Polymorphism Patterns in Two Tightly Linked Developmental Genes, *Idgf1* and *Idgf3*, of *Drosophila melanogaster*

Martina Žurovcová*,† and Francisco J. Ayala*,[1]

*Department of Ecology and Evolutionary Biology, University of California, Irvine, California 92697 and †Institute of Entomology, Czech Academy of Sciences, České Budějovice 37005, Czech Republic

Manuscript received July 5, 2001
Accepted for publication June 7, 2002

ABSTRACT

A new developmental gene family, recently identified in *D. melanogaster*, has been called imaginal disc growth factors (IDGF) because the proteins promote growth of cell lineages derived from imaginal discs. These are the first genes reported that encode polypeptide factors with mitotic activity in invertebrates. Characteristics such as similar arrangement of introns and exons, small size, and different cytological localization make this family an excellent candidate for evolutionary studies. We focus on the loci *Idgf1* and *Idgf3*, two genes that possess the most distinctive features. We examine the pattern of intra- and interspecific nucleotide variation in the sequences from 20 isogenic lines of *D. melanogaster* and sequences from *D. simulans* and *D. yakuba*. While MK, HKA, and Tajima's tests of neutrality fail to reject a neutral model of molecular evolution, Fu and Li's test with outgroup and McDonald's test suggest that balancing selection is modulating the evolution of the *Idgf1* locus. The rate of recombination between the two loci is high enough to uncouple any linkage disequilibrium arising between *Idgf1* and *Idgf3*, despite their close physical proximity.

A number of gene loci have been sequenced in Drosophila in the effort to elucidate patterns of nucleotide variation within and between species and to gain insight into the processes that maintain the variation and its contribution to adaptation. Most studies have focused on genes involved in cell metabolism. Whether or not the observed patterns of variation hold for other genes, such as regulatory or developmental ones, remains unknown. Relatively few evolutionary genetics investigations have concerned genes directly involved in early developmental stages of Drosophila. Interestingly, each gene has manifested distinctive patterns. The *transformer* gene (*tra*; Walthour and Schaeffer 1994) is highly polymorphic with respect to the amino acid replacements, but less so for silent polymorphism. The early-development gene *decapentaplegic* (*dpp*; Richter *et al.* 1997) appears to be constrained at the amino acid level with typical levels of silent polymorphism, while the *legless* or *runt* gene (*run*; Labate *et al.* 1999) exhibits expected levels of interspecific divergence, but low intraspecific polymorphism. One study has involved a group of closely related developmental genes, the *Ras* oncogenes (*Dras*; Gasperini and Gibson 1999). The duplication of these genes is ancient (their origin predates the origin of the major animal phyla), but the genes are extremely conserved and exhibit fairly similar levels of polymorphism, which may not be unexpected, although the three *Dras* loci have different cytological positions.

Recently, a new developmental gene family has been identified in *Drosophila melanogaster* (Kawamura *et al.* 1999). These genes are the first ones reported that encode polypeptide factors with mitotic activity in invertebrates. The family has been named imaginal disc growth factors (IDGF), because its proteins promote growth of cell lineages derived from imaginal discs. The IDGF proteins are structurally related to chitinases from which they may have evolved. The *Idgf* gene family has five (putatively six; P. J. Bryant, personal communication) members, all similar in length (from 2 to 3 kb, including introns) and with a similar arrangement of introns and exons (Figure 1). The first known member, the locus *Chit* (originally named *ds47*) was isolated by Kirkpatrick *et al.* (1995) and localized on the second chromosome at 53D. Kawamura *et al.* (1999) showed that three other genes, *Idgf1–3*, form a tight cluster at 36A on the same chromosome, but *Idgf4* is on the X chromosome at 9A. The putative sixth member, named *Idgf5*, is again on the second chromosome, at 55C.

We have initiated the study of the *Idgf* gene family by focusing on *Idgf1* and *Idgf3*. These are localized near the well-studied *Adh*, which can be used as a reference gene. *Idgf1* and *Idgf3* are tightly linked (they are ~4 kb apart), which implies that they may share their evolutionary history to a great extent. An intriguing feature of the *Idgf* genes is the organization of their introns. In Drosophila, genes possessing introns usually have a longer first intron, while the second intron, counted

[1]*Corresponding author:* Department of Ecology and Evolutionary Biology, University of California, 321 Steinhaus Hall, Irvine, CA 92697-2525. E-mail: fjayala@uci.edu

along the direction of transcription, is shorter (Kriventseva and Gelfand 1999). The pattern of introns in *Idgf3* is just the opposite, as the first intron is the shorter one. *Idgf1* differs from all the other *Idgf* genes by having only one intron.

We investigate patterns of polymorphism and divergence of *Idgf1* and *Idgf3*. Although high levels of polymorphism are displayed in both loci, *Idgf3* is evolving neutrally, while in *Idgf1* the heterogeneity of distribution of the polymorphisms suggests that balancing selection is affecting this gene. Both *Idgf1* and *Idgf3* are in a region with estimated intermediate frequency of recombination, which is sufficient to uncouple possible linkage disequilibrium between the two genes.

MATERIALS AND METHODS

Drosophila strains and genomic DNA preparation: We used genomic DNA from 20 lines isogenic for the second chromosome that was kindly provided by M. Aguadé. The lines represent a random sample derived from flies collected in Montblanc (Tarragona, Spain; Aguadé 1998, 1999). The strain of *D. yakuba* is from the National Drosophila Species Stock Center (Bowling Green, OH; no. 14021-0261.0) and *D. simulans* is line 5F from the laboratory of F. J. Ayala. Genomic DNA was extracted from 10 adult flies with the QIAamp tissue kit (QIAGEN, Chatsworth, CA) using the manufacturer's protocol.

Polymerase chain reaction (PCR), cloning, and nucleotide sequencing: For each line an ~2.0-kb region encompassing the *Idgf1* transcriptional unit was amplified using two primers (forward primer 5′-TGCAGACCCCTAAAAGTTGAG-3′ and reverse primer 5′-GCAGGGTCAAAACGTTGTGAC-3′). An ~2.5-kb region encompassing *Idgf3* was similarly amplified (forward primer 5′-CCAATTCCCGTGCTAAGTGTC-3′, reverse primer 5′-GACCGATTCGCCCAGACGTG-3′). PCR reactions were performed in a 100-μl volume of the ExTAKARA buffer containing 2.5 units of ExTAKARA Taq polymerase, 0.5 μM each of the forward and reverse primers, 0.2 mM dNTP, and 5 μl of genomic DNA. The cycling parameters for the amplification were initial denaturation 94° for 30 sec, 35 cycles of denaturation at 94° for 10 sec, annealing at 57° or 64° for 15 sec, and elongation at 72° for 2 min 30 sec. PCR products were purified with the Wizard PCR preps purification system (Promega, Madison, WI) and used for all analyses. PCR products from *D. simulans*, *D. yakuba*, and some *D. melanogaster* lines were cloned using the TA cloning kit (Invitrogen, San Diego) as a means of checking the sequences obtained directly as PCR products.

DNA sequencing was done with an ABI model 377 autosequencer using the Big Dye Terminator ready reaction kit according to the manufacturer's protocol (Perkin-Elmer, Norwalk, CT). We directly sequenced the purified PCR products. All sequences were determined for both strands with use of 10–12 overlapping internal primers. Sequences have been deposited in GenBank under accession nos. AF394691–AF394734.

Nucleotide alignment and statistical analyses: Sequences were edited and aligned both manually and with the assistance of the AutoAssembler, EditSeq, and MegAlign program packages. Statistical, phylogenetic, and molecular evolutionary analyses were conducted using DnaSP version 3.52 (Rozas and Rozas 1999), MEGA version 2b.2 (Kumar *et al.* 2001), PAUP (Swofford 1998), DNA Slider (McDonald 1998), and Recomb-Rate (Comeron *et al.* 1999). For comparisons with the other genes of the family, we used GenBank sequences (accession nos. AF102236–AF102239, AE003799, and U13825). In all statistical analyses, indels as well as complex nucleotide substitutions were excluded.

