Universitat de Barcelona

Facultat de Biologia
Departament de Genètica

Evolución molecular de los genes del sistema olfatorio *OS-E* y *OS-F* en diferentes especies de Drosophila

Alejandro Sánchez-Gracia

Barcelona, Noviembre de 2005

3. Resultados

Capítulo 3.1. Patterns of nucleotide polymorphism and divergence in the odorant-binding protein genes *OS-E* and *OS-F*: analysis of the melanogaster species subgroup of Drosophila

3.1.1 Resumen

En este artículo se describe el análisis de la variabilidad nucleotídica en una región de unas 7kb que contiene los genes *OS-E* y *OS-F* en cuatro especies del subgrupo melanogaster de Drosophila, *Drososphila melanogaster*, *D. simulans*, *D. mauritiana* y *D. erecta*, así como en una población natural de *Drosophila melanogaster*. Los resultados indican que los genes *OS-E* y *OS-F* están presentes en las cuatro especies investigadas y mantienen su estructura génica. Las estimas de divergencia nucleotídica soportan la hipótesis de que ambos genes son funcionales, aunque presentan diferencias en su tasa evolutiva y en su constricción funcional. La variación nucleotídica en el gen *OS-E* presenta un patrón inusual: existe un exceso de sustituciones no sinónimas fijadas, y un pico de polimorfismo a lo largo de la región que incluye el gen. Los resultados se desvían significativamente del modelo neutro y sugieren la acción de la selección natural en la historia evolutiva de estos dos genes del sistema olfativo.

Patterns of Nucleotide Polymorphism and Divergence in the Odorant-Binding Protein Genes *OS-E* and *OS-F*: Analysis in the Melanogaster Species Subgroup of Drosophila

Alejandro Sánchez-Gracia, Montserrat Aguadé and Julio Rozas[1]

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

Manuscript received April 11, 2003
Accepted for publication July 25, 2003

ABSTRACT

The *Olfactory Specific-E* and *-F* genes (*OS-E* and *OS-F*) belong to the odorant-binding protein gene family, which includes the general odorant-binding proteins and the pheromone-binding proteins. In *Drosophila melanogaster*, these genes are arranged in tandem in a genomic region near the centromere of chromosome arm 3R. We examined the pattern of DNA sequence variation in an ∼7-kb genomic region encompassing the two *OS* genes in four species of the melanogaster subgroup of Drosophila and in a population sample of *D. melanogaster*. We found that both the *OS-E* and the *OS-F* gene are present in all surveyed species. Nucleotide divergence estimates would support that the two genes are functional, although they diverge in their functional constraint. The pattern of nucleotide variation in *D. melanogaster* also differed between genes. Variation in the *OS-E* gene region exhibited an unusual and distinctive pattern: (i) a relatively high number of fixed amino acid replacements in the encoded protein and (ii) a peak of nucleotide polymorphism around the *OS-E* gene. These results are unlikely under the neutral model and suggest the action of natural selection in the evolution of the two odorant-binding protein genes.

THE olfactory system of terrestrial animals has an extreme sensitivity and specificity. It can detect and discriminate a large number of olfactory signals, the odorants. Olfactory perception is accomplished by specialized bipolar sensory neurons that extend their dendrites into an aqueous medium: the olfactory mucus in vertebrates and the sensillar fluid in insects (Stern and Marx 1999). Hence, the airborne molecules must traverse the aqueous space that separates neuronal cells from the external air and stimulate the odorant receptors (Steinbrecht 1969, 1996). These receptors are located on the dendritic membrane of the sensory neurons (Buck and Axel 1991; Vosshall *et al.* 1999).

The odorant-binding proteins (OBPs) are abundant low-molecular-weight proteins that bind and solubilize hydrophobic odorants (or pheromones) in the vertebrate olfactory mucus and in the insect sensillar lymph. These small globular proteins are synthesized and secreted by some accessory cells surrounding the sensory neurons. In insects, the OBP family includes the general odorant-binding proteins (GOBPs) and the pheromone-binding proteins (PBPs), which are not homologous to vertebrate odorant-binding proteins (Vogt and Riddiford 1981; Pelosi and Maida 1995).

Despite the low sequence similarity among different insect OBPs, most of these proteins exhibit a similar distribution of conserved hydrophobic residues with a nearly identical predicted secondary structure. Most proteins of this family contain six highly conserved cysteines located in similar positions of the protein (Pikielny *et al.* 1994). In Lepidoptera, these cysteines are involved in disulfide bridges in both PBPs and GOBPs (Scaloni *et al.* 1999). The similar distribution of cysteine residues in both groups of OBPs suggests that the disulfide-bridge pairing might be a general feature of this family of molecules in insects.

Although the specific function of OBPs in olfaction is still unknown, they seem to play an important role in olfactory coding. It has been shown that several OBPs have different odorant specificities and are present in distinct subsets of antennal sensilla (Pelosi and Maida 1995). Additionally, genes encoding olfactory receptors with different binding specificities are also expressed in specific areas of the olfactory organ (Vosshall *et al.* 2000). These observations suggest that these proteins might participate in odor detection by restricting the spectrum of odorants accessible to the underlying receptors. In addition to the established functions of OBPs as carrier molecules and in concentrating hydrophobic odorants in the aqueous medium, it has also been proposed that these proteins could participate in the deactivation of the odorant stimulus (Pelosi and Maida 1995).

In *Drosophila melanogaster*, 51 putative members of the OBP family have been identified (Hekmat-Scafe *et al.* 2002; Vogt *et al.* 2002). Two of theses proteins, OS-E

Sequence data from this article have been deposited in the EMBL/GenBank Data Libraries under accession nos. AJ574644, AJ574762–AJ574774 (*D. melanogaster*), AJ563750 (*D. mauritiana*), AJ567753 (*D. simulans*), and AJ574775–AJ574776 (*D. erecta*).

[1] *Corresponding author:* Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Diagonal 645 08028, Barcelona, Spain. E-mail: jrozas@ub.edu

(olfactory-specific E) and OS-F (olfactory-specific F), colocalize at the same restricted area of the ventrolateral region of the antenna. The *OS-E* and *OS-F* genes (named *Obp83a* and *Obp83b* in Hekmat-Scafe *et al.* 2002), which have a similar intron-exon organization, are arranged in tandem in cytological band 83CD of the third chromosome. The two encoded proteins are highly conserved (72% amino acid identity in the mature protein; Pelosi and Maida 1995) except for a region in the C-terminal domain of the proteins, which has been named the heterogeneous region (*hr*; Hekmat-Scafe *et al.* 2000). The close physical proximity and the high degree of sequence similarity of the two coding regions seem to reflect a recent gene duplication event (McKenna *et al.* 1994; see, however, Hekmat-Scafe *et al.* 2000). Two hypotheses have been proposed to explain the low amino acid conservation of the *hr* region: (i) the *hr* region might form a putative binding site for the odorant molecule and (ii) the *hr* region might be a putative contact site for the olfactory receptor (Hekmat-Scafe *et al.* 2000).

Since olfaction is essential for survival and reproduction, genes involved in olfactory perception have likely evolved by the action of positive natural selection. Indeed, recognition and discrimination of olfactory signals are critical for finding food sources and for the reproduction of individuals; furthermore, certain chemoreceptive processes, like pheromone perception, contribute to critical evolutionary processes such as reproductive isolation and speciation. In fact, positive natural selection has been proposed to be involved in the evolution of PBPs of the moth Chortstoneura (Lepidoptera; Willett 2000) and also in the evolution of the OBPs of fire ants and other closely related species, in which these proteins could control some aspects of social organization (Krieger and Ross 2002). Here, we analyze DNA variation at the *OS-E* and *OS-F* genes in four species of the melanogaster subgroup of Drosophila (*D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. erecta*) and also in a natural population of *D. melanogaster* to infer the evolutionary history of these genes. We found that the two genes are present in all surveyed species and thus originated from an ancient duplication event; nevertheless, these genes differ in their functional constraint. We show that the *OS-E* gene region has very distinctive evolutionary patterns, specifically, (i) an accumulation of fixed amino acid replacements in the OS-E protein of *D. melanogaster* and (ii) an atypical pattern of nucleotide polymorphism. These results suggest that positive natural selection was likely involved in the evolution of this gene.

MATERIALS AND METHODS

Fly stocks: Fourteen *D. melanogaster* isochromosomal strains for the third chromosome were used; these strains were obtained from flies collected in a natural population of Montemayor, Spain, with crosses with the TM6/MKRS balancer stock (Cirera and Aguadé 1997; Ramos-Onsins and Aguadé 1998). A highly inbred *D. simulans* line (S40; from a natural population in Montblanc, Spain), obtained by 10 generations of sib mating, was also used (Rozas *et al.* 2001). Additionally, one line of each *D. mauritiana* and *D. erecta* kindly provided by F. Lemeunier were included in the present study.

DNA extraction, PCR amplification, and DNA sequencing: Genomic DNA from the *D. melanogaster* lines was CsCl purified (Bingham *et al.* 1981). DNA from *D. simulans*, *D. mauritiana*, and *D. erecta* was extracted from a single individual by using a modification of protocol 48 in Ashburner (1989). In *D. melanogaster*, an ∼4.7-kb genomic region that includes the complete coding region of both the *OS-E* and *OS-F* genes, the intergenic region, and 174 bp of the *OS-E* 5′ flanking region was amplified by PCR (Saiki *et al.* 1988; Figure 1). An additional 2-kb region upstream of the *OS-E* gene was PCR amplified in 13 of the *D. melanogaster* lines (in all lines except line M47). The amplified fragments were purified with Qiaquick columns (QIAGEN, Chatsworth, CA) and subsequently sequenced by using several oligonucleotides designed at intervals of ∼400 nucleotides. The sequenced fragments were separated on ABI PRISM 377 and 3700 automated DNA sequencers. For each line, the DNA sequence was determined on both strands.

For *D. simulans* and *D. mauritiana*, the same 4.7-kb region was amplified and sequenced by using several of the primers designed for *D. melanogaster* and, for the more divergent DNA regions, by primer walking. In *D. erecta*, only the *OS-E* and *OS-F* genes were PCR amplified and sequenced. In this species, primers for amplification and sequencing were designed on the most conserved regions of the genes among the other three species and also by the primer walking technique.

Data analysis: DNA sequences were assembled using the SeqEd version 1.0.3 program (Applied Biosystems, Foster City, CA). Sequences were multiply aligned with the ClustalW program (Thompson *et al.* 1994), and the initial alignment was optimized manually. The MacClade program, version 3.06 (Maddison and Maddison 1992), was used to edit the DNA sequences for further analyses. The secondary structure of the OS-E and OS-F proteins was inferred by using the PHD and PROF secondary structure prediction programs (Rost 2001). The DNA divergence among the studied species was estimated as *K*, the number of nucleotide differences per site corrected according to Jukes and Cantor (1969). Phylogenetic analysis was performed using the neighbor-joining algorithm (Saitou and Nei 1987) implemented in the MEGA version 2 program (Kumar *et al.* 2000) and by the maximum-likelihood method (Felsenstein 1993). The bootstrap analysis was based on 1000 replicates. Coding DNA sequences at internal nodes of the phylogenetic tree were reconstructed by the maximum-likelihood ancestral reconstruction approach using codon substitution models (Goldman and Yang 1994; Yang *et al.* 1995). From the ancestral sequences, we estimated the number of synonymous and nonsynonymous substitutions in each branch. All these analyses were performed using the *codeml* program included in the PAML 3.0 software (Yang 1997).

The DnaSP version 3.98 program (Rozas and Rozas 1999) was used for most intraspecific and some interspecific analyses. The level of DNA polymorphism was estimated as the per-site nucleotide diversity (π; Nei 1987), the Watterson parameter θ (Watterson 1975), and the haplotype diversity (Hd; Nei 1987). Codon bias was measured as the effective number of codons (ENC; Wright 1990), which measures the deviation from equal usage of synonymous codons.

The recombination parameter *c* (in Drosophila, $c = 2Nr$, where *N* is the effective population size and *r* is the recombination rate per generation between adjacent sites) was estimated using three different methods. The Hudson (1987) method estimates *c* from the variance of the average number of nucleo-

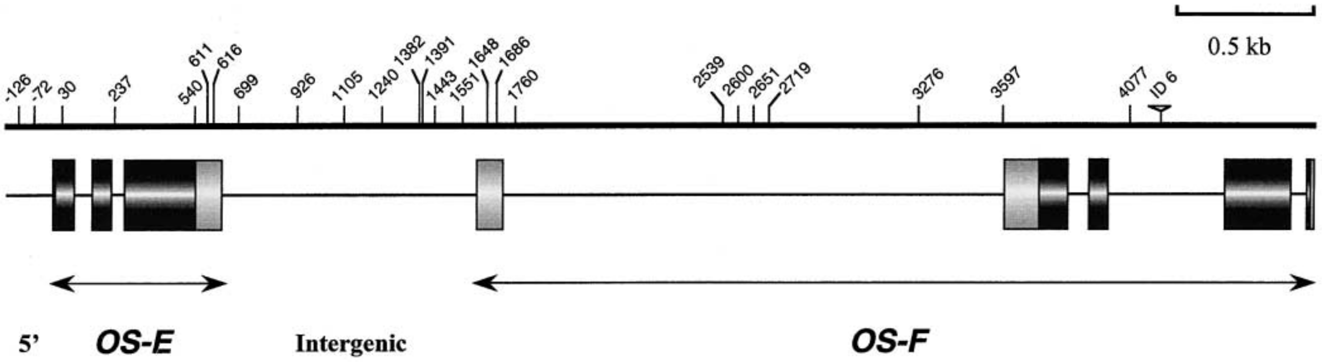


Figure 1.—Physical map of the *OS* region. Solid and shaded boxes indicate translated and untranslated exons, respectively. Numbers above the thick line indicate the position of the nucleotide polymorphisms detected in *D. melanogaster*. Nucleotide sites are numbered from the translation initiation site of the *OS-E* gene. The ~2-kb region upstream of the *OS-E* gene is not included in the figure.

tide differences between pairs of sequences, while the Hudson and Kaplan (1985) method estimates *c* from the minimum number of recombination events in the sample (R_M). The latter method requires the use of coalescent simulations to estimate *c*. An estimate of *c* based on the *D. melanogaster* recombination map was also obtained, assuming (i) that the recombination rate of the *OS* region (which is located on band 83CD) is 0.16×10^{-8} (see Comeron *et al.* 1999) and (ii) that *N* is 10^6 for *D. melanogaster*.

Statistical tests: The Tajima (1989) and the Fu and Li (1993) tests were used to contrast whether the polymorphism frequency distribution (frequency spectrum) conforms to neutral expectations. The overall genetic association between polymorphic sites was determined by the Z_{nS} (Kelly 1997), Wall's *B* and *Q* (Wall 1999), and Z_A (Rozas *et al.* 2001) statistics. The confidence intervals of these statistics were obtained by computer simulations (10,000 replicates) on the basis of the coalescent process assuming a large constant population size (Kingman 1982; Hudson 1983, 1990; Rozas and Rozas 1999). Coalescent simulations were conducted using different values of the recombination parameter and conditioning on the number of segregating sites.

The Hudson-Kreitman-Aguadé (HKA) test (Hudson *et al.* 1987) was conducted to assess whether the levels of polymorphism and divergence between species correlated, as expected under neutrality. The test was carried out using the 5′ *Adh* gene region (Kreitman and Aguadé 1986) as a neutral evolving region.

The Kolmogorov-Smirnov statistic (D_{KS}) was used to test for heterogeneity in the ratio of polymorphism to divergence along the surveyed DNA region. The test is based on the maximum absolute difference between the observed and the expected cumulative number of polymorphic sites (Sokal and Rohlf 1995; McDonald 1998), and it is generally the most powerful test for regions with two areas with very different levels of variation (McDonald 1998). The Wu and Li (1985) relative-rate test was used to test for heterogeneity in the nucleotide substitution rate among lineages. This method is based on the standardized difference of the corrected estimates of the number of substitutions per site between two lineages. The K2WULI program (Jermiin 1996) was used to perform this analysis. For nonsynonymous sites we used the χ^2 test due to the small number of substitutions.

RESULTS

Interspecific analysis: We have identified the *OS-E* and *OS-F* genes in the four species studied (*D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. erecta*). Moreover, both the intron-exon structure and the physical distance between genes are maintained across these species. Our results contrast with the previous report of Hekmat-Scafe *et al.* (2000), where the *OS-E* gene was not detected in either *D. simulans* or *D. mauritiana*. Nevertheless, the methodology used in their survey, a restriction-enzyme-based analysis, likely precluded its detection. Figure 2 shows the multiple alignment of the amino acid sequences encoded by the *OS-E* and *OS-F* genes. The six highly conserved cysteines of the OBP family are present in all OS proteins, except for the second cysteine of the OS-E protein in *D. erecta* that was replaced by a tryptophan. Moreover, the PHD and PROF programs predicted that all OS proteins are helical rich. To obtain clues on the function of specific parts of these proteins, the predicted structure of OS-E and OS-F was compared with that obtained for the pheromone-binding protein of *Bombix mori* (BmPBP). This protein is also a member of the OBP family and its three-dimensional (3D) structure has been determined by X-ray crystallography (Sandler *et al.* 2000). The distribution of the predicted α-helices along the OS proteins (Figure 2) is nearly identical to that found for the BmPBP.

Nucleotide divergence between species was estimated using only the DNA sequence fragment clearly alignable among all species (Table 1); the *D. melanogaster* line M2 was used for this analysis. In general, nucleotide divergence was higher in the *OS-E* than in the *OS-F* region. Despite this difference, both genes have a very similar and quite low codon bias, with an average ENC value for all species equal to 50.55 for *OS-E* and to 45.28 for *OS-F*. In both genes, higher K_S than K_A values were detected. In the *OS-E* gene, divergence estimates were higher at synonymous than at noncoding sites.

Figure 3 shows the neighbor-joining trees reconstructed for the *OS-E* and *OS-F* genes (the same topology is obtained using the maximum-likelihood approach). In the *OS-E* tree, the branch leading to the *D. melanogaster* lineage was rather long. We conducted a relative-rate test (Wu and Li 1985), using *D. erecta* as the outgroup,

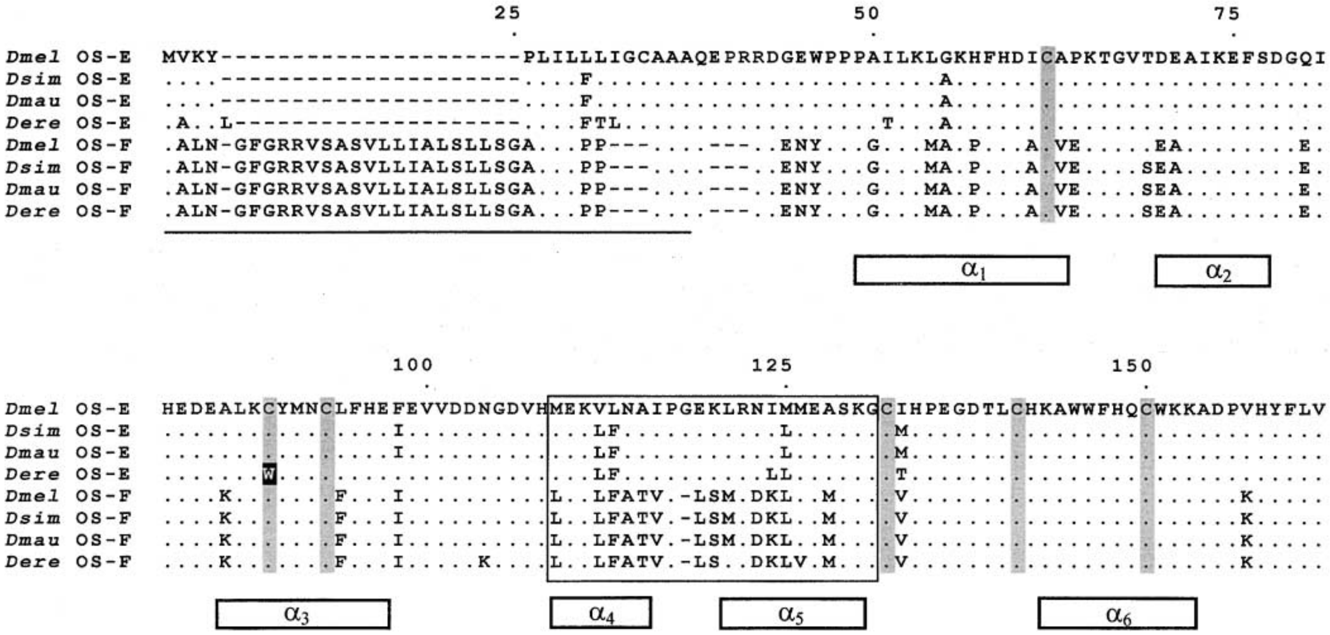

25
50
75
Dmel OS-E MVKY---------------------PLILLLIGCAAAQEPRRDGEWPPPAILKLGKHFHDICAPKTGVTDEAIKEFSDGQI
Dsim OS-E ....---------------------....F......................A..........................
Dmau OS-E ....---------------------....F......................A..........................
Dere OS-E .A..L--------------------....FTL...............T..A..........................
Dmel OS-F .ALN-GFGRRVSASVLLIALSLLSGA...PP---....----..ENY...G...MA.P...A.VE.....EA........E.
Dsim OS-F .ALN-GFGRRVSASVLLIALSLLSGA...PP---....----..ENY...G...MA.P...A.VE....SEA........E.
Dmau OS-F .ALN-GFGRRVSASVLLIALSLLSGA...PP---....----..ENY...G...MA.P...A.VE....SEA........E.
Dere OS-F .ALN-GFGRRVSASVLLIALSLLSGA...PP---....----..ENY...G...MA.P...A.VE....SEA........E.
α1
α2
100
125
150
Dmel OS-E HEDEALKCYMNCLFHEFEVVDDNGDVHMEKVLNAIPGEKLRNIMMEASKGCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV
Dsim OS-E ...............I..........LF..........L......M..............................
Dmau OS-E ...............I..........LF..........L......M..............................
Dere OS-E .......W..................LF..........LL.....T..............................
Dmel OS-F ....K.......F...I.........L..LFATV.-LSM.DKL..M...V.....................K.....
Dsim OS-F ....K.......F...I.........L..LFATV.-LSM.DKL..M...V.....................K.....
Dmau OS-F ....K.......F...I.........L..LFATV.-LSM.DKL..M...V.....................K.....
Dere OS-F ....K.......F...I.....K...L..LFATV.-LS..DKLV.M...V.....................K.....
α3
α4
α5
α6

Figure 2.—Amino acid sequence alignment of the OS-E and OS-F proteins. Amino acids identical to the first sequence are indicated by a dot. Shaded residues indicate the six cysteines highly conserved in the OBP family. The Cys-to-Trp replacement (OS-E protein of *D. erecta*) is also indicated. The line underneath the aligned sequences indicates the predicted signal peptide. The location of the predicted α-helices is indicated by open boxes below the sequences. The box in the sequences delimits the *hr* region (see the Introduction). *Dmel*, *D. melanogaster*; *Dsim*, *D. simulans*; *Dmau*, *D. mauritiana*; *Dere*, *D. erecta*.

to determine whether the numbers of substitutions in the *D. melanogaster* and in the *D. mauritiana* (or *D. simulans*) lineages were significantly different. This test revealed that the *OS-E* region (including coding and noncoding sites) evolves faster in the *D. melanogaster* than in the *D. mauritiana* and *D. simulans* lineages ($z = 2.052$, $P = 0.021$ for *D. mauritiana*; $z = 1.395$, $P = 0.081$ for *D. simulans*). Equivalent z results were obtained when only the coding region was used ($z = 2.006$, $P = 0.024$ for *D. mauritiana*; $z = 1.994$, $P = 0.023$ for *D. simulans*). In fact, the significantly higher number of substitutions accumulated in the *D. melanogaster* lineage was mainly due to nonsynonymous substitutions ($P = 0.058$). All amino acid replacements fixed in the *D. melanogaster* lineage are conservative, *i.e.*, with very low physicochemical distance values (Grantham 1974).

