were originally obtained from the Species Stock Center and the Bloomington Stock Center. A D. madeirensis stock was kindly provided by M. Aguade. All stocks were reared on standard Drosophila medium at room temperature.

**DNA Sequencing:** PCR products were sequenced directly using an ABI 377 automated sequencer and BigDye Terminator chemistry (Applied BioSystems, Foster City, CA).

**RNA Extraction, cDNA Preparation, and RT-PCR:** Poly(A+) RNA was prepared from whole flies or larvae using a MicroPoly(A) kit (Ambion, Austin, TX). DNA was cloned for reverse transcriptase-PCR and rapid amplification of cDNA ends (RACE) was performed from this RNA using the SMART RACE cDNA amplification kit (CLONTECH, Palo Alto, CA). SuperScript II reverse transcriptase (GIBCO BRL, Rockville, MD) was used for all RT reactions. Gene-specific primers were used to assay gene expression by RT-PCR on cDNA isolated from larvae (first, second, and third instar), whole adult males, and whole adult females.

**Genomic Library Construction and Screening:** D. subobscura genomic DNA was isolated from adult flies, partially digested with Sau3A1 (New England BioLabs, Beverly, MA), and then phosphorylated with CIAP (Promega, Madison, WI). These fragments were ligated into the Lambda DASH II vector according to the manufacturer’s instructions (Stratagene, La Jolla, CA; T4 ligase was from GIBCO BRL), followed by packaging using Gigapack III Gold packaging reactions (Stratagene, La Jolla, CA). The library was amplified once on plates using XL-1Blue MRA (P2) cells.

Primary and secondary plaque lifts were carried out on Nytan Nylon membranes (Schleicher & Schuell, Keene, NH). The library was screened with a 1-kb CG9010 probe that was PCR amplified from D. melanogaster. Because this probe cross-hybridizes to plaques harboring Gapdh, we used PCR with primers designed for D. subobscura Gapdh to rule out false positives. Phage containing CG9010 were digested with EcoRI and PstI, resolved on a 0.7% agarose gel, Southern blotted, and probed with CG9010. The fragment containing D. subobscura CG9010 was subcloned into pBluescript.

**Genomic Southern Blot Analysis of CG9010:** Southern blot analysis was used to infer copy number of CG9010 in D. subobscura. Genomic DNA (5 μg) was purified from D. melanogaster, D. pseudoobscura, D. guanche, and D. subobscura. These samples were digested with PstI or HindIII (GIBCO BRL), electrophoresed on a 0.7% gel, and Southern blotted to Nytan nylon membranes. These blots were then probed with PCR-amplified 1-kb fragments of D. melanogaster CG9010 and then D. subobscura CG9010.

**Protein Analyses:** One gram of tissue (whole adults and
larvae) was homogenized in 2 ml ice-cold homogenization buffer and then centrifuged. Protein concentration in the supernatant was determined using the Bradford method (Bio-Rad, Hercules, CA). We then applied SDS-PAGE to our samples (10% acrylamide resolving gels, 4% acrylamide stacking gels). Typically, 5 μg of sample was boiled and then loaded in each lane. Gels were run at 70 mA constant current for 30–40 min. Gels were electrophoretically stained with nitrocellulose at 0° for 30–60 min at 100 V constant voltage following by blocking for 1 hr in a TBS-milk solution. Blots were incubated overnight (at 4°) with goat anti-β, melanogaster ADH (courtesy of C. Benyanjati) diluted in TBS-milk solution. Blots were placed in fresh TBS-milk and incubated with the secondary antibody (anti-goat HRP) for 1 hr followed by washing with TBS-Tween. The secondary antibody was visualized with ECL Plus (Amer- sham Biosciences, Piscataway, NJ) followed by autoradiography.

This approach repeatedly worked well for ADH proteins in all species we assayed (D. melanogaster, D. subobscura, D. yakuba, D. pseudoobscura, and D. hydei). Overexposure of the film to the blot would visualize a number of minor bands, but it was impossible to determine which, if any, would correspond to the band of interest.

We also used aldehyde gels to look for residual ADH activity in the D. subobscura CGF. We adapted the protocol of Batterham et al. (1983). Although we tried a variety of conditions, we observed only a single band of activity. This was consistent with what was reported in the literature (see results).

DNA sequence analysis: BLASTN and TBLASTX were used to identify similar sequences from the NCBI databases (Altschul et al. 1997). DNASTAR (DNASTAR, Madison, WI) was used for sequence alignments, contig assembly, and restriction mapping. Accession numbers of previously published data used in this analysis are X55390, X55391, M55545, X60112, U68470, U68469, X60113, U68472, U68471, AF175211, and AE003805.

Basic population genetic analyses were done using either DNAsp (Roza and Roza 1999) or software written by C.D.J. We limited our analysis to regions of high sequence quality. Typically, insertion/deletion polymorphism was ignored in our calculations of population genetics statistics.

Promoter prediction was accomplished using NNPP (Reese 2001) and McPromoter (Ohler et al. 2002). As noted in results, a threshold of 0.8 was used (which is predicted to give a false positive rate of 0.4% for NNPP). Signal peptide prediction used SignalP (Nielsen et al. 1997). No signal peptides were detected.