Since there is evidence for recombination in our data, coalescent simulations implementing the recombination parameter C were conducted to obtain confidence intervals for Tajima's D (Tajima 1989), Fu and Li's D, D^*, F, F^* (Fu and Li 1993), Kelly's ZnS (Kelly 1997), haplotype diversity Hd, and the number of haplotypes h (Nei 1987). We derived $C_{lab} = 4Nc$ (divided by 2 to account for the lack of recombination in males; Przeworski *et al.* 2001) from laboratory measurements of the exchange of flanking markers ($c = 1.611 \times 10^{-8}$ recombination/bp/generation; Comeron *et al.* 1999) and effective population size $N_e = 3 \times 10^5$ (Aquadro *et al.* 2001). Laboratory-based estimates are most likely underestimates since they are based only on crossing over, which does not account for gene conversion (Andolfatto and Nordborg 1998); therefore, these estimates are considered a conservative lower-bound value independent of our observed data.

RESULTS

Relationships within the *IDGF* gene family: Kirkpatrick *et al.* (1995) and Kawamura *et al.* (1999) found and confirmed the homology of the IDGF family polypeptides with mammalian chitinase-related proteins with respect to the estimated catalytic sites, but did not provide any further details about the members of Drosophila *Idgf* gene family. Using the protein and DNA sequences from GenBank and FlyBase, we have compared the exon/intron structures of these genes and inferred their phylogenetic relationships. In pairwise comparisons, the *Idgf* genes have 42–65% identity at the protein level (52–67% at the DNA level). The positions of the introns are similar in the various genes (Figure 1). We applied several different methods to obtain a phylogenetic tree—neighbor-joining (with Kimura two-parameter distance), maximum parsimony (branch-and-bound search method), as well as maximum likelihood (HKY-gamma substitution model), which were conducted on amino acid and DNA sequences (first plus second positions, third positions, all positions considered in separate analyses). All these procedures yielded similar topologies. The phylogenetic tree (Figure 2) shows that the three genes in the cluster on the 2L chromosome arm (*Idgf1*, *Idgf2*, and *Idgf3*) are more similar to each other than to the other three. On the contrary, the two genes on 2R (*Chit* and *Idgf5*) are not more closely related to each other than to any of the other genes. The position of *Chit*, *Idgf5*, and *Idgf4* is ambiguous, suggesting that either *Idgf5* or *Chit* could be the most diverged from all the other *Idgf* genes. Since the tree is unrooted, it is not possible to estimate the most probable ancestor. Using 2.5 mya (million years ago) as the time of split between *D. melanogaster* and *D. simulans* (Powell 1997, pp. 283–284), the average distance of *Idgf1* and *Idgf3* between these species ($d_a = 0.026$) yields a rate of evolution of IDGFs of 5.2×10^{-9} amino acid replacements/site/year. If we use 6 mya as

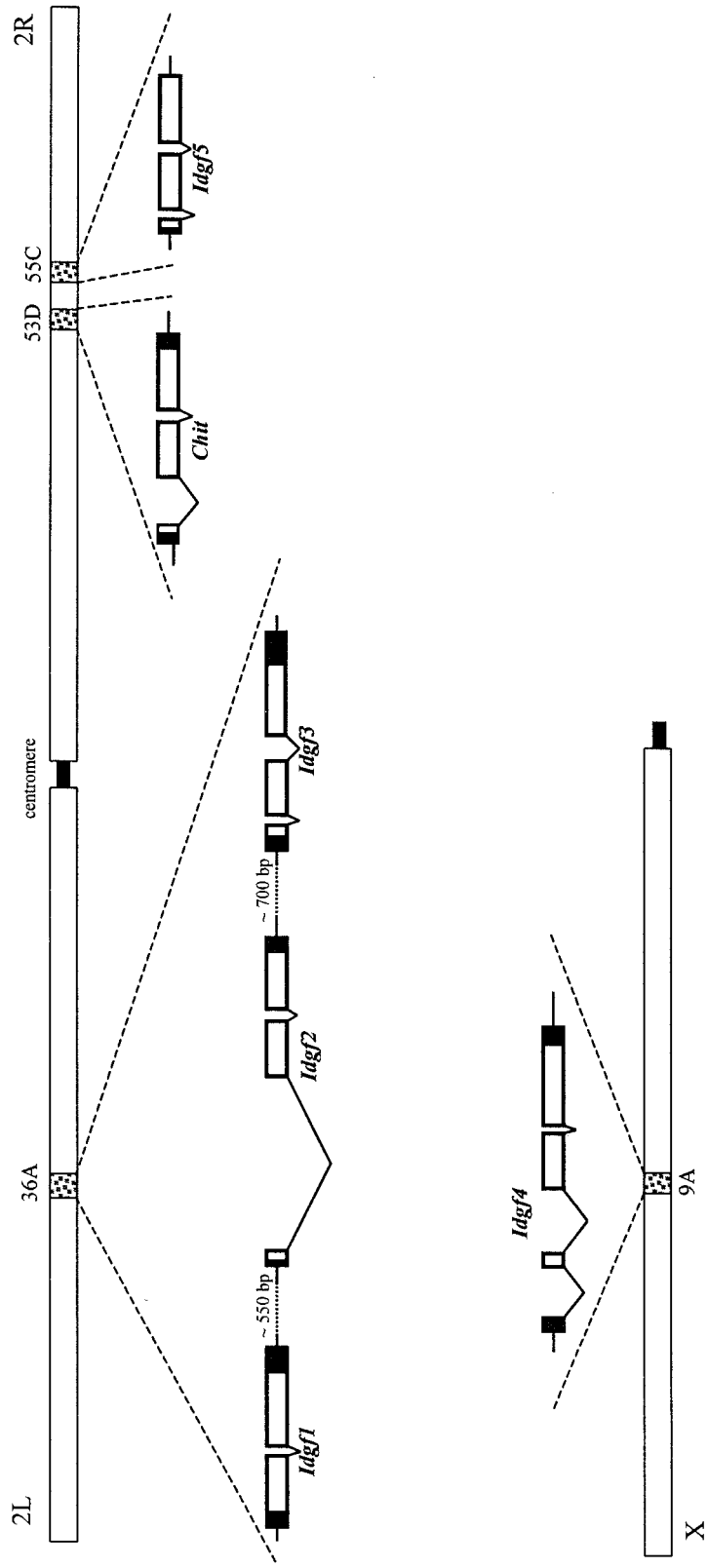

FIGURE 1.—Chromosomal localization and exon/intron structure of the *Idgf* genes. In the schemes of individual genes, exons are indicated by boxes; coding segments are indicated by open boxes, noncoding parts are indicated by solid boxes, and introns are indicated by forked lines. The stippled segments and the numbers above or below them refer to the location of the genes on the cytological maps of chromosomes 2 (top) and X (bottom). Dotted horizontal lines represent DNA segments separating the three genes on 36A, with the length given above the lines.

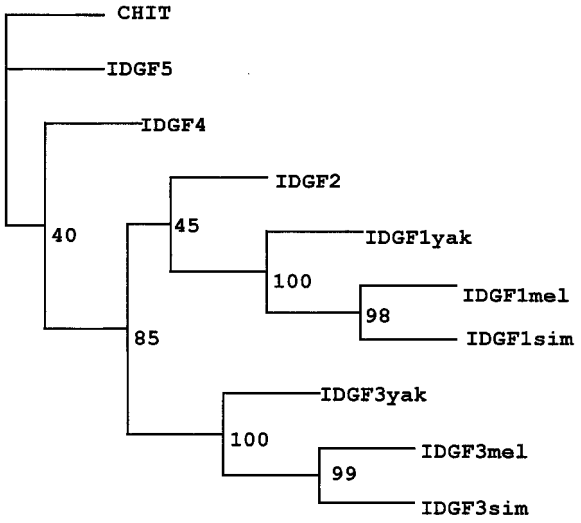

FIGURE 2.—Phylogenetic tree of the *IDGF* family, based on amino acid sequences, using maximum parsimony (branch and bound search, 10,000 replications to obtain the bootstrap values). mel, sim, and yak refer to *D. melanogaster*, *D. simulans*, and *D. yakuba*. The tree is unrooted. The numbers at the nodes are bootstrap values.

the divergence date between *D. yakuba* and *D. melanogaster* (d_a = 0.0525), the rate becomes 4.4×10^{-9}. The origin of the *Idgf1-3* cluster is estimated to be 56 and 61 mya, respectively.