Nucleotide polymorphism in *D. melanogaster*: Figures 1 and 4 show the distribution of DNA polymorphic sites along the 4.7-kb region surveyed. A total of 25 nucleotide polymorphisms (9 of them with singleton variants) and 1 indel polymorphism (6 bp) were detected. All polymorphisms were silent: 1 synonymous polymorphism at site 30 of the *OS-E* coding region and the rest at noncoding positions. Notably, polymorphism at site 540 results in two different stop codons (TAG and TAA) of the *OS-E* gene. Ten different haplotypes (Hd = 0.956) were detected in the 14 lines analyzed.

Estimates of the per-site recombination parameter

TABLE 1

Nucleotide divergence in the *OS* region

Species pair	*OS-E*			*OS-F*		
	K_{NC}	K_S	K_A	K_{NC}	K_S	K_A
Dmel Dsim	0.0871	0.1102	0.0212	0.0603	0.0725	0.0028
Dmel Dmau	0.0682	0.1102	0.0212	0.0580	0.0510	0.0028
Dsim Dmau	0.0448	0.0347	0	0.0288	0.0200	0
Dmel Dere	0.2314	0.3831	0.0446	0.1359	0.1164	0.0112
Dsim Dere	0.2242	0.3514	0.0274	0.1507	0.1051	0.0084
Dmau Dere	0.2088	0.3519	0.0274	0.1367	0.0828	0.0084

K_{NC}, number of noncoding substitutions per noncoding site; K_S, number of synonymous substitutions per synonymous site; K_A, number of nonsynonymous substitutions per nonsynonymous site. *Dmel*, *D. melanogaster*; *Dsim*, *D. simulans*; *Dmau*, *D. mauritiana*; *Dere*, *D.erecta*.

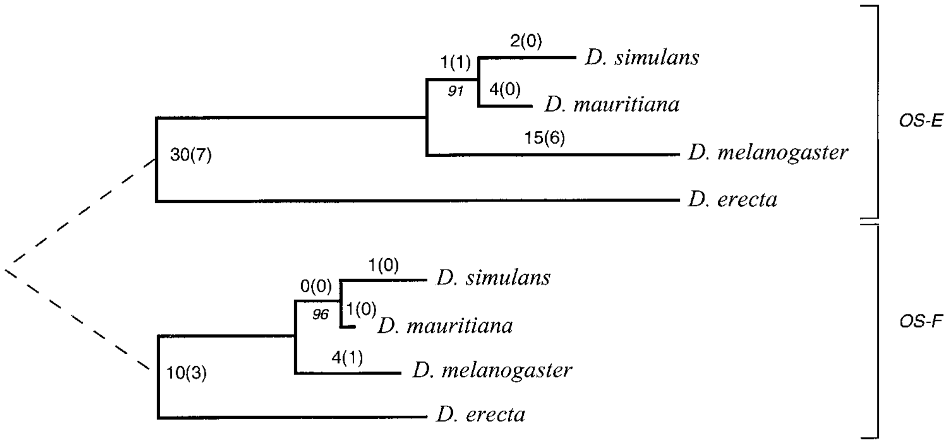


FIGURE 3.—Phylogenetic trees of the *OS-E* and *OS-F* genes. The trees were built using 651 sites (423 coding sites) and 805 sites (462 coding sites) for *OS-E* and *OS-F*, respectively. The numbers of nucleotide substitutions and the number of nonsynonymous changes (in parentheses) are indicated on each branch. The percentages of bootstrap replicates supporting the different nodes are in italics.

[c = 0.0026 from HUDSON (1987), c = 0.0024 from HUDSON and KAPLAN (1985), and c = 0.0032 from comparison of the physical and recombination maps] clearly indicate that the *OS* genomic region shows a reduction from the normal recombination levels (COMERON *et al.* 1999; ANDOLFATTO and PRZEWORSKI 2000). Table 2 summarizes the levels of nucleotide variation estimated separately for the different functional parts of the *OS* region. Estimates of nucleotide diversity for the complete 4.7-kb region (π = 0.0018, silent π = 0.0021) were rather low. However, levels of nucleotide diversity varied considerably along the *OS* region. Silent variation was highest in the *OS-E* gene (π = 0.0081) and lowest in the *OS-F* gene (π = 0.0013). Putative heterogeneity in the distribution of polymorphic to fixed silent sites along the *OS* region was tested by means of the D_{KS} test. Significant heterogeneity (P = 0.03) along the region studied was detected using the most conservative c value (see MCDONALD 1998). We also compared, by means of an HKA test, silent polymorphism and divergence in the total *OS* region (as well as separately for each functional part) and in the 5′ *Adh* region (HUDSON *et al.* 1987). A significant deviation from the neutral prediction was detected for the total *OS* region (P = 0.027). Interestingly, only variation in the *OS-F* region departed significantly from neutral predictions (*OS-E* region, P = 0.276; *OS-F* region, P = 0.017; intergenic region, P = 0.231).

We also conducted Tajima's D and Fu and Li's D and F tests separately for the three functional regions (*OS-E*, intergenic, and *OS-F*) to determine whether the frequency distribution of nucleotide variants departs from that expected under neutrality. For this analysis we used *D. mauritiana* as the outgroup. Under no recombination (which is the most conservative assumption for these tests), the analysis showed a significant deviation only in the *OS-E* region (Table 3). The significantly positive values of the Tajima's D and Fu and Li's F statistics in

Upstream

	-1995	-1959	-1940	-1756	-1755	-1754	-1690	-1163; d1	-1550	-1507	-1245	-1096	-987	-763	-756	-680	-445	-322	-310	-277
M11	G	C	G	C	T	G	A	d	G	C	G	G	T	T	G	C	A	C	G	T
M59	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	.	.	.	.
M2	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	.	.	.	.
M55	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	.	.	.	.
M23	.	.	.	.	.	.	.	d	.	.	.	.	.	A	.	.	.	.	.	.
M54	.	.	.	.	.	.	.	T	.	.	A	A	G	.	.	.	.	.	.	.
M26	.	.	.	.	.	.	.	T	.	.	.	.	.	A	.	.	.	.	.	.
M36	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.
M66	.	.	.	.	.	.	.	T	.	G	.	.	.	A	.	A	.	.	.	.
M13	T	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	T	.
M40	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.
M47																				
M22	T	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	G
M20	.	G	A	G	A	A	G	T	T	.	.	.	.	.	A	.	.	.	T	.
Dsim/Dmau	.	C/T	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	.

5′ *OS-E* *Intergenic* *OS-F*

	-126	-72	30; E1	237	540	611	616	699	926	1105	1240	1382	1391	1443	1551	1648	1686	1760	2539	2600	2651	2719	3276	3597	4077	4189; d6
M11	A	T	G	A	G	T	G	A	T	C	C	C	A	A	G	T	C	T	A	G	T	G	A	A	C	T
M59	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M2	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M55	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
M23	.	.	A	G	.	.	.	.	.	.	.	.	.	.	.	G	G	.	T	A	.	.	.	G	.	.
M54	.	.	A	G	.	.	.	.	.	.	.	.	.	.	A	G	.	.	T	A	C	.	.	G	.	.
M26	.	.	A	G	.	.	.	.	.	.	.	.	.	.	.	G	.	.	T	A	.	.	.	G	.	.
M36	.	.	A	G	.	.	.	.	.	.	.	.	.	.	A	G	.	.	T	A	C	.	.	G	.	.
M66	.	.	A	G	.	.	.	.	.	.	.	.	.	.	.	G	.	C	T	A	C	.	.	G	.	.
M13	.	.	.	.	A	G	T	G	A	.	.	.	.	.	.	G	.	C	.	.	.	.	.	G	.	d
M40	C	A	.	.	A	G	T	G	A	.	.	.	.	.	.	G	.	C	.	.	.	.	.	G	.	d
M47	C	.	.	.	A	G	T	G	A	.	.	.	.	.	.	G	.	C	.	.	.	.	.	G	.	.
M22	C	.	.	.	A	G	T	G	A	.	.	.	.	.	.	G	.	C	.	.	.	.	.	G	.	d
M20	C	.	.	.	A	G	T	G	A	T	.	T	T	G	.	G	.	C	.	.	.	A	T	G	T	.
Dsim/Dmau	.	.	.	.	A	d	T	.	C	.	.	T	.	G	.	G	.	C	T	d	.	.	.	G	.	d

FIGURE 4.—Nucleotide polymorphisms detected in the *OS* region of *D. melanogaster*. Nucleotides identical to the first sequence are indicated by a dot. For length polymorphisms, the nucleotide position refers to the first site affected. d, deletion; d#, deletion of #bp. The last row gives, for the polymorphic positions in *D. melanogaster*, the nucleotide information in *D. mauritiana* and *D. simulans*. E1, exon 1. Information for the additional analysis of the upstream region is also shown.

TABLE 2

Nucleotide variation in different *OS* functional regions

	Sites[a]	S	π	θ	K
5′ *OS-E*	161	2	0.0036	0.0039	0.0148
OS-E					
Coding	423	1	0.0012	0.0007	0.0397
Noncoding[b]	220	4	0.0090	0.0057	0.0793
Silent[b]	307.16	5	0.0081	0.0051	0.0899
Intergenic	929	8	0.0023	0.0027	0.0505
OS-F					
Coding	462	0	0	0	0.0131
Noncoding[b]	2493	10	0.0013	0.0013	0.0470
Silent[b]	2594.33	10	0.0013	0.0012	0.0471
Total silent	3991.5	25	0.0021	0.0019	0.0498
Total	4688	25	0.0018	0.0017	0.0438

S, number of segregating sites; *K*, number of substitutions per site between *D. melanogaster* and *D. mauritiana*.

[a] Number of sites in the intraspecific data set.

[b] Including intronic and untranslated regions.

TABLE 3

Neutrality tests

	OS-E gene region	Intergenic region	*OS-F* gene region
Tajima's *D*	1.982*	0.641	0.229
Tajima's *D*[a]	1.768*	0.522	1.251
Fu and Li's *D*	1.195	0.130	−0.633
Fu and Li's *F*	1.628*	−0.227	−0.540
Z_A	0.827*	0.453	0.309
Wall's *B*	0.750*	0.428	0.222
Wall's *Q*	1.000***	0.625	0.400

*$0.01 < P < 0.05$; ***$P < 0.001$.

[a] Excluding line M20.

this region reflect an excess of nucleotide variants at intermediate frequencies. For the *OS-F* and intergenic regions, the test statistic values substantially increased when line M20 (which accounts for seven of the nine singletons found in the sample) was removed.

No overall significant association between polymorphic sites (linkage disequilibrium) was detected by the Z_{nS} statistic either in the whole region ($Z_{nS} = 0.257$; $P = 0.541$) or in its different functional parts (results not shown). Nevertheless, a significant association was detected between polymorphic sites in the *OS-E* region (4 of the 10 pairwise comparisons were significant even after the conservative Bonferroni correction). The *OS-E* region also showed significant values of the Z_A and Wall's *B* and *Q* statistics even using the conservative no-recombination assumption (Table 3). These results indicate that nucleotide variation at the *OS-E* region is highly structured (Figure 4).

DISCUSSION

The presence of both the *OS-E* and *OS-F* genes in the four Drosophila species studied, and also in species of the obscura group (A. Sánchez-Gracia, M. Aguadé and J. Rozas, unpublished results), indicates that the DNA duplication event is relatively old. Nucleotide divergence estimates among copies and among species and the phylogenetic trees clearly indicate that the two genes have evolved independently since their origin by gene duplication (*i.e.*, there is no evidence for gene conversion between paralogous copies).

Our analysis has revealed a higher number of synonymous than nonsynonymous substitutions in all phylogeny branches, suggesting that both proteins are under purifying selection. The strength of natural selection, however, differs in the two genes. Indeed, the lower nucleotide substitution rates of the *OS-F* gene indicate an overall higher functional constraint (Tables 1 and 2; Figure 3) and, therefore, would support that the idea that these genes have been functionally diverging since their origin. The detection of the two OS proteins in the *D. melanogaster* sensillar lymph (McKenna *et al.* 1994) also supports the active action of natural selection in the evolution of these genes.

It could be argued, nevertheless, that the different *OS* substitution rates were caused by local mutation rate differences and not by differential functional constraint. We found that divergence estimates at noncoding positions in the *OS-F* region are slightly lower than those in *OS-E* (Table 2). This fact is likely caused by the presence of an extremely conserved DNA fragment in the second intron of *OS-F*; this conserved region had been already identified in other species as distant as *D. virilis* (Hekmat-Scafe *et al.* 2000), although its functional significance is unknown. Levels of nucleotide divergence across noncoding regions, nevertheless, are distributed more homogeneously than in coding regions. This analysis does not support, therefore, that the different evolutionary rates found between the two *OS* genes were caused by putative differences in the silent mutation rate along the *OS* region.

Most *D. melanogaster* OBP family members (in addition to *OS-E* and *OS-F*) are located in gene clusters (Galindo and Smith 2001). This suggests that gene duplication is an important mechanism to increase diversity in this gene family. In fact, the high sequence divergence among functional members of the family suggests the contribution of positive selection to the rapid evolution and functional diversification of these genes (Galindo and Smith 2001). It has been also shown that insect OBPs can form dimers in physiological conditions

(Campanacci *et al.* 1999; Danty *et al.* 1999; Sandler *et al.* 2000). Since the *OS-E* and *OS-F* genes are coexpressed in the same cells, the encoded proteins would be able to form homodimers and also heterodimers. Although the formation of such heterodimers has not been demonstrated, this possibility is suggestive since it might be involved in the evolution of these genes: if heterodimers were more efficient than homodimers, selection might have favored the differentiation and maintenance of the two genes. Certainly, information on the quaternary structure of these OS proteins would be relevant to ascertain the role of putative dimers on the molecular evolution of these genes.

We also found an excess of substitutions at the *OS-E* coding region in the *D. melanogaster* lineage. This excess, largely due to a high number of nonsynonymous changes, could be explained by either a relaxation of natural selection or the action of positive directional selection. Although a reduction in the selective pressure could increase the fixation probability of weakly selected mutations (Ohta 1973, 1992), the reduced levels of nucleotide polymorphism in the *OS-E* coding region would not support this hypothesis. Hence, it seems likely that positive directional selection would have driven these amino acid changes to fixation (see below).

Currently, the 3D structure of the OS-E and OS-F proteins has not been determined. However, two lines of evidence suggest that this structure could be similar to that of the BmPBP obtained by X-ray crystallography. First, the secondary structures predicted for OS-E and OS-F show a remarkable similarity to that previously predicted for the BmPBP, in which the predicted location of α-helices has been confirmed by the 3D structure. Second, there are five highly conserved phenylalanines in BmPBP, with two of them (Phe12 and Phe118) involved in the general (*i.e.*, not specific) binding hydrophobic surface (Sandler *et al.* 2000). In the OS-E and OS-F proteins, there are also five conserved phenylalanines in all surveyed species. Two of these residues (Phe18 and Phe108) are also found in similar positions on the predicted BmPBP hydrophobic surface. Probably the residues constituting the odorant-binding pocket and those involved in the specific binding site of the OS-E and OS-F proteins are in locations equivalent to those described in BmPBP.

A preliminary analysis of OBPs (Plettner *et al.* 2000; Sandler *et al.* 2000; Peng and Leal 2001) has shown that some conservative amino acid changes observed across a number of Lepidoptera species might alter the protein-binding specificity. In particular, replacements among residues such as valine, leucine, isoleucine, or methionine would be responsible for this change in specificity. Remarkably, all amino acid replacements found in the four Drosophila species studied (except the Cys-to-Trp change in *D. erecta*; Figure 2) are conservative, with low physicochemical distance values (Grantham 1974), and are found in protein locations similar to the changes observed in Lepidoptera OBPs. Furthermore, four of the six OS-E amino acid changes fixed in the *D. melanogaster* lineage also involve valine, leucine, isoleucine, and methionine residues. These replacements are located either in the heterogeneous region (Hekmat-Scafe *et al.* 2000) or close to it. Some of these replacements might have been beneficial due to plausible changes in the binding specificity of the protein. Directional positive selection would thus have driven them to fixation.

The action of positive selection should have also left a fingerprint on intraspecific polymorphism and on the ratio of polymorphism to divergence. We have shown that levels of silent nucleotide polymorphism in *D. melanogaster* were reduced, which is consistent with expectations for a low-recombining genomic region (Begun and Aquadro 1992). However, the level of silent nucleotide polymorphism clearly differs between gene copies (Table 2). In fact, the observed number of segregating sites and the results of the Kolmogorof-Smirnov test ($P = 0.03$) clearly define two regions with distinct levels of variation: a left part including the *OS-E* and intergenic regions and a right part that includes the *OS-F* gene region. While the *OS-F* portion has a low level of variation, as expected in regions of reduced recombination, the level of nucleotide variation at the left part is concordant with the average level of silent variation in *D. melanogaster* (∼0.011; Moriyama and Powell 1996). The analysis of the ratio of polymorphism to divergence also shows a clear drop in about the middle of the surveyed region. The results of the HKA test are significant only in the *OS-F* region; in this part, silent nucleotide polymorphism is likely reduced due to the effects of linked selection in concordance with its low recombining environment. In contrast, no significant HKA results were obtained at the left part of the *OS* region, reflecting a much higher level of intraspecific variation than expected in a region of low recombination. In particular, there was a local peak of variation, with most polymorphisms at noncoding regions (Figure 5).

To know whether the increase of variation in *OS-E* is really a peak of variation (*i.e.*, whether it decays also in the upstream region), we sequenced an ∼2-kb region upstream of the *OS-E* gene. The level of variation in this 5′ flanking region ($\pi = 0.002$) was similar to that detected in the *OS-F* region and therefore lower than that in the *OS-E* region. The sliding window analysis of the entire region surveyed (6.7 kb) reveals a peak of variation in the *OS-E* region (Figure 5). Clearly, these results are unlikely not only under the neutral model, but also under simplistic selective models [such as the genetic hitchhiking (Maynard Smith and Haigh 1974) or the background selection (Charlesworth *et al.* 1993) models], or demographic scenarios. Heterogeneity in the recombination rate or in the silent mutation rate along the surveyed region could explain the different pattern of variation observed in the *OS-E* and *OS-F*

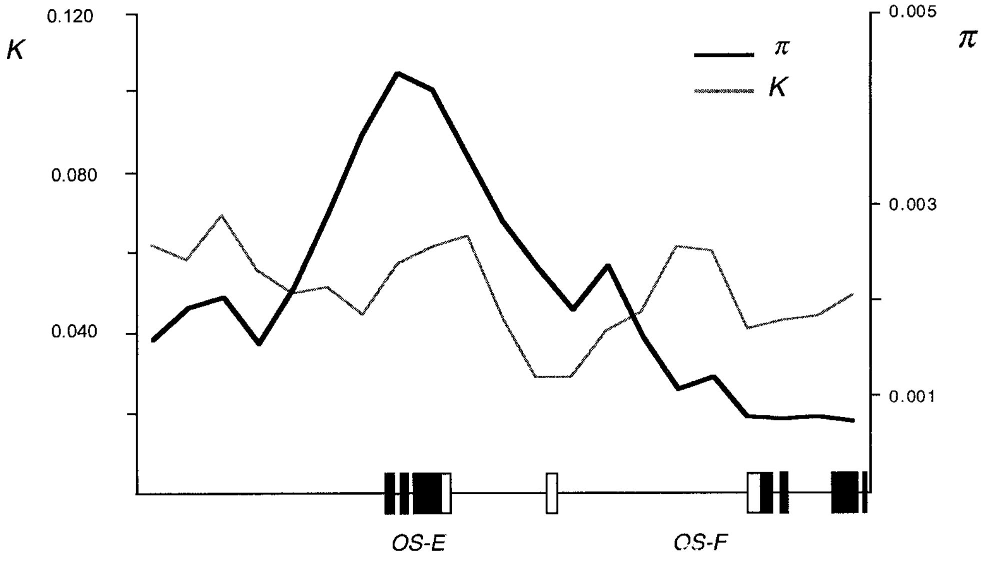


Figure 5.—Sliding window of silent polymorphism in *D. melanogaster* (π) and silent divergence between *D. melanogaster* and *D. mauritiana* (*K*) for the complete *OS* region (including the additional upstream region). Step size, 250 silent sites; window size, 1000 silent sites.

regions. However, it is unrealistic that two closely located genes could differ strongly in their recombination rate. On the other hand, the similar estimates of the silent nucleotide divergence along the *OS* region also do not support the silent mutation rate heterogeneity hypothesis.