Phylogenetic analysis: PAML provides a suite of maximum likelihood-based tools for combining DNA sequence and phylogenetic data to test molecular evolutionary hypotheses (Yang 1997; Yang and Bielawski 2000). We used the phylogeny of Ramos-Ornins et al. (1998). There are three major steps to using PAML: (1) choice of appropriate model, (2) parameterization of that model, and (3) sequential comparison using log-likelihood ratio tests of simpler to more complex models to evaluate if a more complex model provides a significantly better fit to the data. For clarity, steps 1 and 2 will be described here and step 3 will be presented in results.

Evolution of protein-coding regions of CG9010 and Adh-derived sequences were analyzed independently using the codon model (Codeml; Goldman and Yang 1994; Yang 1997). In the following sections, the difference in the log likelihood (Δlnl), for the relevant degrees of freedom, implied a P-value <0.05 and was typically <0.001. Unless noted otherwise, model comparisons involving multiple tests remained significant after Bonferroni corrections. F3X4 codon model fit the CG9010 data the best of the codon models; the estimated codon table fit the Adh data the best. When appropriate and when a significantly better fit to the data was produced, ω, ωs, and α were estimated (see Yang 1997). For the analyses discussed in results, we a priori hypothesize that the lineage created by the formation of the CFG (Adh-Twain) will be undergoing more rapid evolution than the Adh or CG9010 lineages [e.g., hypothesis generation is independent of the data used to test it; see Yang (1997) p. 23]. In several cases, convergence to maximum likelihood estimates was verified by changing the “small difference” parameter (see Yang 1997, p. 19). Reconstruction of ancestral sequences was done using both joint and marginal reconstruction. Ancestral states were identical regardless of method. Note that all amino acid positions are in terms of their position in D. subobscura, not D. melanogaster.

RESULTS

Open reading frame extends 5’ of Adh retrosequence and is homologous to CG9010: Analysis of previously published D. subobscura genomic sequences revealed an open reading frame (ORF) 5’ of the Adh retrosequence. This ORF was contiguous and in the same reading frame as the Adh retrosequence ORF. TBLASTX of the 5’ ORF to the D. melanogaster reference sequence showed that the predicted amino acid sequence from this region was homologous to the translated D. melanogaster protein for CG9010, a broadly expressed, intronless gene that is similar to Gapdh. This region, roughly 300 bases 5’ of the Adh sequence, is homologous to CG9010 in all three species. In D. subobscura, the inferred amino acid sequence of the region from 135 to 441 bases 5’ of the putative retrosequence start codon shows ~60% identity with the amino terminus of the D. melanogaster predicted protein CG9010 (TBLASTX E-value = e-13). A similar region was found in D. madeirensis. The more distantly related D. guanche also contains a homologous region showing ~67% identity with D. melanogaster CG9010 (TBLASTX E-value = e-19).

As mentioned in Luque et al. (1997), the region immediately upstream of the putative retrosequence initiation codon and downstream of the CG9010-like region is similar to obscura group Adh 5’-UTR. It is not clear from sequence comparisons whether this region represents the larval or adult leader variant. The Adh-like region of the retrosequence retained the Adh 3’-UTR and a viable polyadenylation signal. The region 3’ of the Adh-like region 3’-UTR showed no significant similarity to any known gene, although it clearly harbors GEM elements (Vivas et al. 1999). Given the location of CG9010 in D. melanogaster, the D. subobscura CG9010 would be located on chromosome arm E, which is also the location of the Adh retrosequence.

These results suggest that the Adh pseudogene described by Luque et al. (1997) is a chimeric fusion gene that resulted from the retrotransposition of an Adh transcript into a D. subobscura CG9010. Our analysis, however, did not resolve the issue of the frameshift and nonsense mutations associated with at least some copies of the Adh retrosequence. To address this issue, we se-
quenced DNA encompassing 97% of the potential open reading frame of the chimeric fusion gene and some 5’-flanking sequence from 15 lines of *D. subobscura*, one line of *D. guanche*, and one line of *D. madeirensis*. Complete coding sequences were obtained for one *D. subobscura* line and one *D. guanche* line. All sequences had a contiguous open reading frame that included both the CG9010-like region and the Adh-like region. None of the frameshift insertion/deletions observed by Luque *et al.* were observed in our data. Nor were any premature stop codons found in any of the regions surveyed. This suggests that these indels and stop codons may have been sequencing errors in the original article. (If these stop codons and frameshifts do indeed exist, our data suggest that they are rare variants in *D. subobscura*.)

The substitution of the canonical Adh start codon (ATG) by a leucine (CTG) reported by Luque *et al.* (1997) was also observed in our *D. guanche* sequence. From these data, we conclude that this Adh “pseudogene” is likely a novel Adh-derived fusion gene, which we have tentatively named “Adh-Twain.” (Mark Twain’s famous statement, “The rumors of my demise have been greatly exaggerated,” was the inspiration for the name of this gene. We propose the abbreviation AdhT for Adh-Twain.)