We tested the hypothesis of the molecular clock by conducting a test of the homogeneity of substitution patterns between molecular sequences (measured by the disparity index I_D and probability levels assessed by 10,000 Monte Carlo simulations; KUMAR and GADAGKAR 2001). This revealed significant heterogeneity of *Chit vs. Idgf1* and *Idgf2*; *Idgf4 vs. Idgf1-3* cluster; and *Idgf5 vs. Idgf1* and *Idgf4* (data not shown). The time of the duplications cannot be reliably estimated.

Levels of nucleotide diversity and recombination: We have sequenced ~2 kb encompassing the whole coding sequence of *Idgf1* and 2.4 kb encompassing the whole coding region of *Idgf3* in 20 lines of *D. melanogaster*, one line of *D. simulans*, and one line of *D. yakuba*. The nucleotide polymorphic sites are displayed in Figure 3 and summary statistics are given in Table 1.

We found 69 biallelic single nucleotide polymorphisms in *Idgf1*, of which 6 are singletons. In the coding regions, 43 polymorphisms are synonymous and 9 are amino acid replacements. An 11-bp-long deletion occurs in the intron of 11 lines (the 11 bp are present in *D. simulans* and *D. yakuba*). Several indels occur at the 3′ end of the sequenced region, but due to the high variability among the *D. melanogaster* lines (as well as in *D. simulans* and *D. yakuba*), we could not obtain reliable alignments or estimate the precise number of deletions or insertions.

Idgf3 exhibits lower silent variability than *Idgf1*. There are 64 biallelic single nucleotide polymorphisms (16 singletons). In the coding regions, 18 polymorphisms are synonymous and 8 are replacements. A small insertion (3 bp) occurs in the second intron of seven lines. In the noncoding regions, there is one 3-bp insertion (second intron of line 80) and a 14-bp-long deletion (in the 3′ noncoding region of the third exon of line 8). At least four more indels ranging from 1 to 5 bp occur in the adjacent 3′ region.

The first exon of *Idgf1* exhibits more silent variation (33 polymorphisms) than the corresponding exons 1 + 2 of *Idgf3* exhibit (12 polymorphisms). For *Idgf1*, π = 12.86 for total polymorphism, which is much higher than the average observed in *D. melanogaster* of π_{total} = 4.43 (MORIYAMA and POWELL 1996; the highest value reported by these authors is π_{total} = 9.75 for *Amy-p*). Similarly for *Idgf1*, θ_{total} = 10.21 *vs.* the average *D. melanogaster* value of 4.41. The π and θ values of *Idgf3* are also higher than the averages of *D. melanogaster*, but not as high as those of *Idgf1* (see Table 1).

Estimates of the amount of divergence between *D. melanogaster* and *D. simulans* or *D. yakuba* are also given in Table 1. The two genes are about equally divergent, about twice as much from *D. simulans* as from *D. yakuba*, as expected, given the estimated times of divergence of *D. melanogaster* from *D. simulans* (2.5 mya) and from *D. yakuba* (~6 mya; MORIYAMA and POWELL 1996).

The amount of recombination was evaluated using Hudson's rate of recombination $R_{Hud} = 4N\mu$ (HUDSON 1987) and minimum number of recombinations R_M (HUDSON and KAPLAN 1985). Both parameters are higher for *Idgf1* (R_{Hud} = 0.02520 recombination/bp/generation, R_M = 16) than for *Idgf3* (R_{Hud} = 0.01324 recombination/bp/generation, R_M = 11). Recombination rates per gene (R_{Hud} = 48 for *Idgf1* and R_{Hud} = 31.3 for *Idgf3*) are higher than the laboratory-based estimates C_{lab} (18.41 and 22.45), thus confirming the choice of conservative lower-bound values.

Distribution of polymorphisms: Figure 4A displays a sliding window profile of nucleotide diversity (window size = 100 bp, step = 10 bp) that reveals considerable heterogeneity in levels of polymorphism across *Idgf1* and *Idgf3*, especially the former. *Idgf1* shows a high peak toward the middle of exon 1, which has no correspondence in *Idgf3*. Both genes exhibit a high peak in the 3′ noncoding terminal region.

A sliding window profile was also used to examine the distribution of divergence between *D. melanogaster* and *D. simulans* (Figure 4A). This demonstrates that although the shapes are largely coincidental across *Idgf3*, in *Idgf1* there is an apparent region of higher polymorphism and lower divergence in the first exon (approximately between positions 450 and 700).

Table 2 gives the results of the heterogeneity tests (KREITMAN and HUDSON 1991) for ascertaining the uniformity of the distribution of polymorphisms along each gene, tests performed only on silent sites. Contrasts are made successively for exons *vs.* the rest of the sequence,