The anomalous levels of nucleotide variation found in *OS-E* could be explained by the action of some form of balancing selection. Nevertheless, it seems difficult to envisage the target of selection since there is no replacement polymorphism. Yet, selection might act on the RNA stability or on some regulatory elements present at noncoding positions. Furthermore, several other features of the data are consistent with the balancing selection hypothesis. This kind of selection is expected to increase the levels of nucleotide variation and consequently it can skew the frequency spectrum toward intermediate frequencies. The significantly positive Tajima's *D* and Fu and Li's *F* values (Table 3) observed in the *OS-E* region are in agreement with this prediction. Nevertheless, a simple balancing selection model would not easily explain the high number of amino acid changes accumulated in the *D. melanogaster* lineage. A version of the hitchhiking model, the "traffic" model (Kirby and Stephan 1996), might account for both the positive Tajima's *D* values observed in the *OS-E* gene and the increase of nucleotide diversity in the target region. This model considers that the fixation of a favorable mutation might be retarded by the fixation process of another closely located favorable mutation. Under this model, neutral variants could increase in frequency near the selected sites, reaching intermediate frequencies. Eventually, recombination could generate haplotypes with a more favorable combination of mutations that would be driven to fixation. The high structure of genetic variation in the *OS-E* region revealed by the significant linkage disequilibrium values and the Z_A and Wall's *B* and *Q* tests is consistent with the traffic hypothesis. In conclusion, the pattern of nucleotide sequence variation in the *OS* genes is unlikely under the neutral model of molecular evolution and suggests the action of positive natural selection. Further surveys of variation at genes of the olfactory family might contribute to establishing which specific mode of natural selection is acting and thereby to an understanding of the evolutionary meaning and fate of these duplicated genes.

We thank D. Hekmat-Scafe and colleagues for sending us a copy of their manuscript before publication and Serveis Científico-Tècnics, Universitat de Barcelona, for the use of automated sequencing facilities. A.S. was a predoctoral fellow of Universitat de Barcelona. This work was supported by grants PB97-0918 and BMC2001-2906 from Comisión Interdepartamental de Ciencia y Tecnología, Spain, and by grant 2001SGR-00101 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Spain, to M.A.

LITERATURE CITED

Andolfatto, P., and M. Przeworski, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics **156:** 257–268.

Ashburner, M., 1989 *Drosophila: A Laboratory Handbook.* Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Begun, D. J., and C. F. Aquadro, 1992 Levels of naturally occurring DNA polymorphism correlate with recombination rates in *Drosophila.* Nature **356:** 519–520.

Bingham, P. M., R. Levis and G. M. Rubin, 1981 Cloning of DNA sequences from the white locus of *D. melanogaster* by a novel and general method. Cell **25:** 693–704.

Buck, L., and R. Axel, 1991 A novel multigene family may encode odorant receptors: a molecular basis for odor recognition. Cell **65:** 175–187.

Campanacci, V., S. Longhi, P. Nagnan-Le Meillour, C. Cambillau and M. Tegoni, 1999 Recombinant pheromone binding protein 1 from *Mamestra brassicae* (MbraPBP1). Eur. J. Biochem. **264:** 707–716.

Charlesworth, B., M. T. Morgan and D. Charlesworth, 1993 The effect of deleterious mutation on neutral molecular variation. Genetics **134:** 1289–1303.

Cirera, S., and M. Aguadé, 1997 Evolutionary history of the sex-peptide (Acp70A) gene region in *Drosophila melanogaster*. Genetics **147:** 189–197.

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

Danty, E., L. Briand, C. Michard-Vanhee, V. Perez, G. Arnold *et al.*, 1999 Cloning and expression of a queen pheromone-binding protein in the honeybee: an olfactory-specific, developmentally regulated protein. J. Neurosci. **17:** 7468–7475.

Felsenstein, J., 1993 *Phylogenetic Inference Package (PHYLIP)*, Version 3.5. University of Washington, Seattle.

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

Galindo, K., and D. P. Smith, 2001 A large family of divergent Drosophila odorant-binding proteins expressed in gustatory and olfactory sensilla. Genetics **159:** 1059–1072.

Goldman, N., and Z. Yang, 1994 A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol. Biol. Evol. **11:** 725–736.

Grantham, R., 1974 Amino acid difference formula to help explain protein evolution. Science **185:** 862–864.

Hekmat-Scafe, D. S., R. L. Dorit and J. R. Carlson, 2000 Molecular evolution of odorant-binding protein genes *OS-E* and *OS-F* in Drosophila. Genetics **155:** 117–127.

Hekmat-Scafe, D. S., C. R. Scafe, A. J. Mckinney and M. A. Tanouye, 2002 Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. Genome Res. **9:** 1357–1369.

Hudson, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor. Popul. Biol. **23:** 183–201.

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

Hudson, R. R., 1990 Gene genealogies and the coalescent process, pp. 1–42 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. Futuyma and J. Antonovics. Oxford University Press, New York.

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. Aguadé, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics **116:** 153–159.

Jermiin, L. S., 1996 *K2WuLi*, Version 1.0. Australian National University, Canberra, Australia.

Jukes, T. H., and C. R. Cantor, 1969 *Mammalian Protein Metabolism*. Academic Press, New York.

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

Kingman, J. F. C., 1982 On the genealogy of large populations. J. Appl. Probab. **19A:** 27–43.

Kirby, D. A., and W. Stephan, 1996 Multi-locus selection and the structure of the white gene of *Drosophila melanogaster*. Genetics **144:** 635–645.

Kreitman, M., and M. Aguadé, 1986 Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide recognizing enzyme digests. Proc. Natl. Acad. Sci. USA **83:** 3562–3566.

Krieger, M. J., and K. G. Ross, 2002 Identification of a major gene regulating complex social behavior. Science **295:** 328–332.

Kumar, S., K. Tamura, I. Jakobsen and M. Nei, 2000 *MEGA, Molecular Evolutionary Genetics Analysis*, Version 2.0.

Maddison, W. P., and D. R. Maddison, 1992 *MacClade: Analysis of Phylogeny and Character Evolution*, Version 3.0. Sinauer Associates, Sunderland, MA.

Maynard Smith, J., and J. Haigh, 1974 The hitch-hiking effect of a favorable gene. Genet. Res. **23:** 23–35.

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.

McKenna, M. P., D. S. Hekmat-Scafe, P. Gaines and J. R. Carlson, 1994 Putative *Drosophila* pheromone-binding proteins expressed in a subregion of the olfactory system. J. Biol. Chem. **269:** 16340–16347.

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.

Ohta, T., 1973 Slightly deleterious substitutions in evolution. Nature **246:** 96–98.

Ohta, T., 1992 The nearly neutral theory of molecular evolution. Annu. Rev. Ecol. Syst. **23:** 263–286.

Pelosi, P., and R. Maida, 1995 Odorant-binding proteins in insects. Comp. Biochem. Physiol. B **111:** 503–514.

Peng, G. H., and W. S. Leal, 2001 Identification and cloning of a pheromone-binding protein from the oriental beetle, *Exomala orientalis*. J. Chem. Ecol. **27:** 2183–2192.

Pikielny, C. W., G. Hasan, F. Rouyer and M. Rosbash, 1994 Members of a family of *Drosophila* putative odorant-binding proteins are expressed in different subsets of olfactory hairs. Neuron **12:** 35–49.

Plettner, E., J. Lazar, E. G. Prestwich and G. D. Prestwich, 2000 Discrimination of pheromone enantiomers by two pheromone binding proteins from the gypsy moth *Lymantria dispar*. Biochemistry **39:** 8953–8962.

Ramos-Onsins, S., and M. Aguadé, 1998 Molecular evolution of the *Cecropin* multigene family in Drosophila: functional genes *vs.* pseudogenes. Genetics **150:** 157–171.

Rost, B., 2001 Protein secondary structure prediction continues to rise. J. Struct. Biol. **134:** 204–218.

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

Rozas, J., M. Gullaud, G. Blandin and M. Aguadé, 2001 DNA variation at the *rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics **158:** 1147–1155.

Saiki, R. K., D. H. Gelfand, S. Stoffel, S. J. Scharf, R. Higuchi *et al.*, 1988 Primer-directed enzymatic amplification of DNA with a thermostable polymerase. Science **239:** 487–491.

Saitou, N., and M. Nei, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol. Biol. Evol. **4:** 406–425.

Sandler, B. H., L. Nikonova, W. S. Leal and J. Clardy, 2000 Sexual attraction in the silkworm moth: structure of the pheromone-binding-protein-bombykol complex. Chem. Biol. **7:** 143–151.

Scaloni, A., M. Monti, S. Angeli and P. Pelosi, 1999 Structural analysis and disulfide-bridge pairing of two odorant-binding proteins from *Bombyx mori*. Biochem. Biophys. Res. Commun. **266:** 386–391.

Sokal, R., and F. J. Rohlf, 1995 *Biometry*. W. H. Freeman, New York.

Steinbrecht, R. A., 1969 Comparative morphology of olfactory receptors, pp. 3–21 in *Olfaction and Taste III*, edited by C. Pfaffmann. Rockefeller University Press, New York.

Steinbrecht, R. A., 1996 Are odorant-binding proteins involved in odorant discrimination? Chem. Senses **21:** 719–727.

Stern, D., and J. Marx, 1999 Making sense of scents. Science **286:** 703–728.

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

Thompson, J. D., D. G. Higgins and T. J. Gibson, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Vogt, R. G., and L. M. Riddiford, 1981 Pheromone binding and inactivation by moth antennae. Nature **293:** 161–163.

Vogt, R. G., M. E. Rogers, M. D. Franco and M. Sun, 2002 A comparative study of odorant-binding protein genes: differential expression of the PBP1-GOBP2 gene cluster in *Manduca sexta* (Lepidoptera) and the organization of OBP genes in *Drosophila melanogaster* (Diptera). J. Exp. Biol. **205:** 719–744.

Vosshall, L. B., H. Amrein, P. S. Morozov, A. Rzhetsky and R. Axel, 1999 A spatial map of olfactory receptor expression in the *Drosophila* antenna. Cell **96:** 725–736.

Vosshall, L. B., A. M. Wong and R. Axel, 2000 An olfactory sensory map in the fly brain. Cell **102:** 147–159.

Wall, J., 1999 Recombination and the power of statistical test of neutrality. Genet. Res. **74:** 65–79.

Watterson, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor. Popul. Biol. **7:** 256–276.

Willett, C. S., 2000 Evidence for directional selection acting on pheromone-binding proteins in the genus *Choristoneura*. Mol. Biol. Evol. **17:** 553–562.

Wright, F., 1990 The effective number of codons used by a gene. Gene **87:** 23–29.

Wu, C.-I, and W.-H. Li, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82:** 1741–1745.

Yang, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **15:** 555–556.

Yang, Z., S. Kumar and M. Nei, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics **141:** 1641–1650.

Communicating editor: M. Veuille

Capítulo 3.2. Unusual pattern of nucleotide sequence variation at the *OS-E* and *OS-F* genomic region of *Drosophila simulans*

3.2.1 Resumen

En este capítulo se describe la caracterización, secuenciación y análisis de la variabilidad nucleotídica de la región genómica que incluye los genes del sistema olfativo *OS-E* y *OS-F* en una población europea y otra población africana de la especie *D. simulans*. En las dos poblaciones, los niveles de polimorfismo nucleotídico silencioso son mayores a los detectados en la región homóloga de *D. melanogaster*. Los resultados de los tests de neutralidad basados en la teoría de la coalescencia y utilizando la información intra- e interespecífica de *D. simulans* y *D. melanogaster*, confirman que en el linaje de *D. melanogater* el gen *OS-E* presenta un comportamiento no neutral. Por otro lado, el patrón del polimorfismo nucleotídico de la población europea de *D. simulans* presenta una estructura haplotípica inusual, donde la mitad de las secuencias son prácticamente idénticas (difiriendo en tan sólo una sustitución nucleotídica). Esta estructuración es incompatible con la predicción de la teoría neutralista para poblaciones en equilibrio estacionario. En el trabajo se discute el posible papel de la selección natural positiva, y de factores demográficos, en la generación de esta atípica estructuración de la variabilidad.

Artículo II

Unusual pattern of nucleotide sequence variation at the *OS-E* and *OS-F* gene region of *Drosophila simulans*

Alejandro Sánchez-Gracia and Julio Rozas

2005

(En preparación)

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona,

Barcelona, Spain

INTRODUCTION

Smell is one of the oldest and most important senses of animals. Olfaction allows recognizing and discriminating chemical signals providing essential information to detect and asses food, to identify mating partners and predators, and to establish individual and collective behavior. Positive natural selection can therefore play a major role in the evolution of olfactory system genes; indeed, human olfaction-involved-genes are amongst the most accelerated evolution genes (e.g. GILAD *et al*. 2003; CLARK *et al.* 2003; NIELSEN *et al*. 2005; see also GIMELBRANT *et al*. 2004). Darwinian positive selection, in addition, has been proposed for the evolution of some olfactory system genes in rodents (EMES *et al.* 2004), channel catfish (NGAI *et al*. 1993), salamander (WATTS *et al*. 2004; PALMER *et al.* 2005) and insects (WILLETT 2000; KRIEGER and ROSS 2002, 2005).

The olfactory system constitutes the principal sensory modality of invertebrates, showing a high specificity and sensitivity. Odorant receptors (OR) are located in the external membrane of specialized sensory neurons, extending their dendrites into an aqueous fluid. Hydrophobic odors traverse the fluid space bound to OBPs (Odorant-Binding Proteins), which deliver them close to receptors. The OBP multigene family includes two different families of proteins, the GOBPs (General Odorant-Binding Proteins) that binds and transport general odorants, and the PBP (Pheromone Binding Proteins) specialized in pheromone perception (VOGT and RIDDIFORD 1981; PELOSI and MAIDA 1995). These two families have a monophyletic origin. There is little knowledge about the evolution of Drosophila OBP multigene family. This family contains fifty-one putative members located in clusters and scattered across the genome (HEKMAT-SCAFE *et al*. 2002). Surprisingly, this number is very close to the actual number of the *OR* genes (about 61 members; VOSSHALL 2000). This fact, jointly with the different odorant binding specificities and gene expression patterns (e.g. VOGT *et al.* 1991, 1999; GALINDO and SMITH 2001; VOGT *et al*. 2002), suggest that OBPs have an important role in the olfactory coding and not merely odorant carriers.

We have studied the molecular evolution (at intraspecific and interspecific levels) of two members of the OBP gene family (the *OS-E* and *OS-F* genes) in different Drosophila species (SÁNCHEZ-GRACIA *et al*. 2003; see also HEKMAT-SCAFE *et al*. 2000). These genes likely originated by an old gene duplication event (likely >40 myr), and interestingly still maintain a high degree of conservation at the gene structure, amino acid, and nucleotide levels. In this study, we detected a significant gradient of the nucleotide polymorphism levels along the *OS*-region of *D. melanogaster*, and an excess of fixed amino acid replacements in the lineage of this species (in the *OS-E* gene). Although the results are unlikely for a neutral evolving region, we could not discriminate among the different selective scenarios that might accommodate the data.

Here, we analyze levels and patterns of DNA polymorphism and divergence along the *OS*-gene region in two populations of *D. simulans* to provide new insights on the evolution of these olfactory genes and, particularly, to assess if the evolutionary pattern observed in *D. melanogaster* is a species-specific feature or, on the contrary, it is shared by other Drosophila species. We found that levels of silent variation, and recombination rates, are significantly higher in *D. simulans* than in *D. melanogaster*. Additionally we also detected that the European population of *D. simulans* is highly structured, with a very unusual haplotype configuration. We discuss these findings with their implications on the molecular population genetics of *D. simulans*.

MATERIALS AND METHODS

Drosophila strains: Twenty-two highly inbreed *D. simulans* lines (obtained after 10 generations of sib mating) randomly sampled from two natural populations were surveyed: eleven lines from a European population (Montblanc, Spain; S lines), and eleven from an African sample (Maputo, Mozambique; MZ lines) (ROZAS *et al*. 2001). Present analysis also includes the fourteen lines of *D. melanogaster* (Córdoba, Spain; M lines), one line of *D. mauritiana* and another of *D. erecta* reported in SÁNCHEZ-GRACIA *et al*. (2003).

DNA extraction, PCR amplification, and DNA sequencing: Genomic DNA of *D. simulans* was extracted using a modification of protocol 48 from ASHBURNER (1989). For the European sample, a ~5kb fragment (named as fragment 1) including the *OS-E* and *OS-F* genes and their intergenic region, was amplified by PCR (SAIKI *et al.* 1988), while in the African lines, the amplified fragment (a ~2kb fragment named here as fragment 2) only included the *OS-E* transcription region, the intergenic region and the first untranslated exon of the *OS-E* gene. The amplified fragments were cycle-sequenced and separated on a Perkin-Elmer (Norwalk, CT) ABI PRISM 377 automated DNA sequencer, following the manufacturer's instructions. For each line, the DNA was sequenced on both strands. The new reported nucleotide sequences have been deposited in the EMBL nucleotide sequence database under accession nos. XXXXXX-XXXXXX.

Data analysis: Nucleotide sequences were assembled using the SeqEd 1.03 software (Applied Biosystems), multiple aligned with Clustal X (THOMPSON *et al.*

1997), and edited with the MacClade version 3.06 program (MADDISON and MADDISON 1992). Phylogenetic analysis was performed using the neighbour-joining algorithm (SAITOU and NEI 1987) implemented in the MEGA version 2 program (KUMAR *et al*. 2001). The bootstrap analysis was based on 1000 replicates. We estimated the number of synonymous and nonsynonymous substitutions in each branch of the tree by using the *codeml* program implemented in PAML package version 3.14 (YANG 1997).

The DnaSP version 4.0 program (ROZAS *et al.* 2003) was used for most intraspecific and some interspecific analyses. The level of DNA polymorphism was estimated as the per-site nucleotide diversity (π; NEI 1987), the Watterson parameter (θ; WATTERSON 1975), the haplotype diversity (h; NEI 1987) and the number of haplotypes (k, Nei 1987). Nucleotide divergence between species was estimated as K, the number of per site substitutions corrected according to JUKES and CANTOR (1969). The confidence intervals, and P-values, of a number of neutrality tests were obtained by Monte Carlo simulations based on the neutral coalescent process assuming the infinite-sites model in a large constant-size population (KINGMAN 1982a, 1982b; HUDSON 1983, 1990). Coalescent simulations were performed assuming either no intragenic recombination (10000 replicates) or with variable levels of recombination (2000 replicates) (HUDSON 1983; ROZAS *et al.* 2003). This information allows obtaining the empirical distribution of statistics that, in turn, was used to determine their confidence intervals. Computer simulations were carried out either fixing the value of θ ($\theta = 4N\upsilon$, where N is the effective population size and υ is the per-gene mutation rate), or the number of segregating sites. These two methods yielded similar results and thus we will present results based on only the coalescent conditional on the number of segregating sites.

The recombination parameter *C* (in Drosophila, $C = 2N_e r$, where *r* is the per-generation recombination rate for the studied region) was estimated using three different methods. The HUDSON (1987) method estimates *C* (C_H) from the variance of the average number of nucleotide differences. The HUDSON AND KAPLAN (1985) method estimates *C* (C_R) from the minimum number of recombination events in the sample (R_M) by using coalescent simulations. Estimates of *C* based on the *D. simulans* recombination map (C_M) (ANDOLFATTO and PRZEWORSKI 2000) was obtained assuming that $r = 1.04 \times 10^{-8}$ (i.e., assuming that the *OS* and *Gld* regions, which are located in chromosomal bands 83CD and 84D, respectively, have the same recombination rate), and that N_e is 2×10^6. We also used computer simulations to estimate the C_L value (ROZAS *et al.* 2001); that is, the minimum value of *C* compatible at the 5% with the observed R_M value. The effect of intragenic recombination on nucleotide variation was also analyzed using *ZZ* statistic (ROZAS *et al.* 2001).

The TAJIMA (1989), FU and LI (1993) and FU (1997) tests were conducted to contrast whether the DNA polymorphism pattern conforms to the neutral expectations. FAY and WU's (2000) test was used to asses for the presence of high frequency derived nucleotide variants in the sample. The correlation between polymorphism and divergence expected under the neutral model was tested using the HKA test (HUDSON *et al.* 1987). The putative genetic differentiation between populations was determined by a permutation test (1000 replicates) using the S_{nn} statistic HUDSON (2000). We calculated the age of the putative selective sweep assuming that all mutations detected in the sample were new mutations originated after the variation sweep caused by the hitchhiking effect.