**CG9010 exists in *D. subobscura* and *D. guanche***: If CG9010 function was necessary for the ancestor of *D. subobscura*, *D. guanche*, and *D. madeirensis* (which is likely given its similarity to Gapdh and its conservation across Drosophilida), then CG9010 must have duplicated in this lineage. In other words, if insertion of the retrosequence into CG9010 abolished CG9010 function, then an alternative functional copy of CG9010 should exist in the subobscura clade. We used Southern analysis to determine if CG9010 is present in more than one copy in *D. subobscura*. Digestion of genomic DNA with several restriction enzymes followed by blotting and hybridization with the 5’ end of CG9010 showed that CG9010 is single copy in *D. melanogaster* and *D. pseudoobscura*. The CG9010 probe, however, consistently produced two bands in *D. subobscura* (data not shown), supporting the idea that CG9010 duplicated in this lineage. BLAST searches of *D. pseudoobscura* suggest the presence of only one copy CG9010 in the strain used for the genome sequence.

To confirm that our Southern detected a surviving full copy of CG9010, we cloned CG9010. We used several methods to obtain the DNA sequence for the CG9010 homolog in *D. subobscura* and *D. guanche* (see Materials and Methods). The amino acid sequence of the *D. subobscura* homolog of CG9010 shares 91% amino acid identity with *D. pseudoobscura* and 82% amino acid identity with *D. melanogaster*. ESTs from *D. melanogaster* show that CG9010 is transcribed, as do our RT-PCR data from *D. guanche* and *D. subobscura* (Figure 1).

**Adh-Twain is transcribed**: RT-PCR using Adh-Twain-specific primers showed that it is transcribed in *D. subobscura* larvae, adult males, and adult females (Figure 1). The gene is also expressed in *D. guanche* larvae, adult males, and adult females (data not shown). 5’-RACE data from *D. subobscura* and *D. guanche* definitively show that the CG9010-like region of the fusion gene is part of a longer transcript that includes the entire Adh-like region in both *D. subobscura* and *D. guanche*. The RACE data also allowed us to experimentally identify the 5’-UTR. Our 3’-RACE data confirmed that the Adh-Twain transcript terminates near the polyadenylation site of the Adh-like 3’-UTR region.

We used RT-PCR to compare transcription between *D. subobscura* Adh-Twain and *D. subobscura* CG9010. Figure 1 suggests no significant differences in the expression patterns of Adh-Twain and CG9010, although CG9010 maybe slightly less abundant in males.

**Adh-Twain-predicted protein characteristics**: Basic characteristics of the predicted *D. subobscura* ADH-TWAIN protein are presented in Table 1. The predicted protein is ∼40% larger than ADH. Other than size, the most notable difference between ADH-TWAIN protein
vs. ADH and CG9010 is that ADH-TWAIN is much more positively charged at pH 7.0 than either CG9010 or ADH. The ADH-derived portion of ADH-TWAIN also appears to have increased in molecular weight relative to ADH.

The predicted amino acid sequence of ADH-TWAIN is substantially different from that of its ancestors, in both the ADH-derived and the CG9010-derived parts of the protein. For example, the ADH-derived portion of ADH-TWAIN is roughly 50% more diverged from ADH than D. subobscura ADH in comparisons vs. D. melanogaster ADH. Indeed, the ADH-TWAIN ADH-derived sequence is more diverged from D. melanogaster than any of the published Drosophila ADH proteins we have surveyed. These data suggest that it is unlikely that ADH-TWAIN retains significant ADH-like activity. In fact, prior allozyme studies of ADH in D. subobscura provided no evidence of a protein other than ADH that oxidized ethanol (Loukas et al. 1979; Balanya et al. 1994; Castro et al. 1999). Our own allozyme analysis confirmed these previous results and is consistent with the idea that ADH-TWAIN has reduced ability to catalyze the oxidation of ethanol. Of course, this experiment does not rule out the possibility that ADH-TWAIN can oxidize some other alcohol or related molecule (e.g., aldehydes).

Western blots (data not shown) revealed a single band of the size expected for ADH in D. melanogaster, D. subobscura, D. guanche, D. yakuba, D. pseudoobscura, D. virilis, and D. hydei. However, we observed no strong secondary band of the size expected for the predicted ADH-TWAIN. [Similarly, jingwei, which is known to produce a protein in vitro (Zhang et al. 2004), was not visible in our Western blot.] Overexposure of the Western blot revealed a number of minor bands, but it was impossible to determine which, if any, would correspond to the band associated with the predicted ADH-TWAIN. The failure of the ADH antibody to react with ADH-TWAIN is not unexpected, given the very high protein divergence between the ADH-derived portion of ADH-TWAIN and known ADH proteins in Drosophila.

5’ regulatory region of Adh-Twain shows similarity to that of CG9010: If the 5’ end of Adh-Twain is derived from a chromosomal duplication of CG9010, then we expect the 5’ region of Adh-Twain to harbor regulatory elements, such as promoters, derived from those of CG9010 (expression patterns of Adh-Twain and D. subobscura CG9010 are similar; Figure 1). To investigate this possibility, we gathered sequence data for ~600 bases 5’ of the D. subobscura Adh-Twain start codon and ~400 bp of 5’-flanking sequence from D. subobscura CG9010. These sequences were aligned to a sequence from the 5’-flanking region of D. pseudoobscura CG9010. We also used the VISTA browser (http://pipeline.lbl.gov/pseudo/) to qualitatively compare these sequences to D. melanogaster CG9010.