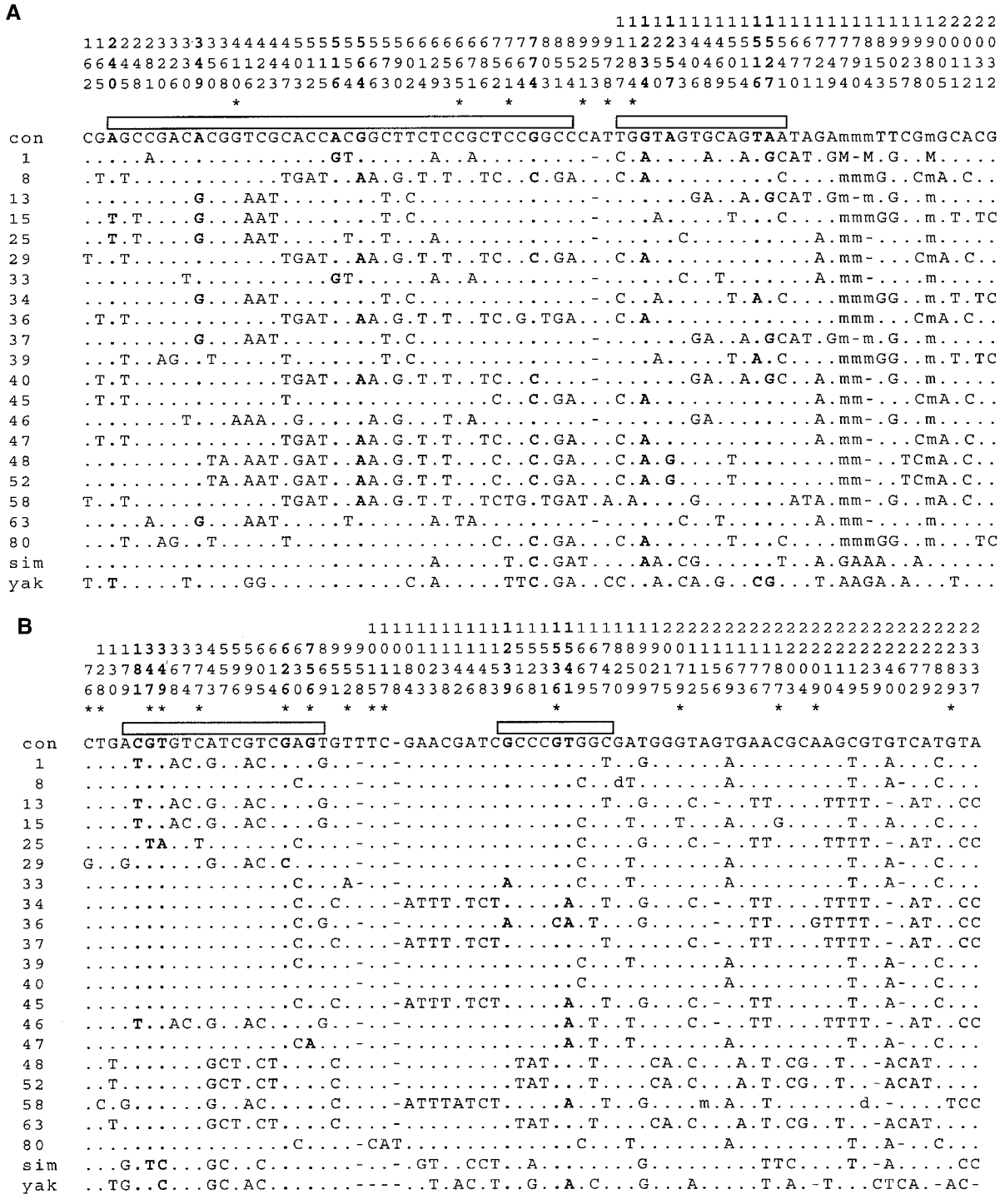

FIGURE 3.—Nucleotide polymorphic sites in the *Idgf1* (A) and *Idgf3* (B) loci in *D. melanogaster* (lines 1–80), *D. simulans* (sim), and *D. yakuba* (yak). Nucleotides identical to those in the consensus (con) sequence are represented by dots, singletons by asterisks, and indels by dashes, and complex mutations are marked on their first site by m. Only those sites that are polymorphic in *D. melanogaster* are shown for *D. simulans* and *D. yakuba*. The first exon starts at position 182 in *Idgf1* and at 21 in *Idgf3*. Replacement polymorphisms are in boldface type. The boxes above mark the coding segments. Recombination (R_M) has been detected in *Idgf1* between the following sites: [281,349], [350,416], [447,503], [554,573], [573,629], [653,671], [686,774], [774,851], [854,1127], [1127,1250], [1406,1448], [1505,1541], [1541,1670], [1671,1721], [1853,1905], and [1905,2005]. Recombination (R_M) has been detected in *Idgf3* between the following sites: [630,869], [869,1239], [1239,1541], [1541,1675], [1675,1859], [1909,2112], [2112,2176], [2176,2214], [2214,2219], [2219,2272], and [2279,2333].

TABLE 1

Nucleotide diversity of *Idgf1* and *Idgf 3* in *D. melanogaster*

Locus	No. of sites	*S*	π	θ	K_{sim}	K_{yak}
Idgf 1						
Nonsynonymous	1008	9	0.00327	0.00252	0.01366	0.02696
Synonymous	309	43	0.04990	0.03924	0.15144	0.25349
Noncoding	588	17	0.00985	0.00815	0.05156	0.11111
Total	*1905*	*69*	*0.01286*	*0.01021*	*0.04745*	*0.08820*
Idgf 3						
Nonsynonymous	1010	8	0.00145	0.00223	0.00963	0.01752
Synonymous	313	18	0.02036	0.01622	0.11455	0.28939
Noncoding	1041	38	0.01182	0.01029	0.06423	0.16926
Total	*2364*	*64*	*0.00852*	*0.00763*	*0.04762*	*0.11883*

Noncoding regions include introns. *S* is the number of polymorphic sites; π and θ are as in Nei (1987). Divergence *K* is the average number of nucleotide differences per site between species (Nei 1987), calculated for the entire sequenced region (1.95 kb for *Idgf1* and 2.39 kb for *Idgf 3*). Subscript *sim* indicates divergence from *D. simulans*, subscript *yak* divergence from *D. yakuba*.

coding *vs.* noncoding, and five (seven for *Idgf3*) adjacent segments, each including the same number of silent sites as the other four. The tests are significant for *Idgf1* mostly due to the unusual polymorphism observed in its exon 1. The only *Idgf3* test that is significant can be attributed to the excess of polymorphisms in its 3′ flanking region.

The McDonald test (McDonald 1996, 1998) allows us to detect heterogeneity in the distribution of variability across the loci by means of a sliding window and the ratio of intraspecific polymorphism to interspecific divergence. Table 3 shows the values of relevant statistics, four indicating statistical significance for *Idgf1*, but not for *Idgf3*. Figure 4B displays the sliding window plots of the ratio of intraspecific polymorphism to interspecific divergence for silent polymorphism. The peak in the profile for *Idgf1* is coincidental with the peak of polymorphism observed in Figure 4A. For Table 3, we used simulations with several values of the recombination parameter (r = 0, 2, 4, 8, 16, 32, 64, 128) and 1000 repetitions. Then we ran another set of simulations, with a narrower range of the recombination parameter but with 10,000 repetitions to get more accurate estimates. The results (Figure 4B and Table 3) show, as in previous tests, significant heterogeneity for *Idgf1* but not for *Idgf3*.

Linkage disequilibrium and recombination: Two tests for intralocus as well as interlocus (*Idgf1 vs. Idgf3*) nonrandom associations were used: the conservative Fisher's exact test for independence between sites (with Bonferroni correction for multiple comparisons) and Kelly's *ZnS* (Kelly 1997). Only informative sites, *i.e.*, those with at least two copies of the rarer variant present in the sample, were included in the analysis. Results are presented in Table 4; the distribution of the significant pairwise comparisons in individual loci is given in Figure 5. There is little evidence for linkage disequilibrium between the two loci (if we exclude the within-locus comparisons, there are only 108 significant comparisons and 0 after the Bonferroni correction). Different patterns can be seen for each of the two loci.

The significant value of *ZnS* for *Idgf3* indicates overall linkage disequilibrium, although the number of significant comparisons after Bonferroni correction is low. Figure 5B shows that the pairs of polymorphic sites with significant linkage disequilibrium do not form any obvious clusters, especially if only the lower levels of P (<0.05) or coding parts of the gene are considered. This is in accordance with the estimated amount of recombination. A minimum number of 11 recombination events was inferred by the four-gamete test (R_M; Hudson and Kaplan 1985), affecting the major part of the sequenced region.

Idgf1 displays a rather different picture. While the *ZnS* test points toward a nonsignificant amount of linkage disequilibrium, 16 comparisons remain significant after the Bonferroni correction of Fisher's test results. Closer inspection reveals an even more distinct pattern, as 12 out of these 16 comparisons are clustered within a very short region of 151 bp (from site 503 to 653). Despite the high minimum number of recombination events (R_M = 16) estimated for *Idgf1*, and despite the fact that 8 of these events are localized in the first exon, considerable clustering of the pairs of polymorphic sites with significant linkage disequilibrium in exon 1 can be observed overall (Figure 5A).

Haplotype structure: Significant linkage disequilibrium may result in haplotype structuring, while recombination counteracts that effect. Therefore, it is not unexpected that the estimates of haplotype diversity and haplotype number (Table 5) for both loci do not deviate from the neutral expectations. However, this applies also to the 503–653 region [number of haplotypes, h = 7; confidence interval = $(5; 11)_{95}$ with C_{lab} = 1.46 recom-