RESULTS

European sample

DNA sequence variation: We surveyed initially a genomic region that included the *OS-E* and *OS-F* genes with their intergenic region (4896 bp, fragment 1; figure 1) in 11 European lines of *D. simulans*. A total of 96 polymorphic nucleotide sites (consisting in 99 mutation events), and 12 insertion/deletion polymorphisms (ranking from 1 to 60 bp in length) were detected. All polymorphisms were silent: 10 were synonymous (6 and 4 in the *OS-E* and *OS-F* coding regions, respectively), while the rest were in noncoding regions (figure 2). All length polymorphisms were in noncoding regions. Table 1 shows estimates of nucleotide variation for the different *OS*-region functional parts. As in previous reports, levels of synonymous variation (at the coding region) were slightly higher than those present at noncoding fragments. Estimates of the silent nucleotide diversity (π_{SIL} = 0.0088) were similar to that obtained in other surveys at the same European population of *D. simulans* (CIRERA and AGUADÉ 1997; AGUADÉ 1998, 1999; ROZAS *et al.* 2001).

Current silent nucleotide variation levels in *D. simulans* were higher than those estimated at the homologous syntenic region of *D. melanogaster* (π_{SIL} = 0.0021). The intraspecific nucleotide variability distribution along the *OS*-region was also quite different between species. 1) *D. simulans* does not show the gradient in nucleotide diversity observed in *D. melanogaster* (SÁNCHEZ-GRACIA *et al*. 2003), and 2) the HKA test (HUDSON *et al.* 1987) was not significant.

Indeed, in *D. simulans*, levels of polymorphism and divergence correlated, as expected by the neutral model, both for the total region as for the different functional parts ($P > 0.05$, figure 2).

We analyze putative departures from the neutral frequency spectrum by a number of neutrality tests. No significant results were obtained by Tajima´s *D*, Fu and Li *D* and *F* and Fay and Wu's *H* tests, even though the atypical nucleotide structure observed in the sample (see below). Nevertheless, all statistics presented positive values (except for Fay and Wu's *H*), reflecting a rather excess number of substitutions segregating at high frequency.

The analysis at the *OS-E* coding region of the relative levels of synonymous and nonsynonymous substitutions within *D. simulans*, and between *D. simulans* and *D. melanogaster*, were significant (MK test, $P = 0.005$). The test, nevertheless, were not significant using the DNA sequence of *D. mauritiana* as divergence data; therefore, this non neutral behavior is caused by a number substitutions fixed on the *D. melanogaster* lineage.

Linkage disequilibrium and recombination: The *ZZ* statistic value in the European sample was positive and statistically significant ($P = 0.001$), evidencing a major role of the intragenic recombination in shuffling nucleotide variation among DNA sequences. Table 2 shows the estimates of the recombination parameter *C* obtained by different methods. The unusual haplotype structure detected at the *OS*-region (see below) could be the responsible for the discrepancy between C_H and C_R estimates (much more dependent on departures of the neutral equilibrium assumptions), and those based on the recombination map, C_M (ANDOLFFATO and PRZEWORSKI 2000). In all cases, nevertheless, recombination levels in *D. simulans* were clearly higher than those for the syntenic region of *D. melanogaster*. For the total sample, 35% of the pairwise comparisons (1506) showed significant linkage disequilibrium values; nevertheless none of them were significant after applying the conservative Bonferroni procedure. Z_{ns} values were, however, highly significant ($P < 0.001$) after applying the conservative C_L value ($C_L = 11.7$).

Haplotype structure: Present data shows a highly structured nucleotide variation pattern. We identified 6 sequences (out of eleven) identical or nearly identical (differing by a single nucleotide substitution; figure 2). This group of sequences was named as haplogroup H#1. The rest of sequences (H#2 haplogroup) harbored, on the contrary, 83 segregating sites (85 mutations) and 12 indels. In addition, two of the later sequences (S18 and S28) show a chimerical pattern, likely caused by a recombination event between these two divergent haplotypes. We performed computer simulations based on the coalescent process to investigate whether this pattern might be compatible with the neutral equilibrium model. The analysis shows that this haplotype structure was clearly unlikely under the neutral model, even using the conservative C_L value in the simulations (table 4). In particular, we found that number of haplotypes and the haplotype diversity levels were significantly reduced. This reduction in the haplotype diversity also generates highly significant F_s (FU 1997) values. We also made coalescent simulations to estimate the probability of observing a given number of identical sequences (or differing by only one segregating site) in the sample (ROZAS *et al*. 2003). This probability value is a function of θ, *C* and the sample size. Again, the results were highly significant (P ($X \geq 6$) < 0.001). These tests were also highly significant ($P < 0.005$) even when it used the strong conservative criteria of no recombination (table 4). This strong haplotype structure extends for all over the *OS* region (including all their functional regions). The reconstructed neighbor-joining tree (figure 4) clearly illustrates two separate clusters (H#1 and H#2 haplogroups) with a reduced level of variation in the H#1 group. The branch lengths also reflect the substantial differences in the population mutational parameter between *D. melanogaster* and *D. simulans*.

African sample

To determine whether this unusual haplotype structure was also present in other populations of *D. simulans*, we extended the analysis to 11 additional sequences from a population of the putative ancestral geographical area for the species, the east African population of Maputo (Mozambique) (LACHAISE *et al.* 1988). Table 3 summarizes the nucleotide variation estimates for the ~2kb comparable sequenced regions (figure 1). In agreements with previous reports, levels of nucleotide diversity were lower in the European (derived) sample. Silent nucleotide divergence between *D. melanogaster* and *D. simulans* (K_{SIL} = 0.088), however, was within the range estimated for other genes (MORIYAMA and POWELL 1996).

In the African sample we did not found the unusual haplotype structure detected in Europa; nevertheless, one line has identical information than that in the H#1 group (figures 5, 6). This different pattern of variation causes that both populations were genetically differentiated [*P* = 0.012; S_{nn} statistic (HUDSON 2000)]. The neighbor-joining tree (figure 6) clearly reflects this pattern of variation. Interestingly, haplotype-based tests are significant (table 4) even after including the African sample (not structured) in the analysis.

DISCUSSION

Nucleotide diversity at the OS-region in D. simulans and D. melanogaster

Previous studies at the *OS*-region in *D. melanogaster* have shown major differences in the evolutionary history of *OS-E* and *OS-F* genes. Here, we have also found that *D. melanogaster* and *D. simulans* show a distinctive evolutionary pattern. Levels of silent nucleotide variation, as well as estimates of the recombination parameter, were higher in *D. simulans* than in *D. melanogaster* (tables 1 and 2). These results can not be attributed to putative changes in the chromosomal location of the *OS*-region between these species; although there is a fixed inversion between *D. melanogaster* and *D. simulans* (3R chromosomal arm: *85F1-93F6*), it does not include the *OS*-region and thus, we can assume that the *OS*-genes are in a syntenic conserved segment [see AULARD *et al.* (2004) for a review]. Comparisons of the genetic and cytogenetic maps between these species, however, revealed marked differences in the recombination rate along the 3R chromosomal arm (TRUE *et al.* 1996). In fact, in *D. simulans* the coefficient of exchange at the *OS*-region is four times higher than in *D. melanogaster*. A reduction in the recombination rate will increase the linkage selection effects, which can be seen as a reduction in the effective population size (MAYNARD SMITH and HAIGH 1974; CHARLESWORTH *et al.* 1993; HILL and ROBERTSON 1966). This factor, with the global higher effective population size proposed for *D. simulans* [AKASHI 1996; see also MOUSSET and DEROME (2004)], might explain current discrepancies in the levels of silent variation between the two species.

The significant results of the HKA test at the *OS*-region in *D. melanogaster* (SÁNCHEZ-GRACIA *et al*. 2003), but not in *D. simulans*, would agree with the linkage selection effect hypothesis. This fact, however, can not explain the significant silent nucleotide variation gradient detected along the *OS* region in *D. melanogaster*. Yet, a N_e reduction might also affect the fixation probabilities of slightly deleterious mutations (e.g. nonsynonymous muations). In fact, in *D. simulans*, we did not detect any nonsynonymous polymorphisms, and only a single fixed replacement (since the split of the two Drosophila species), while at the *OS-E* gene of *D. melanogaster* we observed 6 nonsynonymous fixations (SÁNCHEZ-GRACIA *et al*. 2003). Furthemore, present MK tests clearly indicate that the non neutral behavior, likely caused by an excess of nonsynonymous fixations, occurs in the *D. melanogater* linage and not in *D. simulans*. Therefore, the data might be interpreted as a relaxation of the strength of selection in *D. melanogater*. In this case, nevertheless, the two *OS*-genes should be affected with a ratter same intensity, and it should also affect both fixed and polymorphic changes. However, we did not detect any amino acid replacement segregating at *OS-E* or *OS-F* genes in these species. Consequently, although a putative N_e reduction might explain part of the data, they can not account for the differential behavior of the two *OS* genes of *D. melanogaster*, suggesting that the *OS-E* and *OS-F* genes are evolving under different selective evolutionary forces.

Haplotype structure in D. simulans

We have detected a strong and atypical genetic structure, caused by the presence of a number of lines with an identical sequence (haplogroup H#1).

Several studies have also showed that African (putative ancestral) and European (derived) populations of these species are genetically structured [e.g. *Pgd* (BEGUN and AQUADRO 1994), *runt* (LABATE *et al.* 1999), *In(2L)*t breakpoint (ANDOLFATTO and KREITMAN 2000), *vermilion* and *G6pd* genes (HAMBLIN and VEUILLE 1999; VEUILLE *et al.* 2004) and *rp49-jan-ocn* region (ROZAS *et al.* 2001; PARCH *et al.* 2001; QUESADA *et al.* 2003)]. None of these surveys, however, have show a continuous structured genomic region as longer than the observed in the *OS*-region. These results again could not be explained by putative differences in the effective population size, and are more likely to reflect demographic and/or selective effects.

Demographic factors: There are a number of demographic scenarios that, at priori, might explain this unusual haplotype configuration. *D. simulans*, as *D. melanogaster*, originated in tropical areas and, with the rise of agriculture (i.e. in historical times), spread worldwide as human commensal (DOBZHANSKY 1965; LACHAISE *et al.* 1988). The species, therefore, experienced a number of founder events, and likely adaptive changes, through a period much shorter than the within species MRCA time ($4N_e$ as average). Therefore, the signature of the historical events should still be present in patterns of molecular evolution, and hence, the assumption of neutral stationary equilibrium can be unjustified. HAMBLIN and VEUILLE (1999) suggested that *D. simulans* derived populations could have been generated by a recent admixture of genetically differentiated African populations. ANDOLFATTO and PRZEWORSKI (2000), comparing the population parameters C and θ in 16 independent loci of *D. simulans*, found greater than expected intralocus linkage disequilibrium. They showed that the data does not fit with a symmetric island model, and would need more complex demographic scenarios to be explained. ANDOLFATTO (2001) reexamined the available data and concluded that, although congruent with a simple bottleneck caused by the out-of-Africa, it might be also explained by the presence of an ancient (African) population structure. Besides, WALL *et al*. (2002) found that under reasonable conditions, no simple evolutionary model (a simple hitchhiking or a bottleneck) could explain the North American *D. simulans* data set of BEGUN and WHITLEY (2000).

Several pieces of the *OS* data reflect the signature of a demographic event, either a bottleneck caused by founder effects originated in the out-of-Africa, or a recent population admixture. First, Montblanc and Maputo are clearly genetically differentiated populations. Second, levels of silent nucleotide variation in the European population were lower than in the African sample (table 3). Third, the strong haplotype structure detected in the European sample departs significantly from the neutral equilibrium (figure 4). Under the bottleneck scenario, however, we should not expect two major haplogroups, one with nearly no variation and the other with regular levels of polymorphism. This pattern is more plausible under a model that considers the admixture of two highly differentiated populations. This scenario requires, nevertheless, that one of such populations should harbor no variation. Moreover, present DNA polymorphism pattern could also be caused by more complex demographic scenarios; for example by spatial or temporal fluctuations in the local effective population sizes (see GRAVOT *et al.* 2004). Nevertheless, it is not clear whether this scenario could have occurred in the evolutionary history of the Montblanc population.

Selective factors: The presence of the H#1 haplotype (with nearly zero variation) in a very large fragment (~5kb), could be the fingerprint caused by the increase in frequency of an advantageous mutation in, or close to, the *OS*-region, i.e. a selective sweep (MAYNARD SMITH and HAIGH 1974). This scenario has also been proposed to explain the similar unusual haplotype structure observed in the *rp49* gene (ROZAS *et al.* 2001), and in their closest linked regions in this species (PARSCH *et al.* 2001; QUESADA *et al*. 2003, MEIKLEJOHN *et al.* 2004). These authors found the same haplotype structure both in ancestral and derived populations, with no significant differentiation between them. Moreover, QUESADA *et al.* (2003) observed that the haplotype structure gradually decayed at both sides of the most structured stretch, as expected for a recombining region under positive natural selection. MEIKLEJOHN *et al*. (2004) come to the same conclusion in another survey at the same genomic region. Due that *OS* and *rp49* regions are clearly unlinked (they are located 16 polytene bands apart) these surveys, likely, would be detecting different sweep events.

In contrast to ROZAS *et al*. (2001) data, we found that the unusual haplotype structure is present only in the European population; therefore, the putative selective event might reflect some local adaptations. The microsatellite variability screen at several *D. simulans* populations is consistent with that scenario (SCHOFL and SCHLOTTERER 2004); this study suggests that the number of beneficial mutations seems to be higher in derived populations, and would reflect the adaptive process to new environments. We can not discard, nevertheless, that the selective sweep really originated in the ancestral Africa population. In fact, we also detected one line of the haplotype H#1 in this population (figure 5). In this case, current patterns of variation might be explained by the action of natural selection (in the African population) followed of some demographic events (bottleneck or admixture). The estimates of the time back to the hitchhiking event (~10,000 years for the *OS*-region; ~6,500 years for the *rp49* region), is consistent with this hypothesis.

Most of the nucleotide variation surveys in *D. simulans* have been conducted in North American populations, being the Europe samples exceptionals. In addition, many of these studies analyzed only a few number of sequences, or used inadequate sampling strategies to draw firm conclusions (e.g. ANDOLFATTO and PZREWORSKI 2000; ANDOLFATTO 2001; WALL *et al*. 2002). Interestingly, if we ignore polymorphism data form surveys with less than 8 sequences (within a single population), many of the studies would point to selective forces to explain for most of the data (e.g. ZUROVCOVA and EANES 1999; KERN *et al*. 2002, 2004; SCHLENKE and BEGUN 2003, 2004, 2005; DEROME *et al*. 2004; DUMONT *et al*. 2004; LAZARO 2005; but see also INRVIN *et al*. 1998; SCHMID *et al*. 1999; DUVERNELL and EANES 2000). Therefore, we should consider that, without discarding the probable occurrence of overlapping demographic events, adaptive evolution could have a conspicuous role in *D. simulans* populations. A large-scale genomic survey in different geographically distributed populations of *D. simulans* would be needed to unambiguously determine such major role of the positive selection in this species.

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., 1996 Molecular evolution between *Drosophila melanogaster* and *D. simulans*: reduced codon bias, faster rates of amino acid substitution, and larger proteins in *D. melanogaster*. Genetics 144: 1297-1307.

ANDOLFATTO, P., and M. KREITMAN, 2000 Molecular variation at the *In(2L)t* proximal breakpoint site in natural populations of *Drosophila melanogaster* and *D. simulans*. Genetics 154: 1681-1691.

ANDOLFATTO, P., and M. PRZEWORSKI, 2000 A genome-wide departure from the standard neutral model in natural populations of Drosophila. Genetics 156: 257-268.

ANDOLFATTO, P., 2001 Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*. Mol Biol Evol 18: 279-290.

ASHBURNER, M., 1989 *Drosophila: A laboratory handbook*. Cold Spring Harbor Laboratory Press, New York.

AULARD, S., L. MONTI, N. CHAMINADE and F. LEMEUNIER, 2004 Mitotic and polytene chromosomes: comparisons between *Drosophila melanogaster* and *Drosophila simulans*. Genetica 120: 137-150.

BEGUN, D. J., and C. F. AQUADRO, 1994 Evolutionary inferences from DNA variation at the *6-Phosphogluconate dehydrogenase* locus in natural populations of Drosophila: Selection and geographic differentiation. Genetics 136: 155-171.

BEGUN, D. J., and P. WHITLEY, 2000 Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc Natl Acad Sci U S A 97: 5960-5965.

CHARLESWORTH, B., M. T. MORGAN and D. CHARLESWORTH, 1993 The effect of deleterious mutations on neutral molecular variation. Genetics 134: 1289-1303.

CIRERA, S., and M. AGUADÉ, 1997 Evolutionary history of the sex-peptide *(Acp70A)* gene region in *Drosophila melanogaster*. Genetics 147: 189-197.

CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. D. THOMAS, A. KEJARIWAL et al., 2003 Inferring nonneutral evolution from Human-Chimp-Mouse orthologous gene trios. Science 302: 1960-1963.

DEROME, N., K. METAYER, C. MONTCHAMP-MOREAU and M. VEUILLE, 2004 Signature of selective sweep associated with the evolution of sex-ratio drive in *Drosophila simulans*. Genetics 166: 1357-1366.

DOBZHANSKY, T., 1965 Wild and domestic species of Drosophila, pp. 533-547 in *The Genetics of Colonizing Species*, edited by H. G. A. L. S. DAKER, G. Academic Press, New York and London.

DUMONT, V. B., J. C. FAY, P. P. CALABRESE and C. F. AQUADRO, 2004 DNA variability and divergence at the *Notch* locus in *Drosophila melanogaster* and *D. simulans*: A case of accelerated synonymous site divergence. Genetics 167: 171-185.

DUVERNELL, D. D., and W. F. EANES, 2000 Contrasting molecular population genetics of four hexokinases in *Drosophila melanogaster*, *D. simulans* and *D. yakuba*. Genetics 156: 1191-1201.

EMES, R. D., S. A. BEATSON, C. P. PONTING and L. GOODSTADT, 2004 Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. Genome Res 14: 591-602.

FAY, J. C., and C. I. WU, 2000 Hitchhiking under positive Darwinian selection. Genetics 155: 1405-1413.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693-709.

FU, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915-925.

GALINDO, K., and D. P. SMITH, 2001 A large family of divergent Drosophila odorant-binding proteins expressed in gustatory and olfactory sensilla. Genetics 159: 1059-1072.

GILAD, Y., C. D. BUSTAMANTE, D. LANCET and S. PAABO, 2003 Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet 73: 489-501.

GIMELBRANT, A. A., H. SKALETSKY and A. CHESS, 2004 Selective pressures on the olfactory receptor repertoire since the human-chimpanzee divergence. Proc Natl Acad Sci U S A 101: 9019-9022.

GRAVOT, E., M. HUET and M. VEUILLE, 2004 Effect of breeding structure on population genetic parameters in Drosophila. Genetics 166: 779-788.

HAMBLIN, M. T., and M. VEUILLE, 1999 Population structure among African and derived populations of *Drosophila simulans*: evidence for ancient subdivision and recent admixture. Genetics 153: 305-317.

HEKMAT-SCAFE, D. S., R. L. DORIT and J. R. CARLSON, 2000 Molecular evolution of odorant-binding protein genes *OS-E* and *OS*-F in Drosophila. Genetics 155: 117-127.

HEKMAT-SCAFE, D. S., C. R. SCAFE, A. J. MCKINNEY and M. A. TANOUYE, 2002 Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. Genome Res 12: 1357-1369.

HILL, W. G., and A. ROBERTSON, 1966 The effect of linkage on limits to artificial selection. Genet Res 8: 269-294.

HUDSON, R. R., 1983 Properties of a neutral allele model with intragenic recombination. Theor Popul Biol 23: 183-201.

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

HUDSON, R. R., M. KREITMAN and M. AGUADÉ, 1987 A test of neutral molecular evolution based on nucleotide data. Genetics 116: 153-159.

HUDSON, R. R., 1990 Gene genealogies and the coalescent process, pp. 1-44 in *Oxford Surveys in Evolutionary Biology*, edited by J. ANTONOVICS and D. FUTUYMA. Oxford University Press, Oxford.

HUDSON, R. R., 2000 A new statistic for detecting genetic differentiation. Genetics 155: 2011-2014.

IRVIN, S. D., K. A. WETTERSTRAND, C. M. HUTTER and C. F. AQUADRO, 1998 Genetic variation and differentiation at microsatellite loci in *Drosophila* *simulans*. Evidence for founder effects in new world populations. Genetics 150: 777-790.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21-123 *in Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.

KERN, A. D., C. D. JONES and D. J. BEGUN, 2002 Genomic effects of nucleotide substitutions in *Drosophila simulans*. Genetics 162: 1753-1761.

KERN, A. D., C. D. JONES and D. J. BEGUN, 2004 Molecular population genetics of male accessory gland proteins in the *Drosophila simulans* complex. Genetics 167: 725-735.

KINGMAN, J. F. C., 1982a On the genealogy of large populations. J. Appl. Probab. 19A: 27-43.

KINGMAN, J. F. C., 1982b The coalescent. *Stochastic Processes and Their Applications*. 13: 235-248.

KRIEGER, M. J., and K. G. ROSS, 2002 Identification of a major gene regulating complex social behavior. Science 295: 328-332.

KRIEGER, M. J., and K. G. ROSS, 2005 Molecular evolutionary analyses of the odorant-binding protein gene Gp-9 in fire ants and other Solenopsis species. Mol Biol Evol 22: 2090-2103

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.