Figure 2 shows a multiple alignment of the 5’ regions of CG9010. It is immediately obvious that there are several highly conserved regions. Using a combination of our 5’-RACE data, published cDNA data, interspecific sequence comparisons, and promoter prediction methods (see MATERIALS AND METHODS), we identified the putative 5’-UTRs and putative promoter elements of Adh-Twain and CG9010. Of particular note is the conserved sequence between base 177 and base 357 of the Adh-Twain 5’ sequence in Figure 2. This region is very highly conserved in D. guanche Adh-Twain (data not shown), well conserved in D. pseudoobscura CG9010, and weakly conserved in D. melanogaster. The methods of Reese (2001) and Ohler et al. (2002) both suggest a promoter element in this region of Adh-Twain and CG9010 in D. subobscura. This region shows some similarity to TATA-less promoters seen in D. melanogaster (e.g., Burke and Kadonaga 1997). This conserved region may be important for regulation of Adh-Twain and CG9010 and could potentially contribute to their similar expression patterns. The Adh-Twain 5’-UTR is much larger than the CG9010 5’-UTR. Using parsimony, we infer that the Adh-Twain 5’-UTR represents the derived state and that these sequence insertions arose after duplication of CG9010. The potential biological consequences of this larger UTR are not known, although 5’-UTRs are known to play a critical role in translational regulation (Gray and Wickens 1998). Most other insertions/deletions relative to D. pseudoobscura are shared by D. subobscura CG9010 and Adh-Twain.

DNA polymorphism and divergence in Adh-Twain: We surveyed most of the Adh-Twain open reading frame and ~330 bases of the 5’-UTR and regulatory regions from 15 iso-female lines of D. subobscura. We also col-

**TABLE 1**

<table>
<thead>
<tr>
<th>Characteristics of Adh-Twain protein</th>
<th>Adh-Twain-D. subobscura</th>
<th>Adh-D. subobscura</th>
<th>CG9010-D. subobscura (for region conserved)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Molecular weight (Da)</td>
<td>39664.22</td>
<td>27566.79</td>
<td>10805.75</td>
</tr>
<tr>
<td>Amino acids</td>
<td>356</td>
<td>254</td>
<td>96</td>
</tr>
<tr>
<td>Isoelectric point</td>
<td>9.524</td>
<td>8.349</td>
<td>9.313</td>
</tr>
<tr>
<td>Charge at pH 7.0</td>
<td>19.495</td>
<td>2.523</td>
<td>5.983</td>
</tr>
</tbody>
</table>

Origin and Evolution of Chimeric Fusion Gene 211
Figure 2.—5′-UTR and promoters are similar for Adh-Twain (D. sub Fusion) and CG9010 and conserved across taxa.

lected sequence data from iso-female lines of D. guanche \((n = 1)\) and D. madeirensis \((n = 1)\). These data are summarized in Figure 3 and Table 2.

Several interesting patterns can be seen. First, polymorphism in the 5′-UTR is relatively low for a noncoding region, especially near regions hypothesized to play important regulatory roles. The most variable region is near the junction of CG9010-like and Adh-like sequences, in the area presumably derived from the Adh leader sequence. In contrast, DNA derived from the coding region of Adh has much lower levels of polymorphism. Patterns of polymorphism and divergence are correlated in Adh-Twain, consistent with the hypothesis of differing levels of functional constraint across the gene (see Figure 3A).

There are two in-frame indel polymorphisms in D. subobscura Adh-Twain, both of which are located near the junction of the CG9010-like and Adh-like regions (D1, bases 772–787 and D2, 814–820 on Figure 3). We used parsimony to infer that these were deletions relative to the ancestral CG9010 in D. subobscura and D. guanche (these amino acids are also present in CG9010 in D. melanogaster and D. pseudoobscura). Among the surveyed D. subobscura Adh-Twain alleles there were three deletion haplotypes: (1) no deletions \((1/15)\); (2) D2 only \((3/15)\); and (3) D1 and D2 \((11/15)\). Our D. madeirensis allele has both the D1 and the D2 deletions. D. guanche has neither of these deletions, but does have an in-frame 6-base deletion from base 889 to 895. All of these deletions occur in a region of high nucleotide polymorphism. This suggests that relative to most of the CG9010-like region and most of the Adh-like region, the intersection of these two regions is under relatively low constraint.

We also compared nucleotide divergence between the D. subobscura Adh-Twain and the ancestral D. subobscura CG9010 and Adh genes (gray lines in Figure 3, B and C). We looked at patterns of divergence between the D. subobscura and the D. guanche homologs of CG9010 and Adh (black lines in Figure 3, B and C). Comparison of these two sets of data (black vs. gray lines in Figure 3, B and C) shows a clear disconnect between the sites that tend to diverge between species in the parental genes and those that diverge between the Adh-Twain regions and their parental genes. This result, combined with the correlation between the polymorphism and divergence in Adh-Twain, suggests a shift in the divergence that is conserved in Adh-Twain.

As initially reported by Luque et al. (1997), the Adh region of Adh-Twain shows more codon bias than does Adh in D. subobscura. Table 3 shows that the entire Adh-Twain has substantial codon bias in all three species. Relative to other genes in the obscura group, Adh-Twain
Figure 3.—Major sequence features of \textit{Adh-Twain} compared to patterns of polymorphism and divergence. (A) Comparison of the nucleotide polymorphism observed at \textit{Adh-Twain} in our population of \textit{D. subobscura} (black line) and the nucleotide divergence between \textit{D. subobscura} and \textit{D. guanche Adh-Twain} (gray line). (B) Divergence between \textit{D. subobscura} and \textit{D. guanche CG9010} genes (black line) and divergence between \textit{D. subobscura Adh-Twain} and \textit{CG9010} (gray line). (C) Divergence between \textit{D. subobscura} and \textit{D. guanche Adh} genes (black line) and divergence between \textit{D. subobscura Adh-Twain} and \textit{Adh} (gray line). Both B and C suggest that, relative to \textit{Adh}, a distinctly different set of nucleotides are under constraint in \textit{Adh-Twain}.