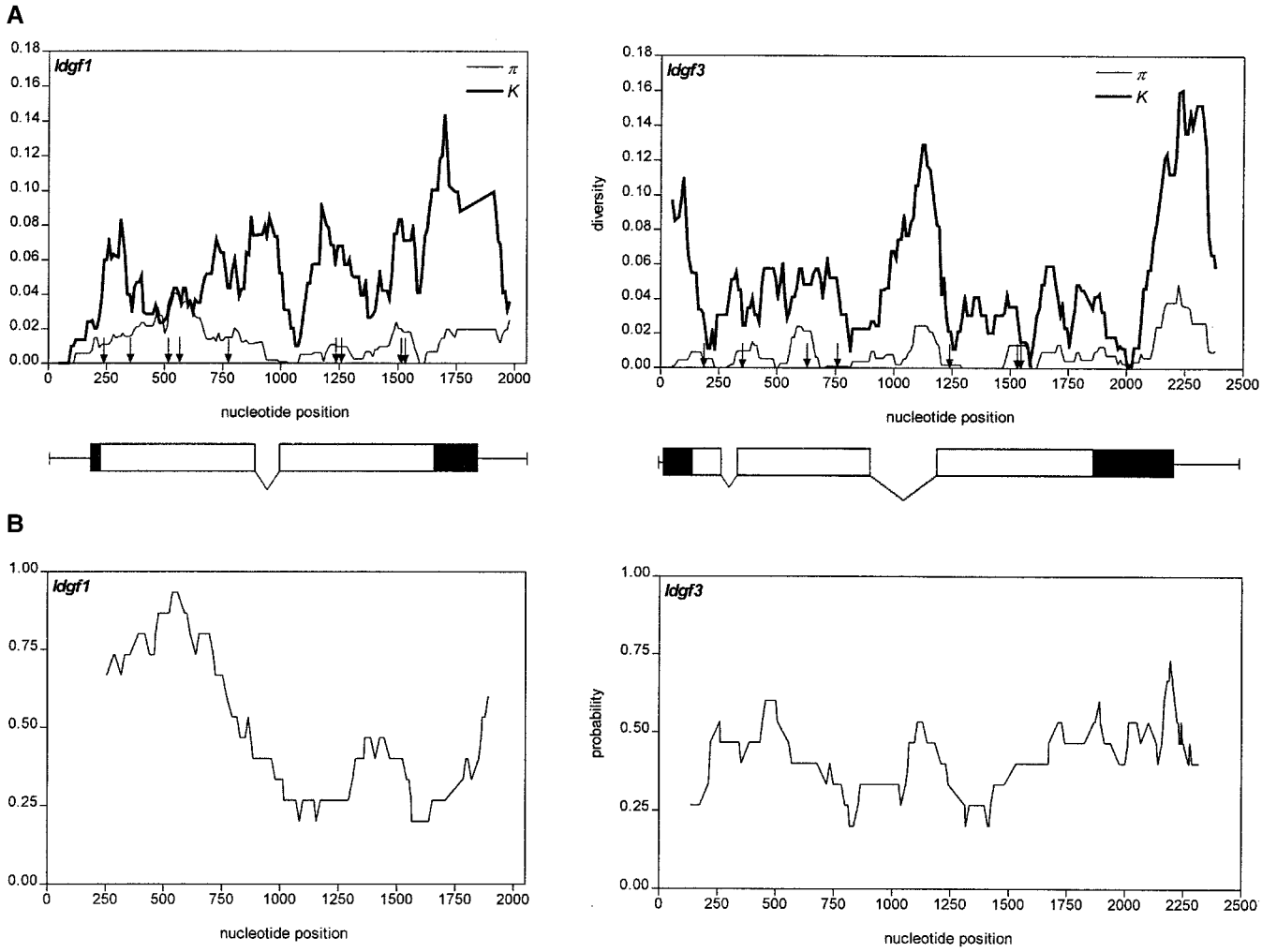

FIGURE 4.—(A) Sliding window profile of polymorphism π within *D. melanogaster* and divergence *K* between *D. melanogaster* and *D. simulans* in *Idgf1* (left) and *Idgf3* (right). Size of the window, 100 bp; step, 10 bp. All variable sites are considered except alignment gaps between *D. melanogaster* and *D. simulans*. Open bars indicate coding regions, solid bars noncoding regions, forked lines introns, and horizontal lines flanking regions. Arrows point to the approximate positions of replacement polymorphisms. (B) Sliding window plots of silent polymorphism-to-divergence ratio in *Idgf1* (left) and *Idgf3* (right). Flanking regions are not included. Window size is 15 bp.

bination/gene; 10,000 coalescent simulations]. Two replacement polymorphisms are within this region—one at position 516 (A/G with the former being the prevalent) and one at position 564 (G/A). In fact, it appears that the latter replacement is associated with three other synonymous polymorphisms (positions 512, 515, and 614) and that those are linked to four more (positions 503, 590, 614, and 653). This combination is found in lines 8, 29, 36, 40, 47, 48, 52, and 58, which could be arbitrarily labeled as haplotype A, while the rest might be labeled as haplotype G. Maximum parsimony analysis was carried out to confirm this partitioning (Figure 6A). Since line 46 was grouped with the haplotype A lines, gene conversion analysis was conducted for the 503–653 region (BETRAN *et al.* 1997). One conversion track was detected as expected in line 46, spanning the sites 512–564. The phylogram (Figure 6A) furthermore revealed that haplotype A is the younger class (in accordance with the derived state of the associated sites as compared with *D. simulans* and *D. yakuba*), with the older haplotype G being more differentiated. Less nucleotide variability is found within the two haplotype classes than between them in the 503–653 region, reflecting the strong linkage disequilibrium in this region (Figure 6B).

Tests of neutrality: We have used the Hudson-Kreitman-Aguadé (HKA; HUDSON *et al.* 1987) and McDonald-Kreitman (MK; MCDONALD and KREITMAN 1991) tests to examine the neutral hypothesis that levels of polymorphism and divergence are correlated and the tests of Tajima (TAJIMA 1989) and FU and LI (1993) to examine the hypothesis that all substitutions at the locus are neutral. Where appropriate, the significance of a test or parameter has been verified by coalescent simulation (10,000 repeats). Table 5 presents the significant tests.

TABLE 2

Heterogeneity tests of *Idgf1* and *Idgf3*

Locus and type of division	χ^2_C	*P*	χ^2_L	*P*
Idgf1				
(Exon/rest of sequence) functional regions	11.45	*	32.78	***
(Coding/noncoding) functional regions	18.30	**	49.18	***
Five regions with equal number of silent sites[a]	11.65	*	32.90	***
Idgf3				
(Exon/rest of sequence) functional regions	5.96	NS	10.04	NS
(Coding/noncoding) functional regions	10.63	NS	22.04	**
Seven regions with equal number of silent sites[a]	4.90	NS	10.58	NS

χ^2_C and χ^2_L are two separate tests according to Kreitman and Hudson (1991). NS, not significant. *Statistically significant $P < 0.05$. **Statistically significant $P < 0.01$. ***Statistically significant $P < 0.001$.

[a] We subdivide *Idgf1* into five regions, but *Idgf3* into seven, because the former is shorter and has only one intron rather than two.

For the HKA test we have contrasted *Idgf1* and *Idgf3* with one another, with loci *Acp26Aa*, *Acp26Ab*, and *Acp29AB* sampled from the same population previously (Aguadé 1998, 1999), and with the *Adh* 5′ end (Hudson *et al.* 1987). Several possibilities were explored: using the whole region sequenced, only the coding regions, and contrasting different parts of the regions (data not shown). Only three tests were significant: the contrast between the *Idgf1* coding region and the *Adh* 5′ region and *Idgf1* contrasted with *Acp26Aa* and *Acp29AB*. Since the two *Acp* loci demonstrate selection, only the former test is taken as truly significant ($P = 0.0460$ for total and $P = 0.0363$ for silent sites). However, it has to be noted that the MK and HKA tests are quite conservative, based on the assumption of no recombination. Since we have ample evidence for high values of recombination in our data, these tests are probably not sensitive enough.

Tests based on the frequency spectrum such as those of Tajima and Fu and Li were far more revealing (Table 5). The values of these statistics are positive and significantly so in the case of *Idgf1*, which would suggest a possibility of balancing selection, recent bottleneck, or population structure affecting these loci.

DISCUSSION

Evolution of the IDGF gene family: Our observations corroborate that the *Idgf* genes in *D. melanogaster* form a small gene family with six members, which has been previously indicated by their homology at the DNA and protein level and by similar functions. *Idgf5* was found with the assistance of a computer search of the *D. melanogaster* genome, but we assume that it also is a functional gene.