LACHAISE, D., M. L. CARIOU, J. R. DAVID, F. LEMEUNIER, L. TSACAS et al., 1988 Historical biogeography of the *Drosophila melanogaster* species subgroup. Evolutionary Biology 22: 159-225.

LAZZARO, B. P., 2005 Elevated polymorphism and divergence in the class C scavenger receptors of *Drosophila melanogaster* and *D. simulans*. Genetics 169: 2023-2034.

MADDISON, W. P., and D. R. MADDISON, 1992 *MacClade: Analysis of Phylogeny and Character Evolution*. Version 3.0. Sinauer, Sunderland, MA.

MAYNARD SMITH, J., and J. HAIGH, 1974 The hitch-hiking effect of a favourable gene. Genet Res 23: 23-35.

MEIKLEJOHN, C. D., Y. KIM, D. L. HARTL and J. PARSCH, 2004 Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. Genetics 168: 265-279.

MORIYAMA, E. N., and J. R. POWELL, 1996 Intraspecific nuclear DNA variation in Drosophila. Mol Biol Evol 13: 261-277.

MOUSSET, S., and N. DEROME, 2004 Molecular polymorphism in *Drosophila melanogaster* and *D. simulans*: what have we learned from recent studies? Genetica 120: 79-86.

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

NGAI, J., M. M. DOWLING, L. BUCK, R. AXEL and A. CHESS, 1993 The family of genes encoding odorant receptors in the channel catfish. Cell 72: 657-666.

NIELSEN, R., C. BUSTAMANTE, A. G. CLARK, S. GLANOWSKI, T. B. SACKTON et al., 2005 A scan for positively selected genes in the genomes of humans and chimpanzees. PLoS Biol 3: e170.

PALMER, C. A., R. A. WATTS, R. G. GREGG, M. A. MCCALL, L. D. HOUCK et al., 2005 Lineage-specific differences in evolutionary mode in a salamander courtship pheromone. Mol Biol Evol 22: 2243-2256.

PARSCH, J., C. D. MEIKLEJOHN and D. L. HARTL, 2001 Patterns of DNA sequence variation suggest the recent action of positive selection in the *janus-ocnus* region of *Drosophila simulans*. Genetics 159: 647-657.

PELOSI, P., and R. MAIDA, 1995 Odorant-binding proteins in insects. Comp Biochem Physiol B Biochem Mol Biol 111: 503-514.

QUESADA, H., U. E. M. RAMIREZ, J. ROZAS and M. AGUADÉ, 2003 Large-scale adaptive hitchhiking upon high recombination in *Drosophila simulans*. Genetics 165: 895-900.

ROZAS, J., M. GULLAUD, G. BLANDIN and M. AGUADÉ, 2001 DNA Variation at the *rp49* gene region of *Drosophila simulans*: Evolutionary inferences from an unusual haplotype structure. Genetics 158: 1147-1155.

ROZAS, J., J. C. SÁNCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI et al., 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487-491.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

SÁNCHEZ-GRACIA, A., M. AGUADÉ and J. ROZAS, 2003 Patterns of nucleotide polymorphism and divergence in the Odorant-Binding Protein genes

OS-E and *OS-F*: Analysis in the melanogaster species subgroup of Drosophila. Genetics 165: 1279-1288.

SCHLENKE, T. A., and D. J. BEGUN, 2003 Natural selection drives Drosophila immune system evolution. Genetics 164: 1471-1480.

SCHLENKE, T. A., and D. J. BEGUN, 2004 Strong selective sweep associated with a transposon insertion in *Drosophila simulans*. Proc Natl Acad Sci U S A 101: 1626-1631.

SCHLENKE, T. A., and D. J. BEGUN, 2005 Linkage disequilibrium and recent selection at three immunity receptor loci in *Drosophila simulans*. Genetics 169: 2013-2022.

SCHMID, K. J., L. NIGRO, C. H. AQUADRO and D. TAUTZ, 1999 Large number of replacement polymorphisms in rapidly evolving genes of Drosophila: Implications for genome-wide surveys of DNA polymorphism. Genetics 153: 1717-1729.

SCHOFL, G., and C. SCHLOTTERER, 2004 Patterns of microsatellite variability among X chromosomes and autosomes indicate a high frequency of beneficial mutations in non-african *D. simulans*. Mol Biol Evol 21: 1384-1390.

STROBECK, C., 1987 Average number of nucleotide differences in a sample from a single subpopulation - a test for population subdivision. Genetics 117: 149-153.

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

THOMPSON, J. D., T. J. GIBSON, F. PLEWNIAK, F. JEANMOUGIN and D. G. HIGGINS, 1997 The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25: 4876-4882.

TRUE, J. R., J. M. MERCER and C. C. LAURIE, 1996 Differences in crossover frequency and distribution among three sibling species of Drosophila. Genetics 142: 507-523.

VEUILLE, M., E. BAUDRY, M. COBB, N. DEROME and E. GRAVOT, 2004 Historicity and the population genetics of *Drosophila melanogaster* and *D. simulan*s. Genetica 120: 61-70.

VOGT, R. G., and L. M. RIDDIFORD, 1981 Pheromone binding and inactivation by moth antennae. Nature: 161-163.

VOGT, R. G., G. D. PRESTWICH and M. R. LERNER, 1991 Odorant-binding-protein subfamilies associate with distinct classes of olfactory receptor neurons in insects. J Neurobiol 22: 74-84.

VOGT, R. G., F. E. CALLAHAN, M. E. ROGERS and J. C. DICKENS, 1999 Odorant binding protein diversity and distribution among the insect orders, as indicated by LAP, an OBP-related protein of the true bug *Lygus lineolaris* (Hemiptera, Heteroptera). Chem Senses 24: 481-495.

VOGT, R. G., M. E. ROGERS, M. D. FRANCO and M. SUN, 2002 A comparative study of odorant binding protein genes: differential expression of the *PBP1-GOBP2* gene cluster in *Manduca sexta* (Lepidoptera) and the organization of *OBP* genes in *Drosophila melanogaster* (Diptera). J Exp Biol 205: 719-744.

VOSSHALL, L. B., 2000 Olfaction in Drosophila. Curr Opin Neurobiol 10: 498-503.

WALL, J. D., P. ANDOLFATTO and M. PRZEWORSKI, 2002 Testing models of selection and demography in *Drosophila simulans*. Genetics 162: 203-216.

WATTERSON, G. A., 1975 On the number of segregating sites in genetical models without recombination. Theor Popul Biol 7: 256-276.

WATTS, R. A., C. A. PALMER, R. C. FELDHOFF, P. W. FELDHOFF, L. D. HOUCK et al., 2004 Stabilizing selection on behavior and morphology masks positive selection on the signal in a salamander pheromone signaling complex. Mol Biol Evol 21: 1032-1041.

WILLETT, C. S., 2000 Evidence for directional selection acting on pheromone-binding proteins in the genus Choristoneura. Mol Biol Evol 17: 553-562.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

ZUROVCOVA, M., and W. F. EANES, 1999 Lack of nucleotide polymorphism in the Y-linked sperm flagellar Dynein Gene *Dhc-Yh3* of *Drosophila melanogaster* and *D. simulans*. Genetics 153: 1709-1715.

TABLE 1

Summary of the nucleotide variation at the *OS*-region in the Montblanc population of *D. simulans*

	5'	*OS-E*	Intergenic	*OS-F*	Total
Silent					
No. sites[a]	75	308.5	934	2609.3	3926.8
S	2	12	18	64	96
π	0.0087	0.0144	0.0070	0.0088	0.0088
θ	0.0091	0.0133	0.0066	0.0088	0.0086
K	0.0194	0.1076	0.0522	0.0462	0.0519
Synonymous					
No. sites[a]		88.5		101.3	189.8
S		6		4	10
π		0.0275		0.0111	0.0188
θ		0.0232		0.0135	0.0180
K		0.1260		0.0578	0.0887
Noncoding					
No. sites[a]	75	220	934	2508	3737
S	2	6	18	60	86
π	0.0087	0.0091	0.0070	0.0087	0.0083
θ	0.0091	0.0093	0.0066	0.0086	0.0081
K	0.0194	0.1000	0.0522	0.0457	0.0499

S, number of segregating sites; *K*, nucleotide divergence between *D. simulans* and *D. melanogaster*

[a], Number of sites in the polymorphism data set

TABLE 2

Estimates[a] of the recombination parameter *C* for the *OS*-region

	D. simulans (*n* = 11)	*D. melanogaster* (*n* = 14)
R_M (C_R)	6 (25)	2 (12)
C_H	3.7	12.5
C_L	11.7	4.5
C_M	186.8	14.7

[a] European population data.

n, sample size

TABLE 3

Summary of the nucleotide variation at the fragment 2 of *D. simulans*

	Maputo	Total
Sample size	11	22
S (*η*)	71 (72)	81 (83)
No. sites	1733	1733
No. silent sites	1398.4	1398.4
π_{SIL}	0.0200	0.0176
K_{SIL}	0.0594	0.0606

S, number of segregating sites; *η*, number of mutations; π_{SIL}, silent nucleotide diversity; K_{SIL}, silent nucleotide divergence between *D. simulans* and *D. melanogaster.*

TABLE 4

Haplotype-based tests for the fragment 2

		Probability		
		$C = 0$	$C = C_L$	$C = C_M$
Montblanc[a]				
I +1	6	<0.001	<0.001	<0.001
k	6	0.005[c]	0.001	<0.001
h	0.86	0.005	0.001	<0.001
Fu's F_s	7.59	0.005	0.001	<0.001
Total[b]				
I +1	7	0.077	0.006	<0.001
k	14	0.227[c]	0.03	0.001
h	0.94	0.141	0.019	<0.001
Fu's F_s	3.62	0.094	0.005	<0.001

[a], Montblanc sample, n = 11; m = 4622; C_L = 11.7

[b], Total sample, n = 22; m = 1733; C_L = 24.8

[c], Strobeck's (1987) S statistic

m, number of sites

FIGURE LEGENDS

Figure 1. Structure of the *OS*-region in *D. melanogaster*. Black and grey boxes indicate the coding regions of the *OS-E* and *OS-F* genes, respectively. White boxes indicate untranscribed exons. Introns are indicated by a V. The two amplification fragments are also indicated.

Figure 2. Nucleotide polymorphisms found at the complete *OS*-region of the Montblanc population. A dot represents nucleotides identical to the first sequence. For length polymorphisms, the position indicates the first site affected. Site 1 corresponds to the first position of the *OS-E* translation start codon. Coding positions are shaded. The two last rows indicate the nucleotide present in *D. mauritiana* and *D. melanogaster* for each polymorphism detected in *D. simulans*. Dashes indicate the absence of the corresponding variant. E, *OS-E* coding region; F, *OS-F* coding region; d, deletion; i, insertion; d#, delection of # bp; i# insertion of # bp; Rec, putative recombinant sequences.

Figure 3. Sliding window of silent polymorphism and divergence along the *OS*-region. Window size, 1000 bp; step size, 250 bp.

Figure 4. Neighbor-joining tree of the *OS*-region reconstructed using Jukes and Cantor corrected distances. *D. simulans* and *D. melanogaster* sequences are depicted as light circles and solid diamonds, respectively. Bootstrap values >90% are shown under the branches, and X/Y values, represent the number of synonymous and nonsynonymous changes, respectively.

Figure 5. Nucleotide polymorphisms detected at the fragment 2 for the two populations of *D. simulans*. African lines are shown as MZ.

Figure 6. Neighbor-joining tree of the *OS*-region (fragment 2). Open and solid circles indicate Montblanc and Maputo sequences, respectively.

FIGURE 1

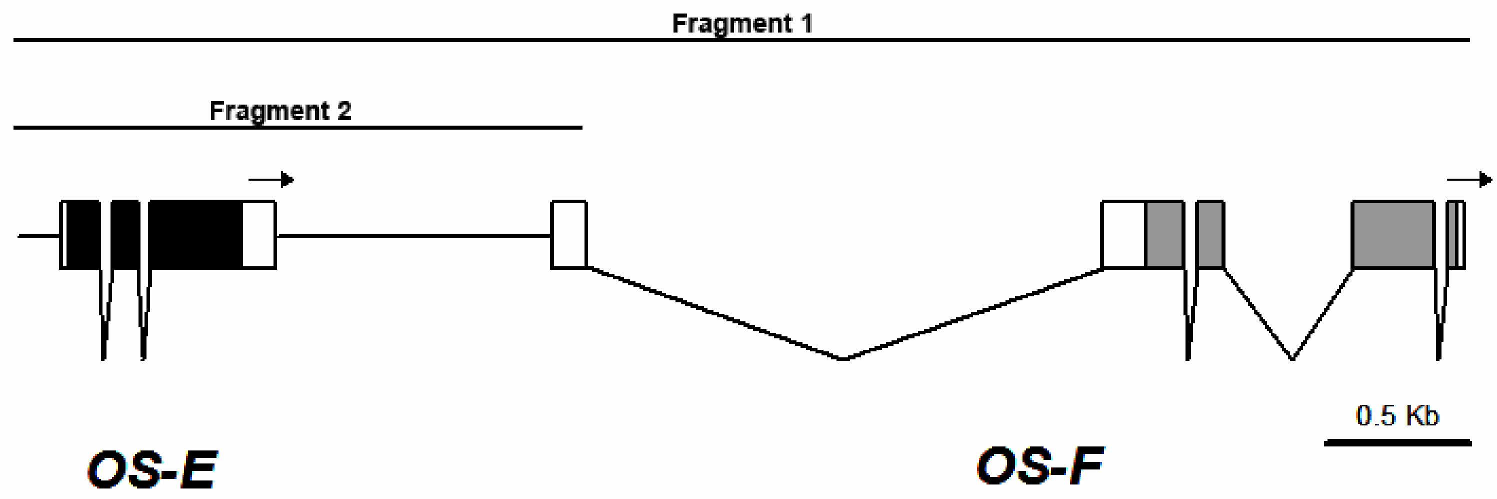


FIGURE 2

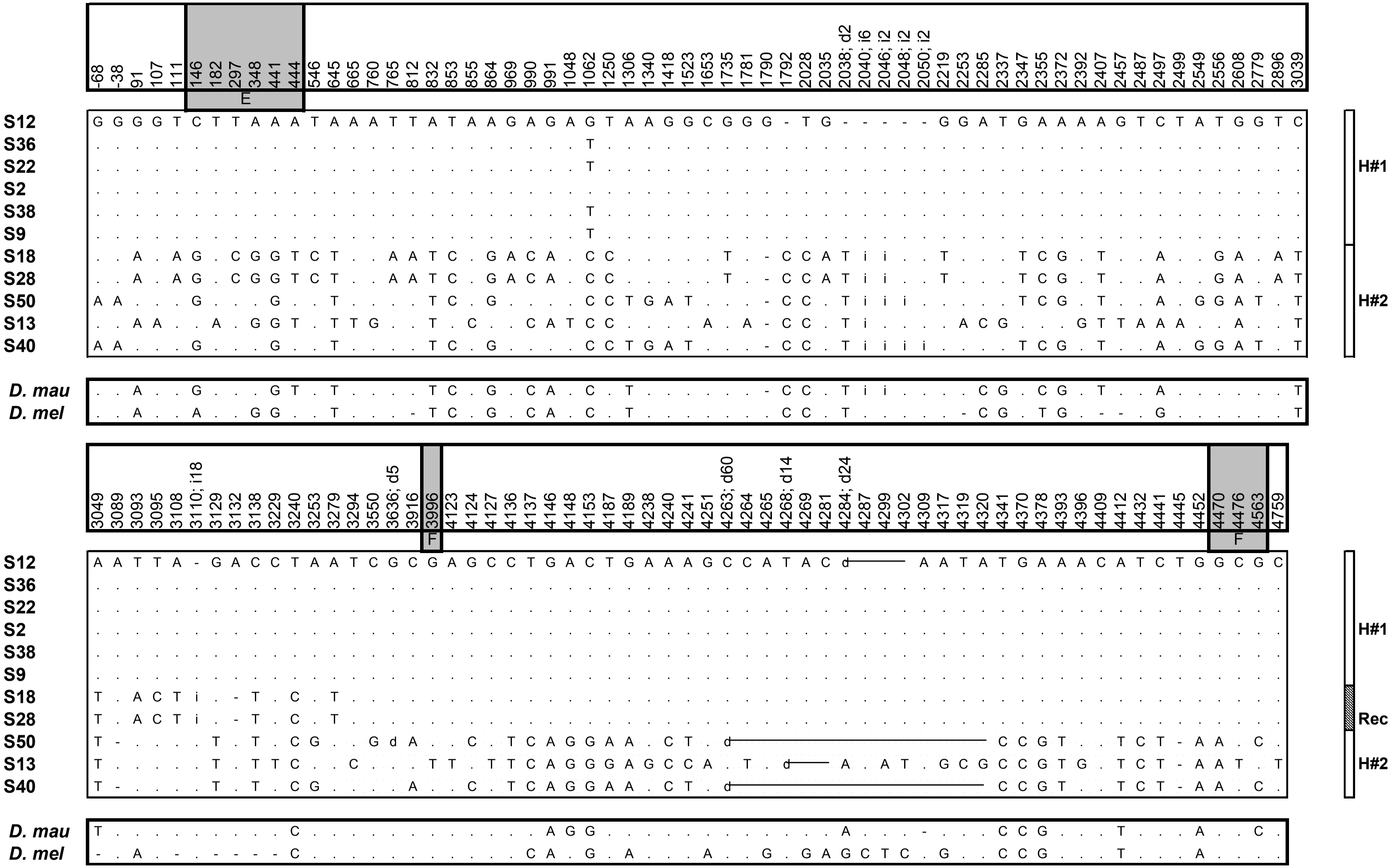

-68 -38 91 107 111 146 182 297 348 441 444 546 645 665 760 765 812 832 853 855 864 969 990 991 1048 1062 1250 1306 1340 1418 1523 1653 1735 1781 1790 1792 2028 2035 2038; d2 2040; i6 2046; i2 2048; i2 2050; i2 2219 2253 2285 2337 2347 2355 2372 2392 2407 2457 2487 2497 2499 2549 2556 2608 2779 2896 3039
E
S12 G G G G T C T T A A A T A A A T T A T A A G A G A G T A A G G C G G G - T G - - - - - G G A T G A A A A G T C T A T G G T C
S36 T
S22 T
S2
S38 T
S9 T
S18 . . A . A G . C G G T C T . . A A T C . G A C A . C C . . . . . T . - C C A T i i . . T . . . T C G . T . . A . . G A . A T
S28 . . A . A G . C G G T C T . . A A T C . G A C A . C C . . . . . T . - C C A T i i . . T . . . T C G . T . . A . . G A . A T
S50 A A . . . G . . . G . . T . . . . T C . G . . . . C C T G A T . . . - C C . T i i i . . . . . T C G . T . . A . G G A T . T
S13 . . A A . . A . G G T . T T G . . T . C . . C A T C C . . . . A . A - C C . T i . . . . A C G . . . G T T A A A . . A . . T
S40 A A . . . G . . . G . . T . . . . T C . G . . . . C C T G A T . . . - C C . T i i i i . . . . T C G . T . . A . G G A T . T
H#1
H#2
D. mau . . A . . G . . . G T . T . . . . T C . G . C A . C . T . . . . . . - C C . T i i . . . . C G . C G . T . . A . . . . . . T
D. mel . . A . . A . . G G . . T . . . - T C . G . C A . C . T . . . . C C . T . . . . . - C G . T G . - - . G . . . . . . T
3049 3089 3093 3095 3108 3110; i18 3129 3132 3138 3229 3240 3253 3279 3294 3550 3636; d5 3916 3996 4123 4124 4127 4136 4137 4146 4148 4153 4187 4189 4238 4240 4241 4251 4263; d60 4264 4265 4268; d14 4269 4281 4284; d24 4287 4299 4302 4309 4317 4319 4320 4341 4370 4378 4393 4396 4409 4412 4432 4441 4445 4452 4470 4476 4563 4759
F
F
S12 A A T T A - G A C C T A A T C G C G A G C C T G A C T G A A A G C C A T A C d A A T A T G A A A C A T C T G G C G C
S36
S22
S2
S38
S9
S18 T . A C T i . - T . C . T
S28 T . A C T i . - T . C . T
S50 T - . . . . T . T . C G . . G d A . . C . T C A G G A A . C T . d C C G T . . T C T - A A . C .
S13 T . . . . . T . T T C . . C . . . T T . T T C A G G G A G C C A . T . d A . A T . G C G C C G T G . T C T - A A T . T
S40 T - . . . . T . T . C G . . . . A . . C . T C A G G A A . C T . d C C G T . . T C T - A A . C .
H#1
Rec
H#2
D. mau T . . . . C . . . . A G G . . . . A . . - . . C C G . . . T . . . A . . C .
D. mel - . . A . - . - - - - C . . . . C A . G . A . . . A . . G . G A G C T C . G . . C C G . . . T . . . A . . . .