\textit{Adh-Twain} appears to be a highly biased gene (Powell 1997). Codon bias, however, has decreased in the \textit{CG9010} region of \textit{Adh-Twain}, relative to \textit{CG9010}.

\textbf{Patterns of recent molecular evolution in \textit{Adh-Twain}:} The McDonald-Kreitman (MK) test is commonly used to detect directional selection acting on amino acid sequences (McDonald and Kreitman 1991). We used polymorphism data from \textit{D. subobscura} and divergence data from all three species in our tests. Table 4 shows that most analyses do not reject neutral evolution. The unpolarized MK test of \textit{D. subobscura} and \textit{D. guanche} for the \textit{Adh}-derived region of \textit{Adh-Twain} is significant at a $P < 0.05$ threshold without correcting for multiple tests. If we restrict our analysis to only fixations that occurred along the \textit{D. subobscura} lineage—as estimated by PAML as well as by parsimony (see materials and methods)—the MK test is no longer significant. In short, comparisons of the patterns of divergence and polymorphism among \textit{extant Adh-Twain} homologs shows no evidence of recent adaptive evolution.

Kern \textit{et al.} (2002) recently proposed a method for detecting directional selection using polymorphism data combined with two outgroup sequences. We used PAML to reconstruct ancestral sequences for \textit{Adh-Twain} (see materials and methods). Using this approach we could analyze all nucleotide fixations that had occurred along the \textit{D. subobscura} lineage since the \textit{D. subobscura}-\textit{D. madeirensis} speciation event. Consistent with the MK test data, we found no evidence for recent directional selection driving these fixations (analysis not shown).

\textbf{Rapid early evolution of \textit{Adh-Twain}:} Given the striking differences in the spatial patterns of divergence between \textit{Adh-Twain} and its \textit{CG9010} and \textit{Adh} ancestors (Figure 3, B and C), we also applied a contingency table analysis to compare the \textit{Adh-Twain} divergence and polymorphism to that of its ancestor genes (Table 4). This is not a canonical MK test as it compares paralogous loci within a species, rather than homologous loci between species. A significant disconnect between polymorphism and divergence does not mean we can reject
the neutral model, as we cannot rule out a fundamental shift in the neutral or nearly neutral mutation rate for one or more of these genes after the duplication events. Nevertheless, this is a useful way to detect a substantial change in the substitution rate at nonsynonymous or synonymous sites. The data indicate that there is no significant difference between the CG9010 ancestor and the CG9010-derived part of Adh-Twain. The Adh region, however, is marginally significant in the unpolarized comparison and highly significant in the polarized comparison. In the polymorphism data, there are \( \sim 2.5 \) synonymous polymorphisms per nonsynonymous polymorphism. In contrast, in the polarized divergence data we estimate only 0.333 synonymous fixations per nonsynonymous fixation—an eightfold difference. By comparison, Adh is estimated to have 2.6 synonymous fixations for every nonsynonymous fixation. Fisher’s exact test shows that the ratio of synonymous to nonsynonymous fixations in Adh is not significantly different from the ratio of synonymous to nonsynonymous polymorphisms in Adh-Twain (\( P = 0.999 \)). Both of these ratios are strikingly different from the ratio of synonymous to nonsynonymous fixations in the Adh-like regions of Adh-Twain (Adh-Twain fixations vs. Adh-Twain polymorphism, \( P = 0.0017 \); Adh-Twain fixations vs. Adh fixations, \( P = 0.0024 \)). These highly significant results suggest rapid amino acid divergence of Adh-Twain from its paralog Adh contrasts with our earlier comparisons of orthologous Adh-Twain sequences. From these results, we hypothesize that adaptive protein evolution in Adh-Twain occurred early in its history, prior to the speciation events leading to D. subobscura/D. guanche/D. madeirensis.

As the comparison of paralogous genes differs from the comparison of orthologous genes in many ways (e.g., time, genomic location), we applied the more suitable phylogenetic approach implemented in PAML (Yang 1997) to test the hypothesis that Adh-Twain evolved rapidly shortly after its formation (solid bars in Figure 4). Specifically, we estimated the nonsynonymous to synonymous rate ratio (\( d_S/d_K \)) for various branches of the Adh-Twain gene tree. In these analyses we separately investigated CG9010 and Adh, the corresponding regions of Adh-Twain.

We followed Yang (1998) and Yang et al. (2000) to test for rate heterogeneity between CG9010 and the CG9010-derived region of Adh-Twain (see Materials and Methods). Specifically, we compared the fit of the data to three hypotheses: (1) all branches were evolving at the same rate (“one-ratio” model 0 in PAML), (2) all branches are evolving at different rates (“free-ratio” model 1), and (3) the branches after the formation of Adh-Twain are evolving differently from the rest (“branch-specific” model 2). The free-ratio model fit the data better than the one-ratio model (model 0, \( \ln l = -951.5 \); model 1, \( \ln l = -937.1 \); \( 2\Delta lnl = 28.8, \text{ d.f.} = 9, P = 0.0007 \)). Clearly, the model of a single rate for \( d_S/d_K \) does not fit the CG9010 data.