We have recovered adequate homologs for at least two genes of the *Idgf* gene family from *D. simulans* and *D. yakuba*, which suggests that some of the duplications that gave origin to this gene family predate the split of these species. This is confirmed by phylogenetic analysis (see Figure 2). Phylogeny shows that the genes in the tight cluster in chromosome arm 2L (*Idgf1*, *Idgf2*, and *Idgf3*) are more similar to each other than to the rest, but the two genes in 2R (*Idgf5* and *Chit*) are not closely related. From the exon/intron structure of the *Idgf* genes, it seems likely that the ancestral gene had at least two introns and that subsequently one intron was lost in *Idgf1*.

There is no obvious evidence of concerted evolution acting on this gene family. Homogenization is more

TABLE 3

Heterogeneity tests using polymorphism-to-divergence ratio

Locus	G_{max}	K_R	D_{KS}	V_{IL}	Q_{IL}	G_{mean}
Idgf1	0.0117 (4)**	0.2394 (32)	0.0038 (4)**	0.2096 (32)	0.0834 (24)*	0.0157 (12)**
Idgf3	0.5121 (12)	0.2107 (24)	0.6817 (12)	0.1191 (24)	0.3058 (32)	0.1885 (12)

The numbers displayed are the highest probability values for each test, with the corresponding recombination parameter *R* (in parentheses). G_{max}, maximum sliding G statistic; K_R, runs statistic; D_{KS}, Kolmogorov-Smirnoff statistic; V_{IL}, interval length variance; Q_{IL}, modified interval length variance; G_{mean}, mean sliding *G* statistic. Divergence is between *D. melanogaster* and *D. simulans*. *$P < 0.05$, **$P < 0.01$. Based on McDonald (1998).

TABLE 4

Linkage disequilibrium

Locus	Sites compared	Informative sites	Significant comparisons[a] (%)	*ZnS*[b]	C_{1ab} (per gene)	95% range values
Idgf1	1953	63	307/16 (15.72)	0.1506 NS	18.41	(0.09476; 0.22863)
Idgf3	1128	48	243/5 (21.54)	0.2145 *	22.85	(0.08873; 0.20960)
Idgf1 vs. Idgf3	6105	111	658/20 (10.78)	0.1205 NS	41.264	(0.08354; 0.15729)

C_{1ab}, recombination rate (see MATERIALS AND METHODS for explanation). Range values were obtained by 10,000 coalescent simulations (DnaSP version 3.52; ROZAS and ROZAS 1999). NS, not significant. *$P < 0.05$.

[a] Number of significant comparisons using the Fisher's exact test without/with Bonferroni correction, percentage of significant comparisons in parentheses.

[b] Kelly's parameter *ZnS* (KELLY 1997).

likely to affect genes in tight clusters, but close inspection of *Idgf1*, *Idgf2*, and *Idgf3* does not manifest any evidence of it, in spite of their very close physical linkage. Our overall observations indicate that none of the members from the *Idgf* gene family has become a pseudogene.

Patterns of nucleotide variation in *Idgf1* and *Idgf3*: The levels of polymorphism of *Idgf1* and *Idgf3* are higher than the average values in *D. melanogaster* genes. This contrasts with the situation in other developmental genes investigated so far such as *tra*, *dpp*, *run*, and *Dras* (see Introduction). Rather, the levels of polymorphism of these two genes are comparable to those of genes involved in metabolic pathways, such as *Adh*, *Amy*, or *Est-6* (MORIYAMA and POWELL 1996). Our data came from a single local population (Montblanc in Spain), which raises the possibility that this population may have a distinctive pattern of polymorphism rather than the levels typical of these genes in other populations. There are, however, no grounds for attributing particularly high levels of polymorphism to this population. Rather, several sex-associated genes, investigated in the same population, yield typical levels of polymorphism (AGUADÉ 1998, 1999), which in European populations are lower than those typical in African samples of *D. melanogaster*. Chromosomal location *per se* would not seem likely to

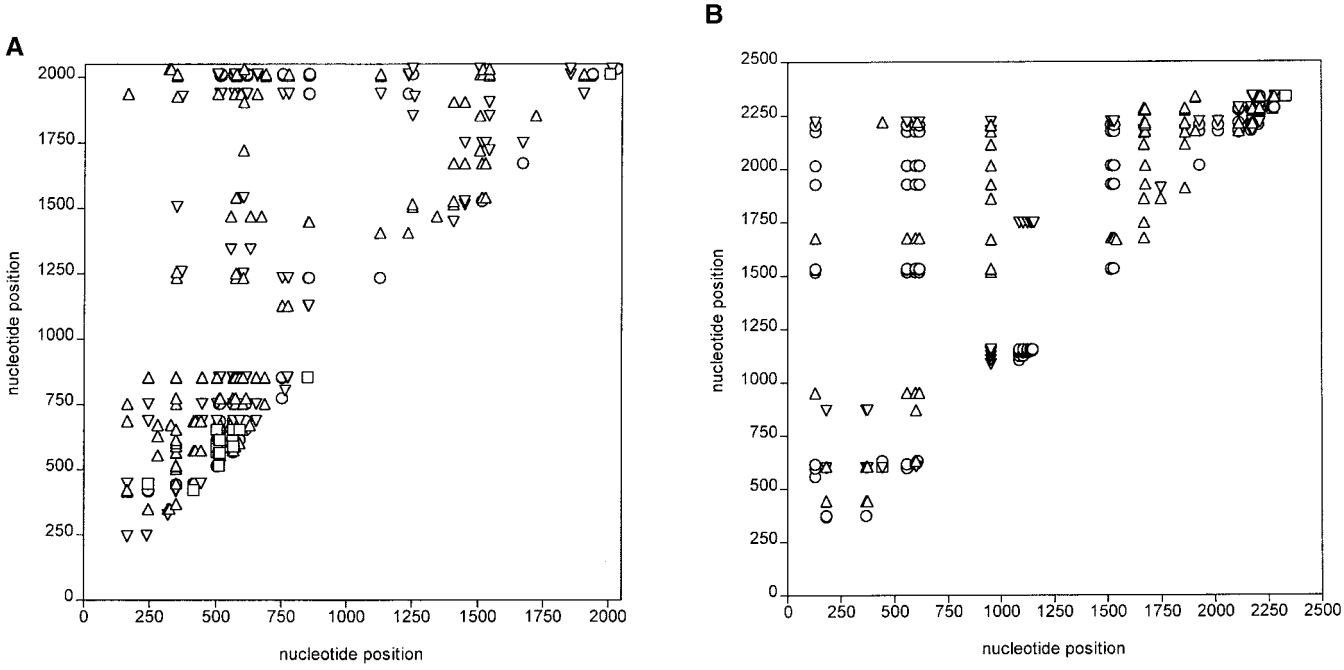

FIGURE 5.—Nonrandom associations (linkage disequilibrium) in *Idgf1* (A) and *Idgf3* (B). Significance was determined by Fisher's exact test for pairwise comparisons among all pairs of informative polymorphic sites. △, $P < 0.05$; ▽, $P < 0.01$; ○, $P < 0.001$; □, significant with Bonferroni correction.

TABLE 5

Tests of neutrality for *Idgf1* and *Idgf3*

	Idgf1		*Idgf3*	
	Parameter value	95% range values	Parameter value	95% range values
Hd	0.989	(0.926; 0.995)	0.979	(0.932; 0.995)
h	18	(12; 19)	17	(13; 19)
Tajima's *D*	1.115	(−1.420; 1.311) NS	0.269	(−1.462; 1.443) NS
Fu and Li's *D*	1.640	(−1.736; 1.358) *	0.600	(−1.976; 1.372) NS
Fu and Li's *F*	1.792	(−2.015; 1.431) *	0.598	(−1.973; 1.558) NS

Estimates of haplotype diversity *Hd* and number of haplotypes *h* (Nei 1987) are based on the whole sequenced regions. Neutrality tests were conducted on the coding parts of the loci, with *D. simulans* as an outgroup. Range values were obtained by coalescent simulations (10,000 repeats; DnaSP version 3.52; Rozas and Rozas 1999). See Table 4 for the parameter C_{lab}. NS, not significant.

account for the high level of polymorphism of these genes since they are tightly linked, yet *Idgf1* is more polymorphic than *Idgf3*.