FIGURE 3

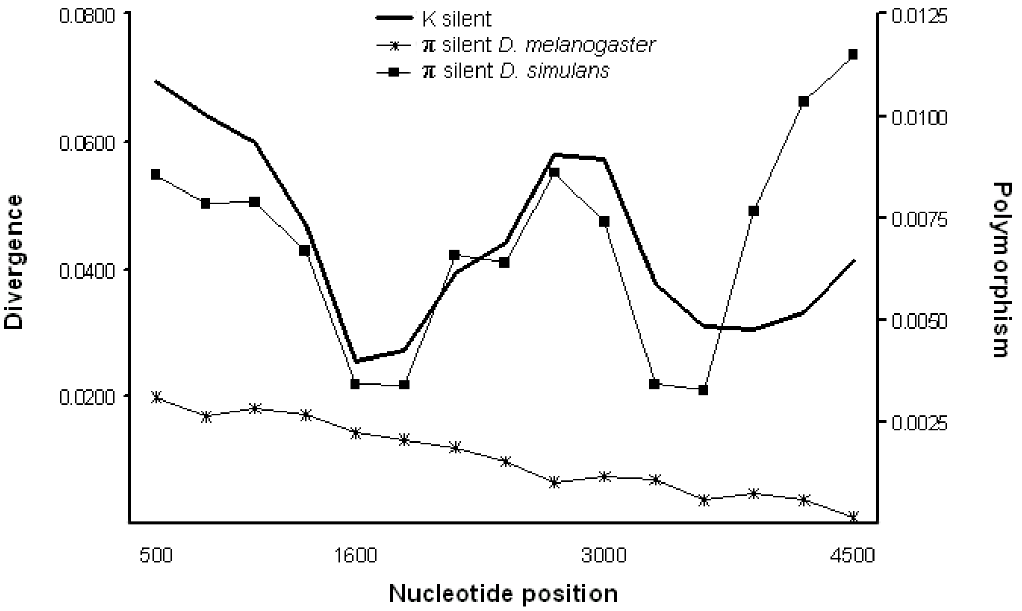
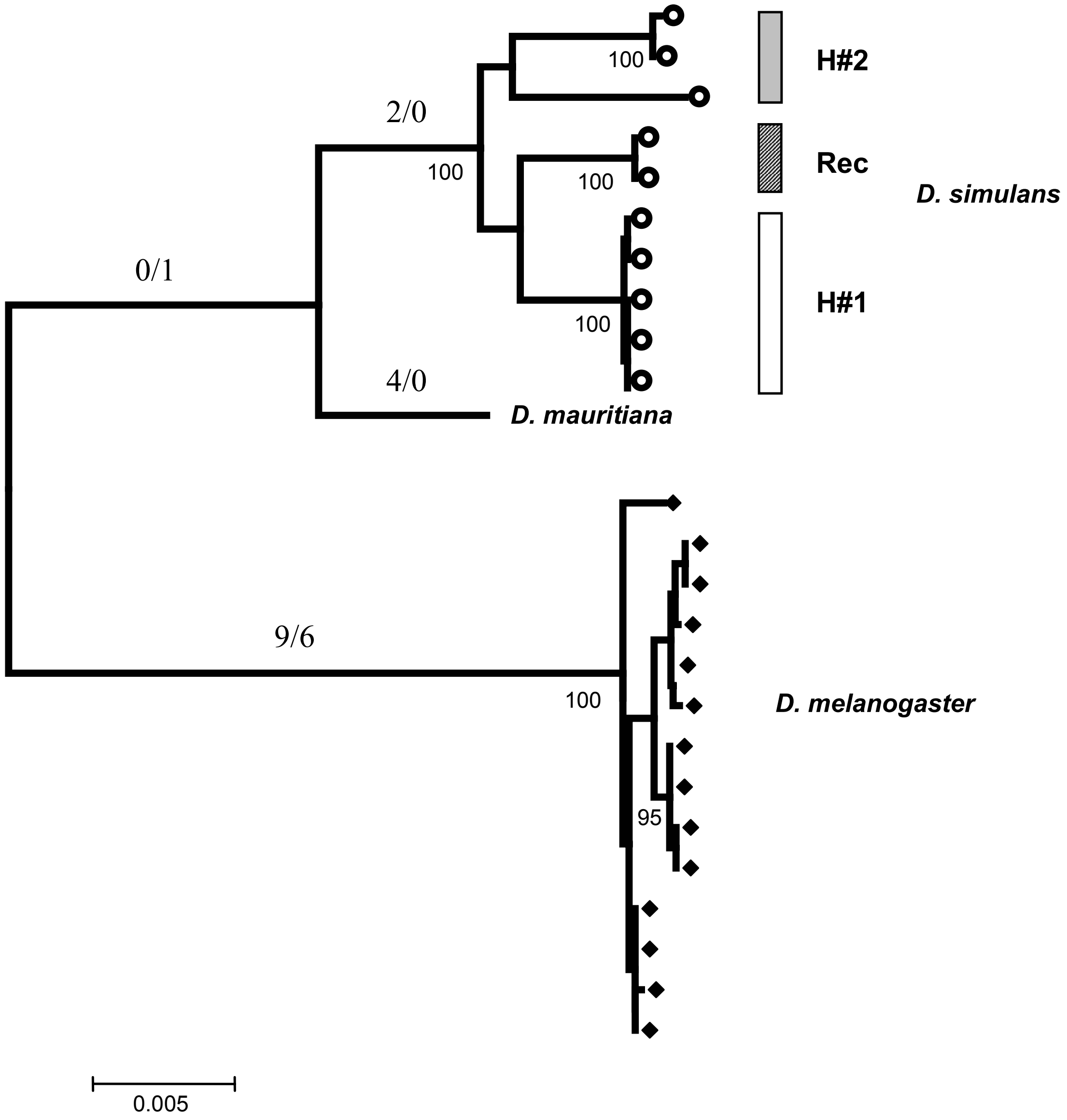
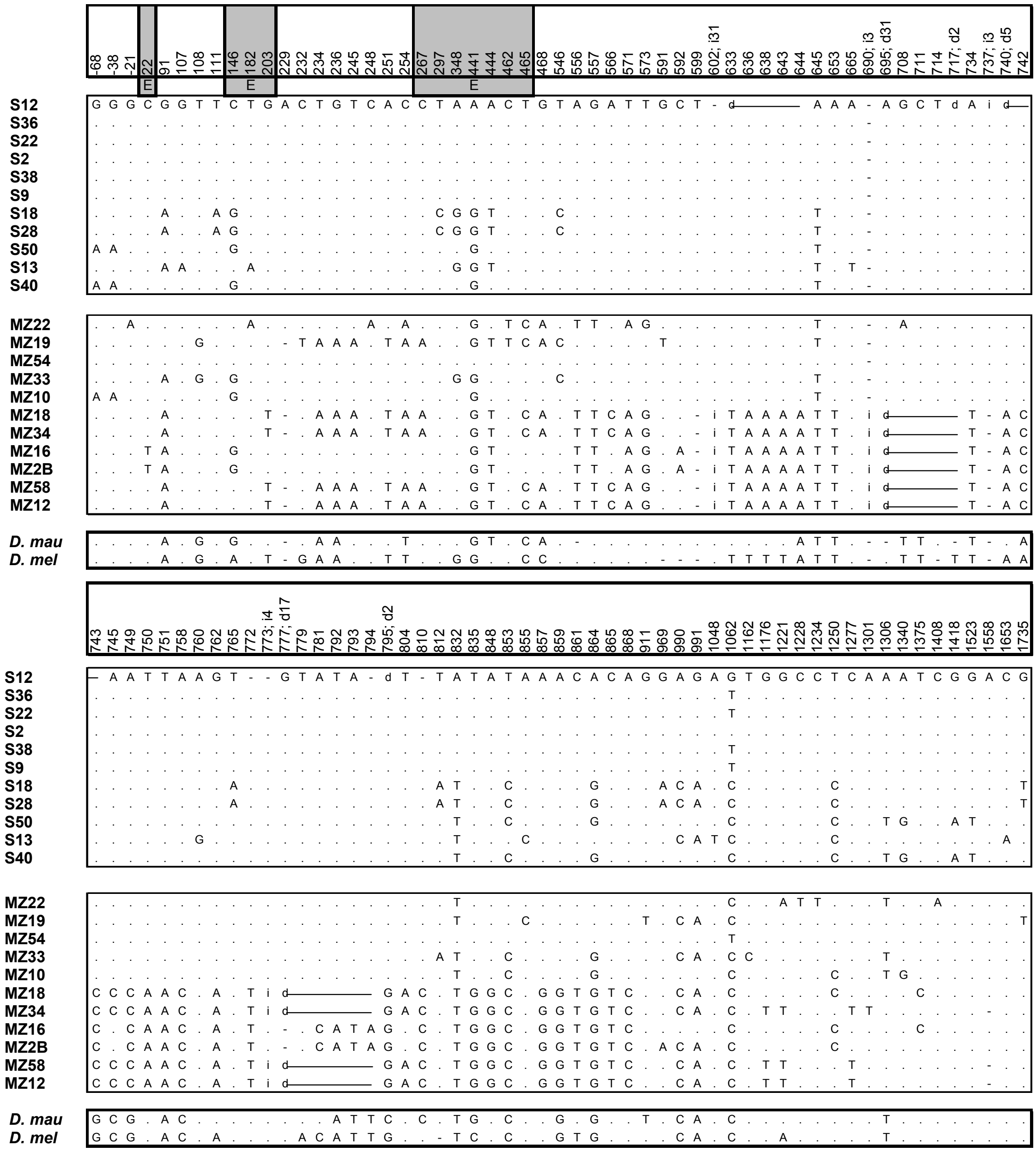

K silent
π silent *D. melanogaster*
π silent *D. simulans*
Divergence
0.0800
0.0600
0.0400
0.0200
Polymorphism
0.0125
0.0100
0.0075
0.0050
0.0025
500
1600
3000
4500
Nucleotide position

FIGURE 4



100
2/0
100
100
0/1
100
4/0
D. mauritiana
H#2
Rec
D. simulans
H#1
9/6
100
D. melanogaster
95
0.005

FIGURE 5



```
      -68 -38 -21 22 91 107 108 111 146 182 203 229 232 234 236 245 248 251 254 267 297 348 441 444 462 465 468 546 556 557 566 571 573 591 592 599 602;i31 633 636 638 643 644 645 653 665 690;i3 695;d31 708 711 714 717;d2 734 737;i3 740;d5 742
                 E                E                                                E
S12   G G G C G G T T C T G A C T G T C A C C T A A A C T G T A G A T T G C T - d———————— A A A - A G C T d A i d——
S36   . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . - . . . . . . . . .
S22   . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . - . . . . . . . . .
S2    . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . - . . . . . . . . .
S38   . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . - . . . . . . . . .
S9    . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . - . . . . . . . . .
S18   . . . . A . . A G . . . . . . . . . . . C G G T . . . C . . . . . . . . . . . . . . T . . - . . . . . . . . .
S28   . . . . A . . A G . . . . . . . . . . . C G G T . . . C . . . . . . . . . . . . . . T . . - . . . . . . . . .
S50   A A . . . . . . G . . . . . . . . . . . . . G . . . . . . . . . . . . . . . . . . . T . . - . . . . . . . . .
S13   . . . . A A . . . A . . . . . . . . . . . G G T . . . . . . . . . . . . . . . . . . T . T - . . . . . . . . .
S40   A A . . . . . . G . . . . . . . . . . . . . G . . . . . . . . . . . . . . . . . . . T . . - . . . . . . . . .

MZ22  . . A . . . . . . A . . . . . . A . A . . . G . T C A . T T . A G . . . . . . . . . T . . - . A . . . . . . .
MZ19  . . . . . . G . . . . - T A A A . T A A . . G T T C A C . . . . . T . . . . . . . . T . . - . . . . . . . . .
MZ54  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . - . . . . . . . . .
MZ33  . . . . A . G . G . . . . . . . . . . . . G G . . . . C . . . . . . . . . . . . . . T . . - . . . . . . . . .
MZ10  A A . . . . . . G . . . . . . . . . . . . . G . . . . . . . . . . . . . . . . . . . T . . - . . . . . . . . .
MZ18  . . . . A . . . . . T - . A A A . T A A . . G T . C A . T T C A G . . - i T A A A A T T . i d———————— T - A C
MZ34  . . . . A . . . . . T - . A A A . T A A . . G T . C A . T T C A G . . - i T A A A A T T . i d———————— T - A C
MZ16  . . . T A . . . G . . . . . . . . . . . . . G T . . . . T T . A G . A - i T A A A A T T . i d———————— T - A C
MZ2B  . . . T A . . . G . . . . . . . . . . . . . G T . . . . T T . A G . A - i T A A A A T T . i d———————— T - A C
MZ58  . . . . A . . . . . T - . A A A . T A A . . G T . C A . T T C A G . . - i T A A A A T T . i d———————— T - A C
MZ12  . . . . A . . . . . T - . A A A . T A A . . G T . C A . T T C A G . . - i T A A A A T T . i d———————— T - A C

D. mau . . . . A . G . G . . - . A A . . . T . . . G T . C A . - . . . . . . . . . . . . A T T . - - T T . - T - . A
D. mel . . . . A . G . A . T - G A A . . T T . . G G . . C C . . . . . . - - - . T T T T A T T . - . T T - T T - A A
```

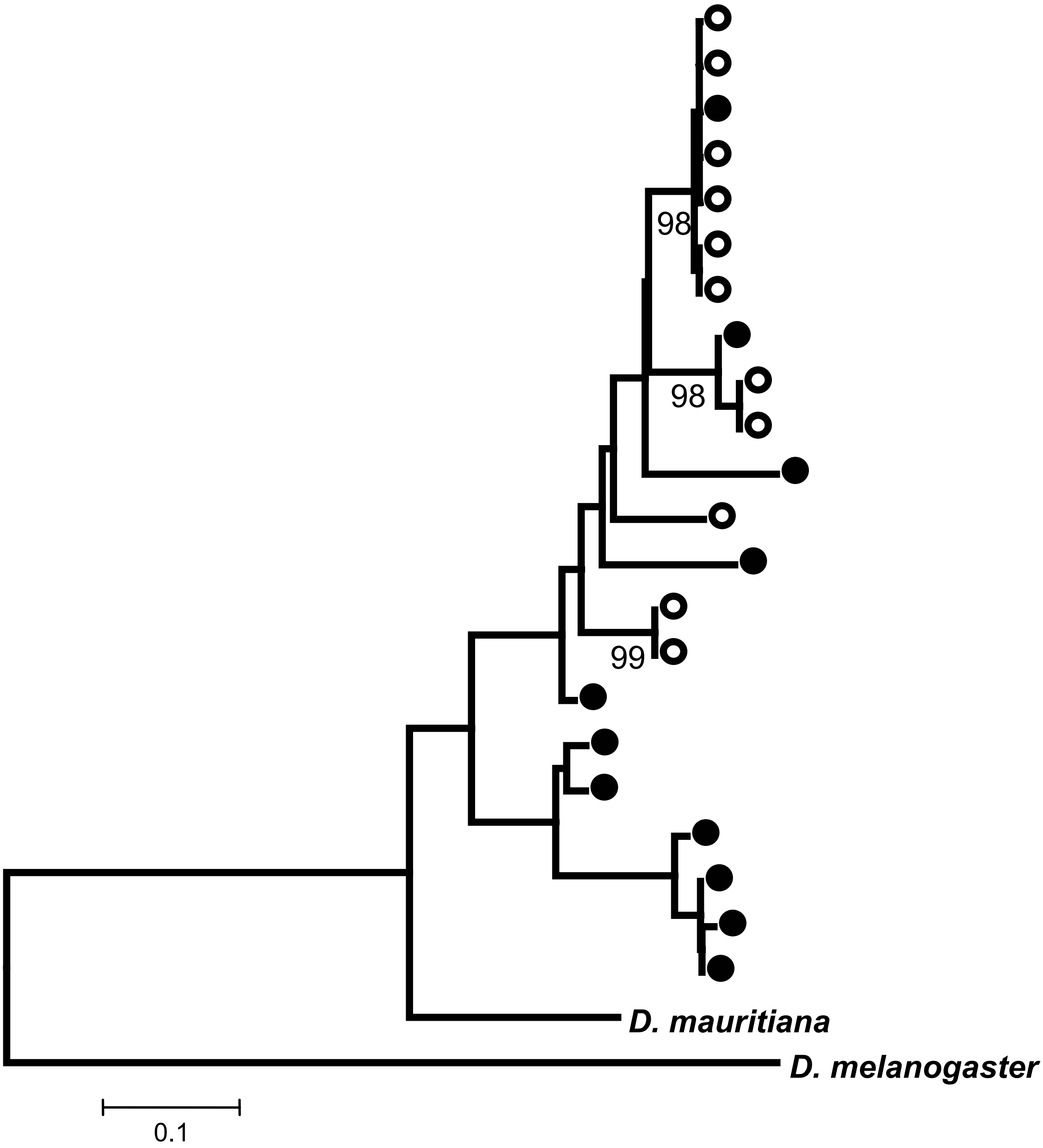
```
      743 745 749 750 751 758 760 762 765 772 773;i4 777;d17 779 781 792 793 794 795;d2 804 810 812 832 835 848 853 855 857 859 861 864 865 868 911 969 990 991 1048 1062 1162 1176 1221 1228 1234 1250 1277 1301 1306 1340 1375 1408 1418 1523 1558 1653 1735

S12   — A A T T A A G T - - G T A T A - d T - T A T A T A A A C A C A G G A G A G T G G C C T C A A A T C G G A C G
S36   . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . . . . . . . . .
S22   . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . . . . . . . . .
S2    . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
S38   . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . . . . . . . . .
S9    . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . . . . . . . . .
S18   . . . . . . . . A . . . . . . . . . . . A T . . C . . . . G . . . A C A . C . . . . . C . . . . . . . . . . T
S28   . . . . . . . . A . . . . . . . . . . . A T . . C . . . . G . . . A C A . C . . . . . C . . . . . . . . . . T
S50   . . . . . . . . . . . . . . . . . . . . . T . . C . . . . G . . . . . . . C . . . . . C . . T G . . A T . . .
S13   . . . . . . G . . . . . . . . . . . . . . T . . . C . . . . . . . . C A T C . . . . . C . . . . . . . . . A .
S40   . . . . . . . . . . . . . . . . . . . . . T . . C . . . . G . . . . . . . C . . . . . C . . T G . . A T . . .

MZ22  . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . . . . . . . C . . A T T . . . T . . A . . . . .
MZ19  . . . . . . . . . . . . . . . . . . . . . T . . . C . . . . . . T . C A . C . . . . . . . . . . . . . . . . T
MZ54  . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . T . . . . . . . . . . . . . . . . .
MZ33  . . . . . . . . . . . . . . . . . . . . A T . . C . . . . G . . . . C A . C C . . . . . . . T . . . . . . . .
MZ10  . . . . . . . . . . . . . . . . . . . . . T . . C . . . . G . . . . . . . C . . . . . C . . T G . . . . . . .
MZ18  C C C A A C . A . T i d——————— G A C . T G G C . G G T G T C . . C A . C . . . . . C . . . . C . . . . . .
MZ34  C C C A A C . A . T i d——————— G A C . T G G C . G G T G T C . . C A . C . T T . . . T T . . . . . . - . .
MZ16  C . C A A C . A . T - . C A T A G . C . T G G C . G G T G T C . . . . . C . . . . . C . . . . C . . . . . .
MZ2B  C . C A A C . A . T - . C A T A G . C . T G G C . G G T G T C . A C A . C . . . . . C . . . . . . . . . . .
MZ58  C C C A A C . A . T i d——————— G A C . T G G C . G G T G T C . . C A . C . T T . . . T . . . . . . . - . .
MZ12  C C C A A C . A . T i d——————— G A C . T G G C . G G T G T C . . C A . C . T T . . . T . . . . . . . - . .

D. mau G C G . A C . . . . . . . . A T T C . C . T G . C . . G . G . . T . C A . C . . . . . . . . T . . . . . . . .
D. mel G C G . A C . A . . . . A C A T T G . . - T C . C . . G T G . . . . C A . C . . A . . . . . T . . . . . . . .
```

FIGURE 6



Capítulo 3.3. Patterns of nucleotide and chromosomal polymorphism at the *OS-E* and *OS-F* genes of *Drosophila subobscura*

3.3.1 Resumen

En el presente artículo se investiga la variación nucleotídica de la región que incluye los genes *OS-E* y *OS-F* en dos ordenaciones cromosómicas ($O_{[3+4]}$ y O_{3+4+23}), presentes en dos poblaciones naturales de *Drosophila subobscura*. Como los genes se encuentran dentro del fragmento invertido su localización difiere entre las dos ordenaciones cromosómicas y la recombinación se puede ver reducida en los heterocariotipos. El análisis de los niveles y patrón del polimorfismo de DNA son consistentes con un origen monofilético de la inversión. A pesar de de que las dos ordenaciones cromosómicas están fuertemente diferenciadas, existen evidencias de que ha habido flujo genético entre las mismas, probablemente provocado por eventos de conversión génica. Los tests basados en la distribución de las frecuencias de las mutaciones indican que las ordenaciones cromosómicas no se encuentran en equilibrio, sino que probablemente aún reflejan el efecto de la expansión poblacional que se produjo después del origen de la inversión. Bajo esta hipótesis se ha estimado que la inversión O_{23} se originó hace unos 0.25 millones de años.

Artículo III

Patterns of nucleotide and chromosomal polymorphism at the *OS-E* and *OS-F* genes of *Drosophila subobscura*

Alejandro Sánchez-Gracia and Julio Rozas

2005

(En preparación)

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona,

Barcelona, Spain

INTRODUCTION

Chromosomal inversion polymorphism is a common feature in the genus Drosophila, and likely one of the best studied systems in population genetics. Three-quarters species of the genus harbor polymorphic inversions in natural populations, and 60% of them are paracentric, i.e. the inversion does not include the centromere (POWELL 1997). There are strong evidences supporting the adaptive character of the inversion polymorphism: DOBZHANSKY (1948, 1950) population cage experiments, and the correlation between latitude and the frequency of chromosomal arrangements detected both in the Old World and in the recently colonized areas of North and South America (PREVOSTI *et al*. 1988; KRIMBAS and POWELL 1992) clearly supports this adaptive character. The genetic content of the inversion, and interactions among loci included in the inverted fragment, would likely play a major role in the maintenance and evolution of chromosomal polymorphism. However, it is not fully understood the specific selective mechanism. There are two major hypotheses: i) the coadaptation hypothesis (DOBZHANSKY 1948, 1950) that proposes that a favorable combination of genes would act harmoniously (epistatic interactions) within a local population; in this sense, the study of SCHAEFFER *et al*. (2003), showing evidences for epistasis among genes included in *D. pseudoobscura* inversions, will support this hypothesis. On the other hand, ii) the position effect model suggests that the inversion breakpoints would alter the expression patterns -or the function- of the genes close or within the breakpoints. A putative example for this mechanism has been recently reported in a *D. buzzati* chromosomal inversion (PUIG *et al*. 2004).