The branch-specific model, which has one \( d_S/d_K \) rate for the fusion gene-related lineages and one rate for all other lineages, is a significantly better fit to the data than the one-ratio model (model 0 vs. model 2, \( \ln l = -943.1 \); \( 2\Delta lnl = 16.8, \text{ d.f.} = 1, P < 0.0001 \)). In contrast, the free-ratio model does not fit the data better than the two-ratio branch-specific model (\( 2\Delta lnl = 12, \text{ d.f.} = 8, P = 0.1512 \)). A three-ratio branch-specific model—in which there is one rate from the branch immediately

---

### TABLE 2

<table>
<thead>
<tr>
<th>Category</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of sequences</td>
<td>15</td>
</tr>
<tr>
<td>No. of bases surveyed</td>
<td>1428</td>
</tr>
<tr>
<td>Total sites (excluding gaps)</td>
<td>1381</td>
</tr>
<tr>
<td>No. of polymorphic (segregating) sites, S</td>
<td>73</td>
</tr>
<tr>
<td>No. of haplotypes, NHap</td>
<td>14</td>
</tr>
<tr>
<td>Polymorphism</td>
<td></td>
</tr>
<tr>
<td>Nucleotide diversity, ( \pi )</td>
<td>0.01378</td>
</tr>
<tr>
<td>( \theta ) (per site)</td>
<td>0.01626</td>
</tr>
<tr>
<td>Average no. of nucleotide differences</td>
<td>19.029</td>
</tr>
<tr>
<td>among D. subobscura alleles</td>
<td></td>
</tr>
<tr>
<td>Divergence</td>
<td></td>
</tr>
<tr>
<td>No. of fixed differences between</td>
<td>44</td>
</tr>
<tr>
<td>D. subobscura and D. guanche</td>
<td></td>
</tr>
</tbody>
</table>

### TABLE 3

<table>
<thead>
<tr>
<th>Region</th>
<th>Species</th>
<th>ENC</th>
<th>CBI</th>
<th>Nucleotides</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>D. guanche</td>
<td>37.3</td>
<td>0.679</td>
<td>288</td>
</tr>
<tr>
<td>CG9010</td>
<td>D. subobscura</td>
<td>37.8</td>
<td>0.679</td>
<td>288</td>
</tr>
<tr>
<td>D. madeirensis</td>
<td>42.4</td>
<td>0.581</td>
<td>288</td>
<td></td>
</tr>
<tr>
<td>D. subobscura*</td>
<td>38.6</td>
<td>0.617</td>
<td>288</td>
<td></td>
</tr>
<tr>
<td>Adh</td>
<td>D. guanche</td>
<td>48.1</td>
<td>0.456</td>
<td>762</td>
</tr>
<tr>
<td>D. madeirensis</td>
<td>47.9</td>
<td>0.449</td>
<td>762</td>
<td></td>
</tr>
<tr>
<td>D. subobscura</td>
<td>46.8</td>
<td>0.455</td>
<td>762</td>
<td></td>
</tr>
<tr>
<td>Adh</td>
<td>D. guanche</td>
<td>42.0</td>
<td>0.530</td>
<td>762</td>
</tr>
<tr>
<td>D. madeirensis</td>
<td>40.6</td>
<td>0.533</td>
<td>762</td>
<td></td>
</tr>
<tr>
<td>D. subobscura*</td>
<td>41.7</td>
<td>0.534</td>
<td>762</td>
<td></td>
</tr>
<tr>
<td>All*</td>
<td>D. subobscura*</td>
<td>43.7</td>
<td>0.479</td>
<td>1499</td>
</tr>
<tr>
<td>Adh-Twain</td>
<td>D. guanche</td>
<td>41.5</td>
<td>0.496</td>
<td>1050</td>
</tr>
<tr>
<td>D. madeirensis</td>
<td>43.7</td>
<td>0.485</td>
<td>1050</td>
<td></td>
</tr>
<tr>
<td>D. subobscura*</td>
<td>43.7</td>
<td>0.479</td>
<td>1050</td>
<td></td>
</tr>
</tbody>
</table>

ENC, effective number of codons; CBI, codon bias index. *Population mean of D. subobscura alleles. GC content of this region is 53%.
Interestingly, a four-ratio model, where all models of sequence evolution for the D. subobscura Adh-Twain region of CG9010 and one rate for all others—does not fit the data better than the two-ratio model (2Δlnl = 0.2, d.f. = 1, P = 0.62). Interestingly, a four-ratio model, where all branches after the CG9010 duplication are free to vary their \(d_s/d_S\) ratio, fits the data marginally better than the two-ratio model (2Δlnl = 7.4, d.f. = 2, P = 0.024). We are cautious of this later result given the large number of tests performed. Figure 4 illustrates these results.