The sliding window profile shows that for either gene the distribution of polymorphisms across the locus is not uniform. In each locus is a peak indicating high variability at the 3′ and adjacent intergenic region. This may not be surprising, because this region is not transcribed, and thus may be less constrained. Intragenic variability of *Idgf3* is more or less homogeneous, but the first exon of *Idgf1* exhibits much more variation than any other part of the gene exhibits. This variability is expressed in several ways: loss of an intron and a high number of point mutations, including one replacement polymorphism in the exact putatively catalytic site, which is conserved in all other *Idgf* genes. The change replaces polar serine (AGC) with nonpolar glycine (GGC), which is a not-very-unusual alteration. According to Harlow and Lane (1988), a change of this sort happens in 13% of all serine replacements by another amino acid. In *D. melanogaster*, both of these codons were described as preferred (Akashi 1995). Serine in this position is conserved in *D. simulans* and *D. yakuba*, and we have observed the replacement in only 2 out of the 20 *D. melanogaster* lines, which may be interpreted as a relatively recent event or as a transitional stage of a deleterious mutation, on its way to be removed completely. But it remains to be determined whether the replacement is adaptive, neutral, or deleterious; phenotypic analysis may answer that.

Although we could not check the sampled lines for inversions, previous work indicates that *In(2L)t* has a frequency <5% in this otherwise standard population (M. Aguadé, personal communication). Since inversions usually suppress recombination, it is unlikely that they could account for the high variability of *Idgf1* and *Idgf3*.

There is considerable nonrandom association between pairs of informative polymorphic sites within both *Idgf1* and *Idgf3*, but almost no linkage disequilibrium between them. While in *Idgf3* the linkage disequilibrium can be explained by mutation-drift balance, in *Idgf1* a cluster of linked sites was detected, defining two haplotype classes A and G. Haplotype A is most likely recent, which is shown by the phylogenetic analysis. The phylogram also corroborates the division into two haplotypes, so that the categories are not completely arbitrary. There are several possibilities as to how haplotype A may have originated: epistatic or balancing selection or demographic processes could be the cause of the observed pattern.

The two selection models may be distinguished on the basis of the pattern of variability within and between the two haplotype classes in that 151-bp region and the amount of recombination between them (Kirby and Stephan 1995). Epistatic selection would lead to highly differentiated haplotypes, with recombinants between these classes being removed. The amount of diversity within each haplotype would be less than the amount of diversity between them. Contrary to that, recent balancing selection would result in little variation within the younger haplotype class, while the older one should be more variable. Figure 6B may point to epistatic selection (more variability found between the haplotypes), but the estimates of nucleotide diversity favor balancing selection. While the overall diversity of the 151-bp region ($\pi = 0.03845$) is higher than the estimates within the haplotypes, the presumably older haplotype G exhibits variability about 2.7 times higher ($\pi_G = 0.01613$) than that of the younger haplotype A ($\pi_A = 0.00589$).

Demographic effects such as the expansion of a few haplotypes after a founder event or recent incorporation of one haplotype into the population offer another explanation. Such history would affect the whole genome in similar fashion. Our data do not support this, as the locus in close physical proximity, *Idgf3*, does not exhibit a parallel pattern. However, the latter model still seems a plausible explanation for the high number of synonymous polymorphisms observed in the first codon of *Idgf1*. The occurrence of the new haplotype

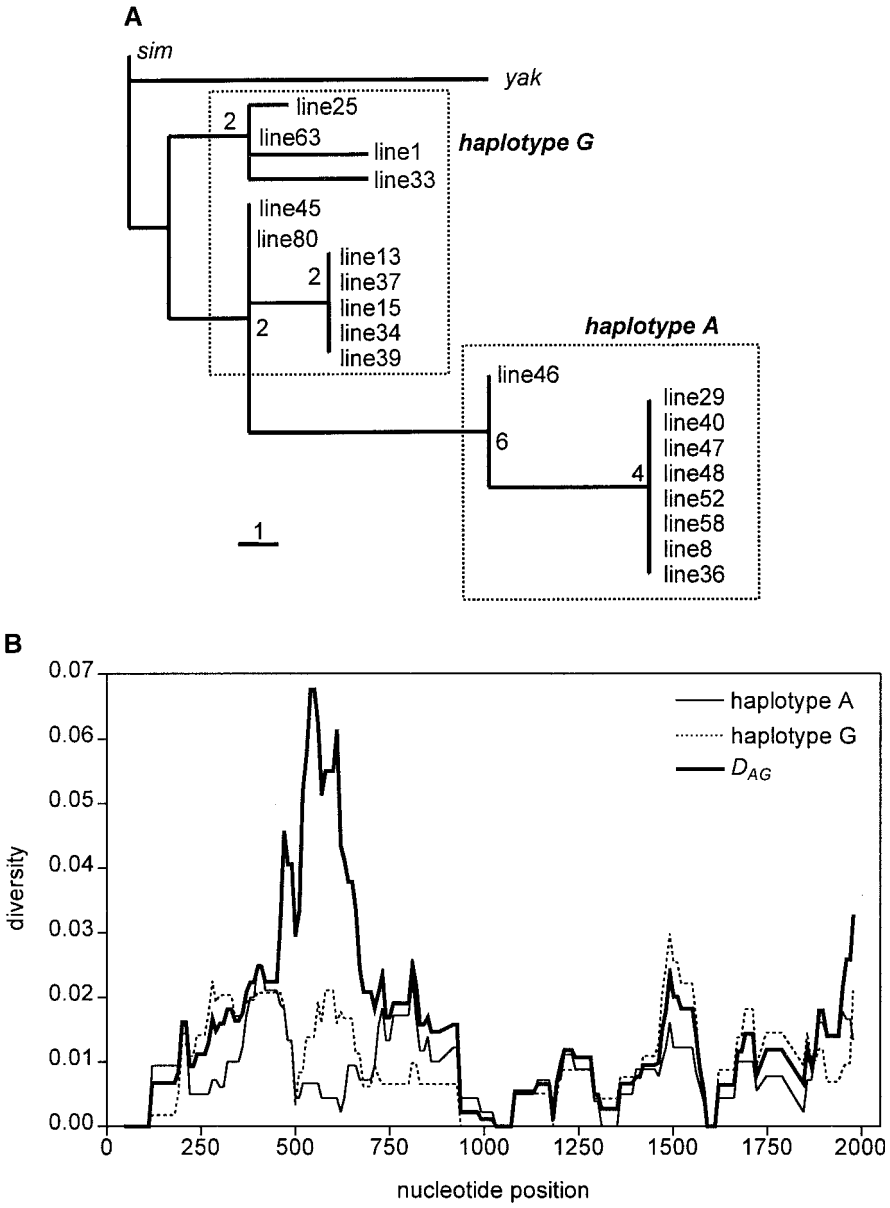

FIGURE 6.—(A) Phylogram of the most parsimonious tree for the 151-bp region spanning sites 503–653. Amino acid replacements are weighted 3:1 against silent changes. *D. simulans* and *D. yakuba* are used as outgroups to root the tree. Numbers adjacent to each branch are the numbers of inferred mutational steps. Dotted outlines indicate the two haplotypes. (B) Sliding window profile of the distribution of nucleotide diversity within and between (thick solid line) haplotype classes. Window size, 100 bp; step, 10 bp.

would have to be on such a time scale that recombination could break the linkage disequilibrium of that new haplotype, leaving just "footprints" in regions under selection. Since we have a sample from only one population, a population process like that cannot be completely excluded.