In the last years a number of studies have analyzed the pattern at nucleotide variation in genes included in chromosomal inversions (e.g. AGUADÉ 1988; ROZAS and AGUADÉ 1990, 1993, 1994; AQUADRO *et al.* 1991; BENASSI *et al.* 1993; POPADIC and ANDERSON 1994, 1995; WESLEY and EANES 1994; POPADIC *et al.* 1995; BABCOCK and ANDERSON 1996; HASSON and EANES 1996; ROZAS *et al.* 1999). These surveys have shown an extensive genetic differentiation between gene arrangements, which is consistent with the reduction in the recombination levels expected in inversion heterozygotes (ROBERTS 1976). Nevertheless, some extent of genetic exchange (either by gene conversion or double crossover) between different gene arrangements has also been detected (ROZAS and AGUADÉ 1990, 1994; ROZAS *et al*. 1999). In *D. subobscura*, the pattern of nucleotide polymorphism in inversion-included genes is consistent with a monophyletic origin of the inversions (ROZAS and AGUADÉ 1990, 1993, 1994; ROZAS *et al*. 1999). These studies have also reported deviations of the nucleotide polymorphism distribution pattern expected for a constant-size population, which likely reflects the population (gene arrangement) expansion process caused by the rapid raise of a new gene arrangement to its equilibrium frequency. Theoretical and empirical studies have suggested that the level of genetic differentiation should be weaker in the central part of the inversion loop than close to the breakpoints (NAVARRO *et al*. 1997; ROZAS *et al*. 1999). Genetic differentiation would decay in function of the genetic exchange rate among arrangements. While the rate of gene conversion might be uniform along the inversion loop, the contribution of double crossovers to genetic exchange would be considerably higher in the central part of the inversion loop. Recently, however, MUNTÉ *et al*. (2005) have showed, in a particular inversion system, a strong genetic differentiation between arrangements, which extends along large chromosomal distances, even affecting the central part of the inversion. The comparative analysis of the nucleotide and chromosomal variation in a genomic region included in a particular inversion could provide, therefore, valuable information about the origin, age and the evolutionary fate of inversion polymorphism systems.

Here, we have studied nucleotide variation at the *OS* region of *D. subobscura* in two different chromosomal arrangements from two populations. In *D. melanogaster*, the *OS* region is located close to the centromere of the 3R chromosomal arm, and includes two odorant binding protein (OBP) genes, *OS-E* and *OS-F* with their intergenic region (figure 1). In *D. subobscura* this region maps on the O chromosome, which is the longest and most polymorphic chromosome for inversions in this species (the O chromosome is homolog to the 3R chromosome arm of *D. melanogaster*). In particular, the *OS*-region is located close to one of the breakpoints of the inversion O_{23}. Here, we have analyzed the patterns of nucleotide variation at the *OS-E* and *OS-F* genes in *D. subobscura* in two different chromosomal inversions. We found that nucleotide variation is clearly not at the steady-state equilibrium, but reflects the expansion process caused by the increase in frequency of the inversion. We used this DNA polymorphism data to study the origin, age and fate of this chromosomal polymorphism system.

MATERIAL AND METHODS

Fly samples: The complete *OS*-region was sequenced in 29 *D. subobscura* isochromosomal lines (15 from El Pedroso, Spain and 14 from Bizerte, Tunisia) previously reported in ROZAS and AGUADÉ 1994 and in ROZAS *et al*. 1999). For each line, the chromosomal gene arrangement had been previously determined (ROZAS *et al*. 1999). The cytological location of the *OS*-region in *D. subobscura* was determined by *in situ* hybridization using a modification of the MONTGOMERY *et al.* (1987) protocol (SEGARRA and AGUADÉ, 1992). Present study also includes one highly inbreed line (after 10 generations of sib mating) of *D. guanche* (kindly provided by G. PERIQUET).

DNA sequencing: Genomic DNA from El Pedroso lines was obtained following KREITMAN and AGUADÉ (1986), while that from Bizerte and of *D. guanche* was extracted by using a modification of protocol 48 from ASHBURNER (1989). In *D. subobscura* a ~4.5 kb region, including the complete *OS-E* coding region and part of the *OS-F* gene, was amplified by PCR (SAIKI *et al*. 1988), using oligonucleotides designed in conserved regions among three species of the melanogaster subgroup of Drosophila (SÁNCHEZ-GRACIA *et al*. 2003, and chapter 3.2). The 3' fragment of the *OS-F* gene was obtained by the inverse-PCR technique (OCHMAN *et al.* 1988). The amplified fragments were cycle-sequenced, using several oligonucleotides designed at intervals of ~400 nucleotides, and separated on a Perkin-Elmer (Norwalk, CT) ABI PRISM 3700 automated DNA sequencer following the manufacture's instructions. For each line, the sequence of both strands was determined. The sequence data from this work have been deposited in the EMBL, Gen Bank, and DDBJ Nucleotide Sequence Databases under the accession numbers XXXXXXXX-XXXXXXXX.

Data analysis: The *OS* nucleotide sequences were multiple aligned by using the Clustal W program (THOMPSON *et al*. 1994). The initial alignments were further optimized using the MacClade program version 3.06 (MADDISON and MADDISON 1992). The DnaSP program version 4.10.3 (ROZAS *et al*. 2003) was used to estimate nucleotide diversity, genetic distances, to detect putative genetic differentiation between populations and gene conversion tracts, and to conduct neutrality tests.

Nucleotide diversity, π, was estimated as the number of nucleotide differences per site (NEI 1987), while nucleotide divergence, *K*, was estimated as the number of nucleotide substitutions per site using JUKES and CANTOR (1969) correction. Estimates of silent nucleotide variation include variability both in the noncoding part as in the synonymous sites of the coding region. DNA divergence between gene arrangements was calculated as D_{xy}, the per site average number of nucleotide substitutions between populations (chromosomal arrangements in our case), and as D_a, the net number of nucleotide substitutions per site between populations (NEI 1987).

The phylogenetic analysis was performed by the neighbor-joining algorithm (SAITOU and NEI 1997) implemented in MEGA version 3 program (KUMAR et al. 2004); bootstrap values were based on 1000 replicates. For the analysis, the DNA sequence of the *OS*-region of *D. guanche* was used as outgroup.

The proportion of nucleotide diversity that is attributable to variation between populations, F_{st}, was estimated following HUDSON *et al*. (1992). The average level of gene flow, N_m, was estimated from F_{st}, assuming the island model with populations at migration-drift equilibrium (WRIGHT 1951; HUDSON *et al*. 1992). The S_{nn} statistic (HUDSON 2000) was computed to test for genetic differentiation between gene arrangements; their statistical significance was assessed using the permutation test based on 1000 replicates. We used the R_2 statistic (RAMOS-ONSINS and ROZAS 2002), which is based on the difference between the number of singleton mutations in a particular DNA sequence and the average nucleotide diversity, for detecting putative recent past population growth events.

We applied the MCDONALD and KREITMAN (1991) test to examine whether the number of polymorphic and fixed synonymous and nonsynoymous substitutions conforms -or deviates- from the neutral theoretical expectations. We also used Tajima's *D* (TAJIMA 1989), Fu and Li's *D* and *F* (FU and LI 1993), Fu's F_S (FU 1997) statistics to test for deviations in the distributon of intraespecific nucleotide variation. The statistical significance, including the confidence intervals of these test-statistics, was obtained by computer simulations (10000 replicates) based on the coalescent process with no recombination (HUDSON 1990). We also performed coalescent simulations with recombination for the within-arrangement analysis. These computer simulations were conducted assuming a conservative value (C_L) of the population recombination parameter *C* ($C = 4N_ec$, where N_e is the effective population size and *c* is the recombination rate per generation for the complete *OS* region); C_L (ROZAS *et al*. 2001) represents, the minimum value of C compatible (at 5%) with the estimated minimum number of recombination events (R_M; HUDSON and KAPLAN 1985). We applied the algorithm described in BETRÁN *et al*. (1997), which is based in the ψ parameter, to identify putative gene conversion tracts in the sample.

RESULTS

Cytological location of the *OS*-region: The *in situ* hybridization on politene chromosomes showed a unique signal located on cytological band 98D at the O chromosome of *D. subobscura* (figure 1). This chromosomal position lies within paracentric inversion O_{23}, which define two major chromosomal classes in the sample, the $O_{[3+4]}$ (including O_{3+4} and O_{3+4+8}) and the O_{3+4+23} (figure 1), where the recombination for the *OS* region is restricted.

Overall nucleotide variation: The surveyed region extends over 5808 bp (5202 bp excluding sites with alignment gaps). The structure of the *OS* genomic region is nearly the same than in *D. melanogaster*; the major difference is in the the size of the intergenic region, ~1kb in *D. melanogaster* and ~2.5kb in *D. subobscura.* Table 1 summarizes the nucleotide polymorphism and divergence levels in the different functional parts of the *OS*-region. We detected 377 segregating sites (representing a minimum of 394 mutations) among the 29 lines of *D. subobscura*; eight of these polymorphisms were nonsynonymous (figure S1). The average nucleotide diversity (π_T) was 0.0112. Figure 2 shows the distribution of silent nucleotide polymorphism and divergence (between *D. subobscura* and *D. guanche*) along the *OS*-region. As detected in SÁNCHEZ-GRACIA et al. (2003), the level of silent variation in the *OS-F* gene was lower than in the *OS-E* and intergenic regions, except for a ~300 bp fragment located in the first intron of the former gene. Polymorphism and divergence, however, are well correlated, as expected under the neutral model. Nevertheless, polymorphic and fixed synonymous and nonsynonymous mutations are uncorrelated (MK test, P = 0.03), being the *OS-E* gene the main responsible for this significant departure (*OS-E*, P = 0.005; *OS-F*, P = 1.000) (table 2).

Nucleotide variation and chromosomal arrangements: Table 3 shows a summary of the genetic differentiation between $O_{[3+4]}$ and O_{3+4+23} classes. The D_{xy} and D_a values (D_{xy} = 0.0125; D_a = 0.0028) indicate a high DNA divergence between arrangements; even so there were 50 shared mutations between arrangements, with no fixed difference between them. We investigate whether the two chromosomal arrangements are genetically differentiated by using the S_{nn} test statistic (HUDSON 2000). The results clearly indicate that the two chromosomal classes are highly differentiated (S_{nn} = 0.965; P < 0.001). We also examined if the number of shared mutations observed between arrangements could have arisen independently in each arrangement or –on the contrary- these mutations would be incorporated by some form of gene flow. For the analysis we assumed a homogeneous mutation rate across silent sites. Results indicated that the observed number of shared mutations are extremely unlikely under the parallel mutation scenario (the expected number of shared polymorphism is 2.67±1.6, P < 0.001). Likely this number of shared polymorphisms should be explained by gene flow between chromosomal classes; in fact, the analysis using BETRÁN *et al*. (1997) algorithm identifies 6 putative gene conversion tracts, with length sizes ranging from 2 to 397 bp (figure S1).

Figure 3 shows the phylogenetic tree for the complete *OS*-region, using the *D. guanche* sequence as outgroup. The sequences form two clear separated clusters, which corresponds with the two chromosomal classes. This topology, clearly indicate the monophyletic origin of the O_{23} chromosomal inversion. Nevertheless, there was a DNA sequence (line TB132) that is located out of the two groups. This sequence, likely, could have incorporated information from other chromosomal inversions (see below).

Table 4 summarizes the nucleotide polymorphism at the *OS*-region in *D. subobscura*, and the divergence with *D. guanche*, separately for each gene arrangement. Levels of silent polymorphism are higher in the $O_{[3+4]}$ (π_{SIL} = 0.0141) than in the O_{3+4+23} (π_{SIL} = 0.0085) sequences. Nevertheless, the pattern of silent diversity along the *OS* region is similar in the two chromosomal classes, and it is also well correlated with that pattern of silent divergence (figure 2). Both chromosomal arrangements present negative TAJIMA'S *D* values, although not significant. The neutral model, however, close to the critical value in both chromosomal classes using FU and LI's *D* and *F* tests (*P* ~0.05, under the conservative assumption of no intra-chromosomal class recombination). If we use the conservative C_L value in the coalescent simulation model, all tests are significant in both gene arrangements (P < 0.05). Since these values might be inflated by gene conversion events, we also estimated TAJIMA'S *D* and Fu and Li's *D* and *F* values after subtracting all nucleotide variants that might have been incorporated in gene conversion tracts; the results, nevertheless, were also significant (results not shown).

We examined the pairwise number of differences (by the mismatch distribution) at the *OS*-region independently for each gene arrangement. Both chromosomal classes showed a similar distribution (figure 4), close to that expected under the expansion model (SLATKIN and HUDSON 1991; ROGERS and HARPENDING 1992). It is known (ROZAS *et al.* 1999), however, that recombination could also generate the mismatch growth distribution in a constant-size population. To circumvent this confounding effect, we applied the R_2 and F_S statistics which likely are less sensitive to recombination than the mismatch distribution-based tests. We obtained significant low R_2 and F_S values (table 4), as expected from a recent severe population growth event.

Since the *OS*-region maps within the inversion 23, we can use DNA polymorphism information to date the origin of this inversion. For the analysis we assumed that this inversion is monophyletic, and that all variation observed within the arrangement originated in this arrangement after the expansion process that followed the origin of the inversion. For such analysis we subtracted all nucleotide variation likely incorporated by gene conversion events, and i) assume that the divergence time between *D. guanche* and *D. subobscura* is 1.8 to 2.8 myr (RAMOS-ONSINS *et al*. 1998), ii) that silent divergence between these species is K_{SIL} = 0.079, and therefore, silent nucleotide substitution per site and per year at the *OS*-region would be $\lambda = 2.19 \times 10^{-8}$ or $\lambda = 1.41 \times 10^{-8}$, respectively. Following the approximation of ROZAS *et al*. (1999) using current silent polymorphism estimates, π_{SIL}= 0.0082, the time for the origin of inversion O_{23} should be about 0.19 or 0.29 millions of years.

DISCUSSION

Our interest in the *OS*-region in *D. subobscura* stemmed from previous data in *D. melanogaster* and *D. simulans* suggesting that variation at this region might be shuffled by positive natural selection. In addition, the cytological location of this region in *D. subobscura* allowed the analysis of the nucleotide variation related to the chromosomal polymorphism in this species.

Silent nucleotide diversity at the *OS*-region of *D. subobscura* was higher than that observed in the European populations of *D. melanogaster* and *D. simulans* (SÁNCHEZ-GRACIA *et al*. 2003; chapter 4). The high number of DNA polymorphisms, and the detected correlation between polymorphism and divergence -as expected under the neutral model- allow us to conduct a fine analysis of the functional constraints along the *OS*-region. Levels of silent nucleotide variation at the *OS-E* gene, and the intergenic region, were higher than that observed at the coding region of the *OS-F* gene and in some parts of their first large intron (figure 2). A similar pattern was already observed in the DNA variation analysis of *D. melanogaster* and *D. simulans* species. Therefore, the results of *D. subobscura* corroborates the higher evolutionary rate of the *OS-E* gene, and the presence of at least two non coding extremely conserved regions in the large intron of the *OS-F* gene. These regions likely should constitute important regulatory regions, and deserves further investigation.

Results of the MK test show a significant excess of fixed nonsynonymous substitutions in the *OS-E* gene between *D. subobscura* and *D. guanche*.

However, when we repeated the MK analysis using the divergence data of *D. madeirensis*, we did not found any departure form the neutral expectations (see also chapter 3.4). Therefore, the significant results of the MK test might be attributable to an excess of nonsynonymous fixations in the *D. guanche* lineage. These amino acid changes might be advantageous, and their fixation could have been promoted by positive selection. Nevertheless, the reduced effective population size of the *D. guanche* (LLOPART and AGUADÉ 1999; PÉREZ *et al.* 2003) might have increased the fixation probability of nearly neutral mutations (as unpreferent synonymous mutations or amino acid replacements), due to a relaxation in the strength of selection (OHTA and KIMURA 1971; OHTA 1972). Only under the positive selection scenario, however, the number of nonsynonymous substitutions per nonsynonysmous site (K_A) can be higher than the number of synonymous substitutions per synonymous site (K_S), i.e., K_A/K_S >1. At the *OS-E* gene of the *D. guanche* lineage, on the contrary, the synonymous rate (K_S = 0.083) is higher than the nonsynonymous rate (K_A= 0.024) and, thus, we can not discard that the fixed amino acids were nearly neutral mutations. Therefore, there are not patent evidences for positive selection in the whole analysis of the gene. Nevertheless, it is possible that positive selection would act only on a few amino acid positions; in this case, testing for positive selection using estimates of synonymous and nonsynonymous substitutions at the whole gene would be highly conservative. This possibility might be investigated by using different approaches, as the ML analysis using codon evolution models that account for heterogeneity in the distribution of functional constraints along the sequence, which have much more power to detect the footprint of natural selection (see chapter 3.4; YANG *et al*. 2000; ANISIMOVA *et al.* 2001, 2002).

Here we have detected a significant genetic differentiation between gene arrangements, as expected by the suppression of recombination in heterocariotypes. This suppression, however, is not complete; in fact the number of shared substitutions between arrangements is higher than that expected for an independent accumulation of mutations. That is, the reduction in recombination occurs in spite of some forms of gene flow (double crossing over or gene conversion) between gene arrangements. Actually, we identified several gene conversion tracts between chromosomal classes, confirming that this mechanism is involved in the genetic exchange among polymorphic chromosome inversions (ROZAS and AGUADÉ 1994; NAVARRO-SABATÉ *et al*. 1999; ROZAS *et al*. 1999; MUNTÉ *et al.* 2000). Nevertheless, the phylogenetic tree of the *OS*-region clearly shows two differentiated groups, as expected by the major role of the suppression of recombination between gene arrangements, and supports the monophyletic origin of this inversion. These results are in concordance with other nucleotide variation studies on the same polymorphic inversion system (ROZAS *et al*. 1999). It should be noticed, however, that one of the $O_{[3+4]}$ sequences is not within their chromosomal group in the tree. Although we have not identified any gene conversion tract in this sequence, it might have incorporate genetic information from other gene arrangements not analyzed in the present study.

The rapid increase in the frequency of an inversion, from their origin (monophyletic) to its equilibrium frequency can be envisaged as a population growth event; therefore, it could leave recognizable signatures in the pattern of DNA molecular diversity for some periods of time (ROGERS and HARPENDING 1992; HARPENDING 1994). Population growths lead to star-shaped genealogies and unimodal (Poisson-like) distributions of the pairwise number of differences (mismatch distribution). In the present study, the shape of mismatch distribution (in both gene arrangements) fit well with the Poisson-like distribution (figure 4). However, intragenic recombination, by shuffling nucleotide variation among DNA sequences, can also generate smooth mismatch distributions (Poisson-like) on a constant-size population. The R_2 and F_s coalescent-based neutrality statistical tests are more powerful than other tests in detecting population growth events, and are less sensitive to recombination than mismatch distribution-based statistics (RAMOS-ONSINS and ROZAS 2002). In addition, for small sample sizes, as is our case, the R_2 statistic is even more powerful than F_s (RAMOS-ONSINS and ROZAS 2002). The analysis showed negative values of these statistics, as expected by a population (chromosomal arrangement) growth.

Consequently, gene arrangements are not in steady-state equilibrium and still reflect the expansion event. Under this scenario we estimated that the time to origin of the inversion would be about 0.19 or 0.29 Myr. Since an inversion would need at least 10^7 generations to reach their equilibrium frequency (NAVARRO *et al*. 2000), (in *D. subobscura* would correspond to 2 myr; five generations per year, ASHBURNER 1989; POWELL 1997), present estimates of the times for the origin of the inversion O_{23} are consistent with a non-equilibrium state of the O_{3+4+23} chromosomal arrangement. These times are slightly lower than those estimated for other inversions of the O chromosome, but agree with the evolutionary history of the O chromosome inversion polymorphism system of *D. subobscura* (ROZAS *et al*. 1999).

ACKNOWLEDGEMENTS

We thank CARMEN SEGARRA for her valuable contribution in performing the *in situ* hybridization.

LITERATURE CITED

AGUADÉ, M., 1988 Restriction map variation at the *Adh* locus of *Drosophila melanogaster* in inverted and noninverted chromosomes. Genetics 119: 135-140.

ANISIMOVA, M., J. P. BIELAWSKI and Z. YANG, 2001 Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18: 1585-1592.

ANISIMOVA, M., J. P. BIELAWSKI and Z. YANG, 2002 Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19: 950-958.

AQUADRO, C. F., A. L. WEAVER, S. W. SCHAEFFER and W. W. ANDERSON, 1991 Molecular evolution of inversions in *Drosophila pseudoobscura*: the amylase gene region. Proc Natl Acad Sci U S A 88: 305-309.

ASHBURNER, M., 1989 *Drosophila: A laboratory handbook*. Cold Spring Harbor Laboratory Press, New York.