Not surprisingly, \(d_S/d_s\) is small for CG9010 in most lineages (Figure 4). However, the \(d_S/d_s\) ratio of the CG9010-derived part of Adh-Twain early in its history is close to 1 (\(d_S/d_s = 0.9919\)). The D. guanche branch is clearly evolving slowly and is consistent with functional constraint (\(d_S/d_s = 0.3313\)). The only evidence for \(d_S/d_s\) significantly greater than 1 in the CG9010-derived portion of Adh-Twain is found in the D. subobscura lineage, which has a \(d_S/d_s\) of 2.09. This result is consistent with directional selection acting on the CG9010 portion of the D. subobscura Adh-Twain. This result, however, must be interpreted with caution as the \(d_S/d_s\) ratio is still close to 1 and MK tests described above—which also take into account polymorphism data—were not significant.

We repeated the above analysis for the Adh-derived region of Adh-Twain. Again, the free-ratio model fit the data substantially better than the one-ratio model (model 0, lnl = −2128.27; model 1, lnl = −2083.60; 2Δlnl = 89.34, d.f. = 15, P > 0.0001). We compared the one-ratio and the free-ratio models to several different models of sequence evolution for the Adh-Twain lineages. First, we compared a two-ratio model, with one \(d_S/d_s\) ratio for the Adh-Twain branch and one ratio for all other lineages. The two-ratio model fit better than the one-ratio model, but the free-ratio model fit slightly better than the two-ratio model (two-ratio model vs. model 0, two-ratio model lnl = −2097.46; 2Δlnl = 60.62, d.f. = 1, P > 0.0001; two-ratio model vs. model 1, 2Δlnl = 27.7, d.f. = 14, P = 0.0156). A three-ratio model, with rapid evolution after the formation of Adh-Twain and then a subsequent slowing down after the D. subobscura-D. guanche speciation, fit better than the two-ratio model (three-ratio model vs. two-ratio model, three-ratio model lnl = −2093.19; 2Δlnl = 8.54, d.f. = 1, P = 0.0035). Interestingly, the free-ratio model does not fit the data significantly better than the three-ratio model (2Δlnl = 19.18, d.f. = 13, P = 0.1176). We also compared several other models, none of which were significant improvements over the three-ratio model (analysis not shown).

For the three-ratio model, the \(d_S/d_s\) ratio for the branch immediately after the formation of Adh-Twain cannot be calculated, as \(d_s\) is 0. This implies a \(d_S/d_s\) ratio much greater than 1. In contrast, \(d_S/d_s\) ratio for...
Figure 4.—Nucleotide substitutions in Adh and CG9010 regions of Adh-Twain and their parental genes. These results are based on our PAML analysis. The solid bars represent the time period during which we hypothesize that directional selection acted on Adh-Twain. Estimates of the number of nonsynonymous and synonymous changes are beneath the branches of the Adh-Twain lineages. In both the Adh-derived and the CG9010-derived regions of the fusion gene there has been an increase in the rates of amino acid substitution, but not synonymous substitutions, relative to the ancestral genes.

The data suggest that CG9010 duplicated prior to the origin of the fusion gene, but after D. subobscura and its relatives split off from D. pseudoobscura. One of these copies subsequently fused with the Adh retrosequence, to produce the fusion gene, Adh-Twain. We do not know precisely when CG9010 duplicated. The fact that only one copy of CG9010 is found in D. pseudoobscura sets an upper bound for the CG9010 duplication of ~8.0–12.0 million years ago (MYA) and a lower bound of 1.8–2.8 MYA (Ramos-Onsins et al. 1998). Second, we do not know which copy of CG9010 participated in the fusion event. We found no evidence for DNA sequence homology to the 3′ end of CG9010 in the sequence data from the fusion gene. This is consistent with two hypotheses. Either the CG9010 target of the Adh retrosequence was a truncated, nonfunctional copy (e.g., Katju and Lynch 1997) or the 3′ end of the CG9010 target is no longer recognizable as a result of extensive molecular evolution.

Several puzzling results from the Luque et al. (1997) original data, including high codon bias and low $d_s/d_a$ ratio compared to the expectation for a putative pseudogene, are explained by our data. The fact that the putative initiation codon of the D. guanche Adh retrose-
Adh canonical has shifted away from secondary alcohols typically catalyzed by Drosophila Adh. The history of the CG9010-derived region of Adh-Twain is less clear. We showed that, unlike the Adh region, a “two-ratio” model was most likely, given our data. There is strong evidence for rapid evolution early in the history of Adh-Twain. There is also weak evidence that the CG9010-derived region is adaptively evolving along the D. subobscura lineage. However, the strong signal of directional selection—many more amino acid changes than silent changes—observed in the early evolution of the Adh region is obscured in the CG9010 region. Instead, the rates of both nonsynonymous and synonymous substitutions were elevated early in the history of this gene, followed by increasing constraint. The simplest interpretation is that the copy of CG9010 that ultimately became part of Adh-Twain evolved quickly under reduced functional constraint right after the initial duplication event, but then became more constrained once it was fused with the Adh retrogene.

Adaptive protein evolution: In general, patterns of polymorphism and divergence at Adh-Twain provide no strong evidence for recent directional selection in D. subobscura. Our analysis of substitution patterns along branches of the genealogy suggests that a “three-ratio” model has the greatest likelihood. In this model, Adh is normally very evolutionarily constrained. Shortly after Adh-Twain was formed, but prior to the splitting of the lineages leading to D. subobscura and D. guanche, there was a burst of adaptive substitutions in the Adh region. Subsequently, the Adh region of the fusion gene has slowed in its rate of evolution, although not nearly as slow as Adh and not along all lineages (e.g., D. guanche).