Selection at *Idgf1* and *Idgf3*: The tests of McDONALD and KREITMAN (1991), TAJIMA (1989), and FU and LI (1993) applied to the *Idgf1* and *Idgf3* loci fail to reject the null hypothesis that most of the variation in these loci is neutral, although some values for *Idgf1* are larger than one (Tajima's *D*, Fu and Li's *D**, and *F**; data not shown), which might suggest some form of balancing selection affecting this locus. The HKA (HUDSON *et al.* 1987) test also is significant for *Idgf1* when only synonymous coding substitutions are taken into account. Further indication of balancing selection in *Idgf1* comes from the significant results obtained with FU and LI (1993) tests, when comparisons are made with *D. simulans*. These tests are among the most powerful, because the outgroup leads to more precise estimation of its parameters. Evidence is also given by the McDonald's test (McDONALD 1996), incorporating six statistics designed to detect heterogeneity in the ratio of polymorphism to divergence across the studied region. Three of them (D_{KS}, G_{max}, and G_{mean}), considered to be the most powerful for identification of the possible pattern caused by balancing selection, were significant with $P(D_{KS}) = 0.0038$, $P(G_{max}) = 0.0117$, and $P(G_{mean}) = 0.0157$.

When the model of balancing selection is estimated from the data, it is desirable to identify which of the linked sites is the one under selection. In genes encoding for allozymes it might be an easier task, provided that the selected site is also responsible for the different

allozyme alleles, such as was demonstrated for *Adh* in *D. melanogaster* (Kreitman 1983). No such simple analysis is possible for the *Idgf* genes, as the exact function of these genes is still under investigation (P. J. Bryant, personal communication). However, in our data, within the 151-bp region under presumed balancing selection there is one candidate replacement polymorphism. It is site 564, which, linked to synonymous polymorphism at position 566, changes the valine (GTG) into isoleucine (ATA). Unlike the replacement polymorphism at the putatively catalytic site 516 (see above), this substitution is against the codon preference (Akashi 1995). Comparison with the crystal structure of IDGF2 can putatively localize the corresponding amino acid into the secondary α-helix outside the triose-phosphate isomerase barrel of the protein, but so far no specific function or activity has been described (Varela *et al.* 2002). Study of samples from different geographic origins, coupled with molecular genetics and biochemistry research of the *Idgf* loci, may help to resolve the evolutionary history of those genes.

We thank Montserrat Aguadé for providing DNA samples, Peter J. Bryant for sharing unpublished data, Joseph Comeron for the software for estimating recombination rate and valuable comments on the neutrality tests, Andrei N. Tatarenkov for insightful discussions and help regarding phylogenetic analysis and other matters, and two anonymous reviewers for helpful critical comments. This research was supported by National Institutes of Health grant no. GM-42397 to F.J.A.

LITERATURE CITED

Aguadé, M., 1998 Different forces drive the evolution of the *Acp26Aa* and *Acp26Ab* accessory gland genes in the *Drosophila melanogaster* species complex. Genetics **150:** 1079–1089.

Aguadé, M., 1999 Positive selection drives the evolution of the *Acp29AB* accessory gland protein in Drosophila. Genetics **152:** 543–551.

Akashi, H., 1995 Inferring weak selection from patterns of polymorphism and divergence as "silent" sites in Drosophila DNA. Genetics **139:** 1067–1976.

Andolfatto, P., and M. Nordborg, 1998 The effect of gene conversion on intralocus associations. Genetics **148:** 1397–1399.

Aquadro, C. F., V. Dumont and F. A. Reed, 2001 Genome-wide variation in the human and fruitfly: a comparison. Curr. Opin. Genet. Dev. **11:** 627–634.

Betran, E., J. Rozas, A. Navarro and A. Barbadilla, 1997 The estimation of the number and the length distribution of gene conversion tracts from population DNA sequence data. Genetics **146:** 89–99.

Comeron, J. M., M. Kreitman and M. Aguade, 1999 Natural selection on synonymous sites is correlated with gene length and recombination in Drosophila. Genetics **151:** 239–249.

Fu, Y.-X., and W.-H. Li, 1993 Statistical tests of neutrality of mutations. Genetics **133:** 693–709.

Gasperini, R., and G. Gibson, 1999 Absence of protein polymorphism in the *Ras* genes of *Drosophila melanogaster*. J. Mol. Evol. **49:** 583–590.

Harlow, E., and D. Lane, 1988 *Antibodies: A Laboratory Manual*. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Hudson, R. R., 1987 Estimating the recombination parameter of a finite population model without selection. Genet. Res. **50:** 245–250.

Hudson, R. R., and N. L. Kaplan, 1985 Statistical properties of the number of recombination events in the history of a sample of DNA sequences. Genetics **111:** 147–164.

Hudson, R. R., M. Kreitman and M. Aguade, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Kawamura, K., T. Shibata, O. Saget, D. Peel and P. J. Bryant, 1999 A new family of growth factors produced by the fat body and active on *Drosophila* imaginal disc cells. Development **126:** 211–219.

Kelly, J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146:** 1197–1206.

Kirby, D. A., and W. Stephan, 1995 Haplotype test reveals departure from neutrality in a segment of the *white* gene of *Drosophila melanogaster*. Genetics **141:** 1483–1490.

Kirkpatrick, R. B., R. E. Matico, D. E. McNulty, J. E. Strickler and M. Rosenberg, 1995 An abundantly secreted glycoprotein from *Drosophila melanogaster* is related to mammalian secretory proteins produced in rheumatoid tissues and by activated macrophages. Gene **153:** 147–154.

Kreitman, M., 1983 Nucleotide polymorphism at the alcohol dehydrogenase locus of *Drosophila melanogaster*. Nature **304:** 412–417.

Kreitman, M., and R. R. Hudson, 1991 Inferring the evolutionary histories of the *Adh* and *Adh-dup* loci in *Drosophila melanogaster* from patterns of polymorphism and divergence. Genetics **127:** 565–582.

Kriventseva, E. V., and M. S. Gelfand, 1999 Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. J. Biomol. Struct. Dyn. **17:** 281–288.

Kumar, S., and S. R. Gadagkar, 2001 Disparity index: a simple statistic to measure and test the homogeneity of substitution patterns between molecular sequences. Genetics **158:** 1321–1327.

Kumar, S., K. Tamura, I. B. Jakobsen and M. Nei, 2001 *MEGA2*: molecular evolutionary genetics analysis software. Bioinformatics **17:** 1244–1245.

Labate, J. A., C. H. Biermann and W. F. Eanes, 1999 Nucleotide variation at the *runt* locus in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **16:** 724–731.

McDonald, J. H., 1996 Detecting non-neutral heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **13:** 253–260.

McDonald, J. H., 1998 Improved tests for heterogeneity across a region of DNA sequence in the ratio of polymorphism to divergence. Mol. Biol. Evol. **15:** 377–384.

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

Moriyama, E. N., and J. R. Powell, 1996 Intraspecific nuclear DNA variation in *Drosophila*. Mol. Biol. Evol. **13:** 261–277.

Nei, M., 1987 *Molecular Evolutionary Genetics*. Columbia University Press, New York.

Powell, J. R., 1997 *Progress and Prospects in Evolutionary Biology: The Drosophila Model*. Oxford University Press, New York.

Przeworski, M., J. D. Wall and P. Andolfatto, 2001 Recombination and the frequency spectrum in *Drosophila melanogaster* and *Drosophila simulans*. Mol. Biol. Evol. **18:** 291–298.

Richter, B., M. Long, R. C. Lewontin and E. Nitasaka, 1997 Nucleotide variation and conservation at the *dpp* locus, a gene controlling early development in Drosophila. Genetics **145:** 311–323.

Rozas, J., and R. Rozas, 1999 DnaSP version 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15:** 174–175.

Swofford, D., 1998 *PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods)*. Sinauer, Sunderland, MA.

Tajima, F., 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics **123:** 585–595.

Varela, P. F., A. S. Llera, R. A. Mariuzza and J. Tormo, 2002 Crystal structure of imaginal disc growth factor-2. J. Biol. Chem. **277:** 13229–13236.

Walthour, C. S., and S. W. Schaeffer, 1994 Molecular population genetics of sex determination genes: the *transformer* gene of *Drosophila melanogaster*. Genetics **135:** 1367–1372.

Communicating editor: J. Hey