A strikingly similar pattern of rapid amino acid evolution shortly after the formation of jingwei was reported by Long and Langley (1993). Interestingly, one of the early, likely adaptive, amino acid changes in Adh-Twain (H190 to Q) occurs at a residue known to affect Adh activity in D. melanogaster mutants (http://www.flybase.org). This result is consistent with the Adh allozyme data from the obscura group, which provided no evidence for canonical Adh activity associated with the fusion gene. This may mean that the enzymatic activity of Adh-Twain has shifted away from secondary alcohols typically catalyzed by Drosophila Adh.

The function of the Adh-Twain protein, however, remains unknown.

Retrogene evolution: If a retrogene is to preserve an ancestral function or acquire a new function, it must avoid premature termination codons and be properly expressed. For CFGs originating by retrotransposition of a “donor” gene into a preexisting “acceptor” gene, reading frame preservation implies an insertion site between acceptor gene codons. Furthermore, it implies removal or avoidance of in-frame stop codons in the 5′-UTR of the donor retrosequence. It is also possible that some ultimately successful fusion genes originating by retrotransposition go through an early stage of loss-of-function prior to restoration of function by new mutations. Distinguishing between these possibilities is problematic as we cannot accurately reconstructing the details of the ancestral state of the insertional mutation (e.g., the retrotransposition of the Adh mRNA in the case of Adh-Twain) in all but the most recently derived CFGs. Nevertheless, the fact that Adh-Twain showed early rapid amino acid evolution and no silent substitutions suggests that this gene was functional immediately after it was formed (Long and Langley 1993 reached a similar conclusion for jingwei).
Adh and the origins of new genes: Recent studies in Drosophila and mammals have shown that retrotransposition of spliced mRNA's is a potent source of gene duplications (Eichler 2001; Betran et al. 2002; Devor and Moffat-Wilson 2003; Long et al. 2003a,b; Emerson et al. 2004). Long and Langley (1993) showed that retrotransposition events such as these can result in the formation of new CFGs, e.g., jingwei. The discovery of three novel Adh-derived Drosophila genes, Adh-Twain, jingwei, and Adh-Finnegan, strongly suggests that Adh-derived novel genes are common in flies.

What aspects of Adh biology might contribute to this phenomenon? Adh in D. melanogaster and other Drosophila can catalyze a variety of substrates ranging from various alcohols to a number of aldehydes (Ashburner 1998). Perhaps the frequency of Adh duplications and Adh-derived CFGs reflects the flexibility and evolutionary potential of the Adh protein or the fact that Adh is one of the most abundant transcripts in Drosophila.

Other factors relating to the structure and molecular biology of Adh may make it more likely to participate in retrotransposition-mediated novel gene formation. For example, Goncalves et al. (2000) identified common characteristics of genes that were the source of retropseudogenes in humans. The coding sequences of the parental genes tended to be relatively short, expressed in a variety of tissues, and have a low G/C content. Their amino acid sequences were highly conserved. Nearly a quarter of these genes produced more than one retropseudogene. Consistent with these observations, the 23 recent retrotransposed genes in D. melanogaster identified by Betran et al. (2002) are shorter than the average D. melanogaster gene (mean retrotransposed gene length, 320 amino acids (aa); mean gene length, 522 aa; median retrotransposed gene length, 230 aa; median gene length, 421 aa). The parental genes of these retrotransposed genes are also often expressed in multiple tissues (Betran et al. 2002). Adh in Drosophila qualitatively fits several of the above patterns as it is relatively short, widely expressed, and highly conserved and has duplicated several times in different Drosophila lineages (Yum et al. 1991; Nurminsky et al. 1996; Powell 1997; Amador and Juan 1999). In any event, the discovery of the obscura Adh-derivered chimeric fusion gene strongly motivates the search for additional novel Adh genes and opens up the possibility of discovering general principles governing the evolution of novel genes.

We thank C. Benyajati (and Uncle Billy) for the Adh antibody. We are also grateful to A. Davis, M. Aguadé, and the Drosophila Species Stock Center for flies. We thank S. Shih for technical assistance and D. Barbarsh for answering many technical questions. We thank M. Long for discussing his unpublished data with us. We also thank I. Dworkin, A. Kern, C. Langley, M. Lawiczak, S. Nuzhdin, T. Schlenke, and B. Wagstaff for discussion. We are grateful to A. Kern and I. Dworkin for comments on the manuscript. We also thank two thoughtful reviewers for many helpful suggestions. D. subobscura cDNA and genomic libraries are freely available from C.D.J. This work was funded by National Science Foundation grants to C.D.J. and D.J.B.

LITERATURE CITED


Burke, T. W., and J. T. Kadonaga, 1997 The downstream core promoter element, DPE, is conserved from Drosophila to humans and is recognized by TAFII60 of Drosophila. Genes Dev. 11: 3920–3931.

Castro, J. A., M. Ramon, A. Picornell and A. Moya, 1999 The genetic structure of Drosophila subobscura populations from the islands of Majorca and Minorca (Balearic Islands, Spain) based on allozymes and mitochondrial DNA. Heredity 83: 271–279.


Communicating editor: T. EickBush