BABCOCK, C. S., and W. W. ANDERSON, 1996 Molecular evolution of the Sex-Ratio inversion complex in *Drosophila pseudoobscura*: analysis of the *Esterase-5* gene region. Mol Biol Evol 13: 297-308.

BENASSI, V., S. AULARD, S. MAZEAU and M. VEUILLE, 1993 Molecular variation of *Adh* and *P6* genes in an African population of *Drosophila melanogast*er and its relation to chromosomal inversions. Genetics 134: 789-799.

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.

DOBZHANSKY, T., 1948 Genetics of natural populations. XVIII. Experiments on chormosomes of *Drosophila pseudoobscura* from differents geographic regions. Genetics 33: 588-602.

DOBZHANSKY, T., 1950 Genetics of natural populations. XIX. Origin of heterosis through natural selection in populations of *Drosophila pseudoobscura*. Genetics 35: 288-302.

FU, Y. X., and W. H. LI, 1993 Statistical tests of neutrality of mutations. Genetics 133: 693-709.

FU, Y. X., 1997 Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147: 915-925.

HARPENDING, H. C., 1994 Signature of ancient population growth in a low-resolution mitochondrial DNA mismatch distribution. Hum Biol 66: 591-600.

HASSON, E., and W. F. EANES, 1996 Contrasting histories of three gene regions associated with *In(3L)Payne* of *Drosophila melanogaster*. Genetics 144: 1565-1575.

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., 1990 Gene genealogies and the coalescent process, pp. 1–42 in *Oxford Surveys in Evolutionary Biology*, Vol. 7, edited by D. FUTUYMA and J. ANTONOVICS. Oxford University Press, New York.

HUDSON, R. R., M. SLATKIN and W. P. MADDISON, 1992 Estimation of levels of gene flow from DNA sequence data. Genetics 132: 583-589.

HUDSON, R. R., 2000 A new statistic for detecting genetic differentiation. Genetics 155: 2011-2014.

JUKES, T. H., and C. R. CANTOR, 1969 Evolution of protein molecules, pp. 21-123 in *Mammalian Protein Metabolism*, edited by H. N. MUNRO. Academic Press, New York.

KREITMAN, M., and M. AGUADÉ, 1986 Genetic uniformity in two populations of *Drosophila melanogaster* as revealed by filter hybridization of four-nucleotide-recognizing restriction enzyme digests. Proc Natl Acad Sci U S A 83: 3562-3566.

KRIMBAS, C. B., and J. R. POWELL, 1992 *Drosophila Inversion Polymorphism*. C. R. C. Press, Boca Raton, FL.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5: 150-163.

LLOPART, A., and M. AGUADÉ, 1999 Synonymous rates at the *RpII215* gene of Drosophila: variation among species and across the coding region. Genetics 152: 269-280.

MADDISON, W. P., and D. R. MADDISON, 1992 *MacClade: Analysis of Phylogeny and Character Evolution*. Version 3.0. Sinauer, Sunderland, MA.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature 351: 652-654.

MONTGOMERY, E., B. CHARLESWORTH and C. H. LANGLEY, 1987 A test for the role of natural selection in the stabilization of transposable element copy number in a population of *Drosophila melanogaster*. Genet Res 49: 31–41.

MUNTÉ, A., M. AGUADÉ and C. SEGARRA, 2000 Nucleotide variation at the *yellow* gene region is not reduced in *Drosophila subobscura*: a study in relation to chromosomal polymorphism. Mol Biol Evol 17: 1942-1955.

MUNTÉ, A., J. ROZAS, M. AGUADÉ and C. SEGARRA, 2005 Chromosomal inversion polymorphism leads to extensive genetic structure: a multilocus survey in *Drosophila subobscura*. Genetics 169: 1573-1581.

NAVARRO, A., BARBADILLA and A. RUIZ, 2000 Effect of inversion polymorphism on the neutral nucleotide variability of linked chromosomal regions in Drosophila. Genetics 155: 685-698.

NAVARRO-SABATÉ, A., M. AGUADÉ and C. SEGARRA, 1999 The relationship between allozyme and chromosomal polymorphism inferred from nucleotide variation at the *Acph-1* gene region of *Drosophila subobscura*. Genetics 153: 871-889.

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

OCHMAN, H., A. S. GERBER and D. L. HARTL, 1988 Genetic applications of an inverse polymerase chain reaction. Genetics 120: 621-623.

OHTA, T., and M. KIMURA, 1971 Behavior of neutral mutants influenced by asociated overdominant loci in finite populations. Genetics 69: 247-260.

OHTA, T., 1972 Evolutionary rate of cistrons and DNA divergence. J Mol Evol 1**:** 150-157.

PÉREZ, J. A., A. MUNTÉ, J. ROZAS, C. SEGARRA and M. AGUADÉ, 2003 Nucleotide polymorphism in the *RpII215* gene region of the insular species *Drosophila guanche*: reduced efficacy of weak selection on synonymous variation. Mol Biol Evol 20: 1867-1875.

POPADIC, A., and W. W. ANDERSON, 1994 The history of a genetic system. Proc Natl Acad Sci U S A 91: 6819-6823.

POPADIC, A., and W. W. ANDERSON, 1995 Evidence for gene conversion in the amylase multigene family of *Drosophila pseudoobscura*. Mol Biol Evol 12: 564-572.

POPADIC, A., D. POPADIC and W. W. ANDERSON, 1995 Interchromosomal exchange of genetic information between gene arrangements on the third chromosome of *Drosophila pseudoobscura*. Mol Biol Evol 12: 938-943.

POWELL, J. R., 1997 *Progress and prospects in evolutionary biology: the Drosophila model*. Oxford University Press, New York.

PREVOSTI, A., G. RIBÓ, L. SERRA, M. AGUADÉ, J. BALAÑÁ, M. MONCLÚS, and F. MESTRES, 1988 Colonization of America by *Drosophila subobscura*: Experiment in natural populations that supports the adaptive role of chromosomal-inversion polymorphism. Proc Natl Acad Sci U S A 85: 5597-5600.

PUIG, M., M. CACERES and A. RUIZ, 2004 Silencing of a gene adjacent to the breakpoint of a widespread Drosophila inversion by a transposon-induced antisense RNA. Proc Natl Acad Sci U S A 101: 9013-9018.

RAMOS-ONSINS, S., C. SEGARRA, J. ROZAS and M. AGUADÉ, 1998 Molecular and chromosomal phylogeny in the obscura group of Drosophila inferred from sequences of the *rp49* gene region. Mol Phylogenet Evol 9: 33-41.

RAMOS-ONSINS, S. E., and J. ROZAS, 2002 Statistical properties of new neutrality tests against population growth. Mol Biol Evol 19: 2092-2100.

ROBERTS, P. A., 1976 The genetics of chromosome aberration, pp. 67-184 in *The Genetics and Biology of Drosophila*, Vol. 1a, edited by M. ASHBURNER and E. NOVITSKI. Academic Press, London.

ROGERS, A. R., and H. HARPENDING, 1992 Population growth makes waves in the distribution of pairwise genetic differences. Mol Biol Evol 9: 552-569.

ROGERS, A. R., 1995 Genetic evidence for a pleistocene population. Evolution 49: 608-615.

ROZAS, J., and M. AGUADÉ, 1990 Evidence of extensive genetic exchange in the *rp49* region among polymorphic chromosome inversions in *Drosophila subobscura*. Genetics 126: 417-426.

ROZAS, J., and M. AGUADÉ, 1993 Transfer of genetic information in the *rp49* region of *Drosophila subobscura* between different chromosomal gene arrangements. Proc Natl Acad Sci U S A 90: 8083-8087.

ROZAS, J., and M. AGUADÉ, 1994 Gene conversion is involved in the transfer of genetic information between naturally occurring inversions of Drosophila. Proc Natl Acad Sci U S A 91: 11517-11521.

ROZAS, J., C. SEGARRA, G. RIBÓ and M. AGUADÉ, 1999 Molecular population genetics of the *rp49* gene region in different chromosomal inversions of *Drosophila subobscura*. Genetics 151: 189-202.

ROZAS, J., M. GULLAUD, G. BLANDIN and M. AGUADÉ, 2001 DNA Variation at the *rp49* gene region of *Drosophila simulans*: evolutionary inferences from an unusual haplotype structure. Genetics 158: 1147-1155.

ROZAS, J., J. C. SÁNCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI et al., 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487-491.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

SÁNCHEZ-GRACIA, A., M. AGUADÉ and J. ROZAS, 2003 Patterns of nucleotide polymorphism and divergence in the Odorant-Binding Protein genes *OS-E* and *OS-F*: analysis in the melanogaster species subgroup of Drosophila. Genetics 165: 1279-1288.

SCHAEFFER, S. W., M. P. GOETTING-MINESKY, M. KOVACEVIC, J. R. PEOPLES, J. L. GRAYBILL et al., 2003 Evolutionary genomics of inversions in *Drosophila pseudoobscura*: evidence for epistasis. Proc Natl Acad Sci U S A 100: 8319-8324.

SEGARRA, C., and M. AGUADÉ, 1992 Molecular organization of the X chromosome in different species of the obscura group of Drosophila. Genetics 130: 513-521.

SLATKIN, M., and R. R. HUDSON, 1991 Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics 129: 555-562.

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

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

WESLEY, C. S., and W. F. EANES, 1994 Isolation and analysis of the breakpoint sequences of chromosome inversion *In(3L)Payne* in *Drosophila melanogaster*. Proc Natl Acad Sci U S A 91: 3132-3136.

WRIGHT, S., 1951 The genetical structure of populations. Ann Eugen 15: 323-354.

YANG, Z., R. NIELSEN, N. GOLDMAN and A. M. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

TABLE 1

Summary of nucleotide variation at the *OS*-region in *D. subobscura*

Silent				
No. sites[a]	311.58	2155	2007.47	4474.05
η	25	219	142	386
π	0.0120	0.0139	0.0119	0.0129
θ	0.0204	0.0259	0.0180	0.0220
K	0.0784	0.0982	0.0578	0.0790
Synonymous				
No. sites[a]	89.58		106.47	196.05
η	14		12	26
π	0.0195		0.0196	0.0196
θ	0.0398		0.0287	0.0338
K	0.1351		0.0924	0.0112
Noncoding				
No. sites[a]	222	2155	1901	3778
η	11	219	130	360
π	0.0089	0.0139	0.0115	0.0126
θ	0.0126	0.0259	0.0174	0,0214
K	0.0565	0.0982	0.0555	0.0773

η, number of mutations; *K*, nucleotide divergence between *D. subobscura* and *D. guanche*. [a], Number of sites in polymorphism data set

TABLE 2

MK test

	Polymorphic		Fixed[a]		
	Syn.	*Nsyn.*	*Syn.*	*Nsyn.*	*P-value*[b]
$O_{[3+4]}$					
OS-E	9	1	8	12	*0.017*
OS-F	9	3	8	2	*1.000*
O_{3+4+23}					
OS-E	6	1	10	12	*0.093*
OS-F	4	3	7	2	*0.59*
Total					
OS-E	14	2	8	12	*0.005*
OS-F	12	5	7	2	*1.000*
OS region	26	7	15	14	*0.033*

Syn, synonymous mutations of the coding region; *Nsyn,* nonsynonymous mutations. [a] substitutions between *D. subobscura* and *D. guanche*. [b] Fisher`s exact test.

TABLE 3

Summary of genetic differentiation

	Shared	Fixed	D_{xy}	D_a	F_{ST}	S_{nn}	ψ
$\mathrm{O}_{[3+4]}$-O_{3+4+23}	50*[a]	0	0.0125	0.0028	0.224	0.965***	0.003

[a] based on the hypergeometric distribution

* 0.01 < P < 0.05, *** P < 0.001

TABLE 4

Nucleotide variation and neutrality tests at *OS*-region in different gene arrangements

	$O_{[3+4]}$	O_{3+4+23}
Sample size	15	14
η	296	154
No. silent sites	4535,12	4639.32
π_s	0.0141	0.0085
K_s	0.0795	0.0789
Tajima's *D*	-1.3385#	-0,8960
Fu and Li's *D*	-1.7717#	-1.7866*
Fu and Li's *F*	-2.0242#	-1.8703#
Fu's F_S	-1.273*	-1.917#
Ramos-Onsins and Rozas R_2	0.0741**	0.0872*

S, number of segregating sites; η, number of mutations; π_s, silent nucleotide diversity; K_s , silent nucleotide divergence between *D. subobscura* and *D. guanche.* # $0.05 < P < 0.10$; * $0.01 < P < 0.05$; ** $P < 0.01$.

FIGURE LEGENDS

Figure 1. (A) *In situ* hybridization on polytene chrormosomes of *Drosophila subobscura* using the complete *OS* region as biotinilated probe. The arrow indicates the hybridization signal. (B) Location of the *OS* and *rp49* regions in different chromosomal arrangements of *D. subobscura*. Shaded bars indicate the regions affected by inversions. O_3 gene arrangement is not present in extant populations of *D. subobscura.*

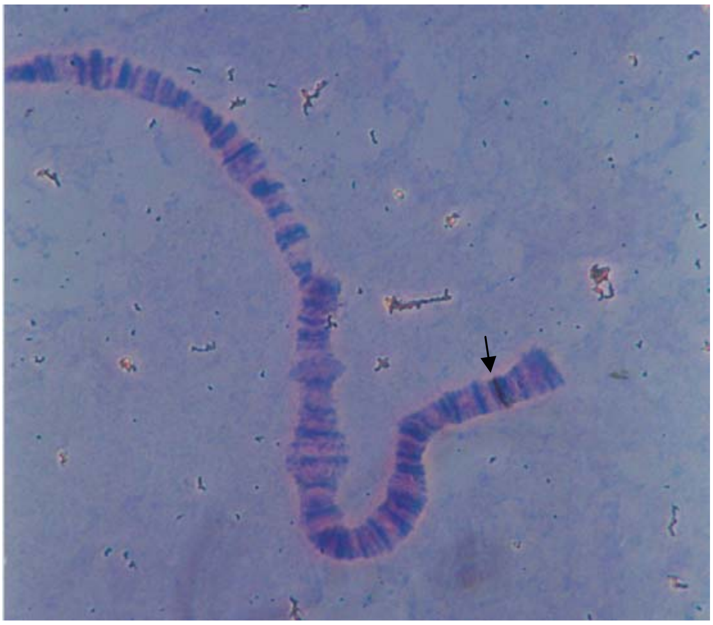
Figure 2. Sliding window of silent polymorphism in *D. subobscura*, and silent divergence between *D. subobscura* and *D. guanche* along the *OS*-region for the total *D. subobscura* data, and for within chromosomal arrangement.

Figure 3. Neighbour-joining tree of the *OS* region of the 29 sequences of *D. subobscura*. Italic numbers indicate the percentages of bootstrap replicates supporting the main nodes. The tree was built using silent nucleotide substitutions and rooted with the *D. guanche* sequence. The distance scale is indicated by a bar.

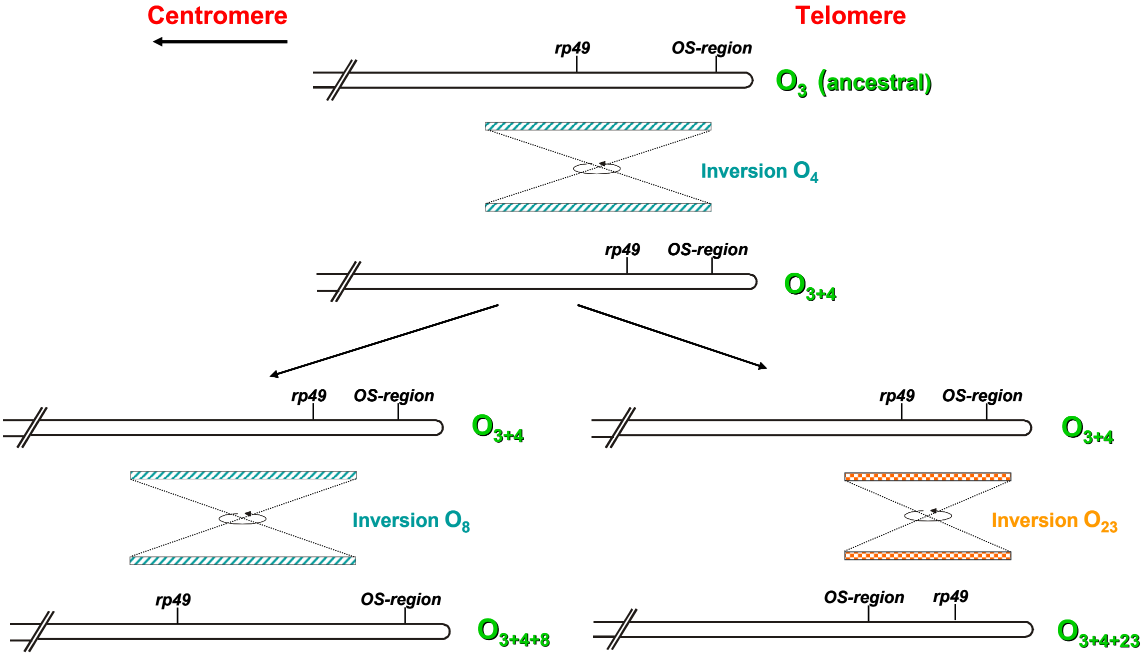
Figure 4. Distribution of the parwise nucleotide differences in the two gene arrangements. The expected distributions were obtained from Equation 3 in ROGERS 1995 and considering $\theta_0 = 0$.

FIGURE 1

A



B



Centromere
Telomere
rp49
OS-region
O3 (ancestral)
Inversion O4
O3+4
Inversion O8
Inversion O23
O3+4+8
O3+4+23

FIGURE 2

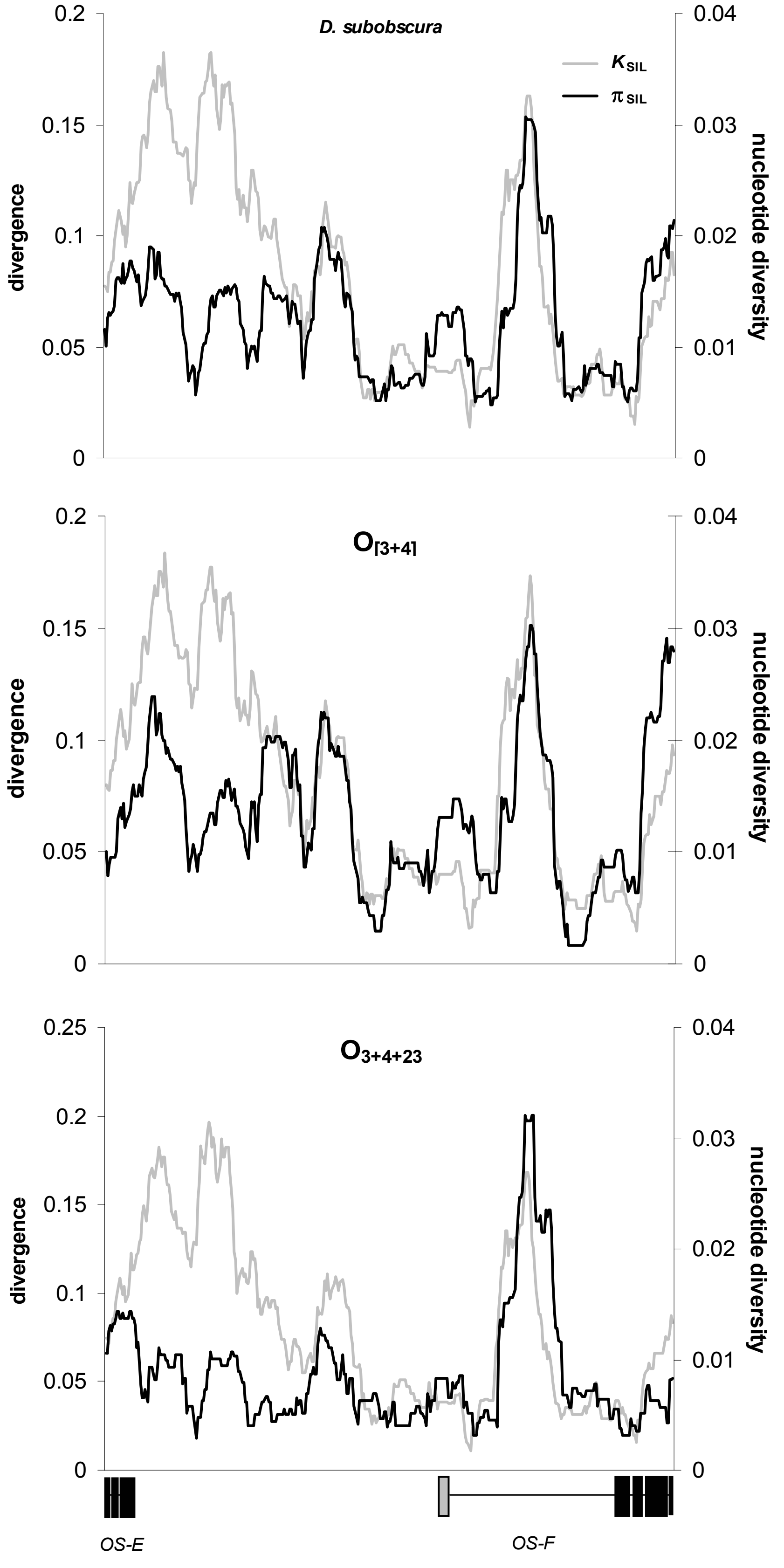

D. subobscura
K_{SIL}
π_{SIL}
divergence
nucleotide diversity
0.2
0.15
0.1
0.05
0
0.04
0.03
0.02
0.01
0
$O_{[3+4]}$
O_{3+4+23}
0.25
OS-E
OS-F

FIGURE 3

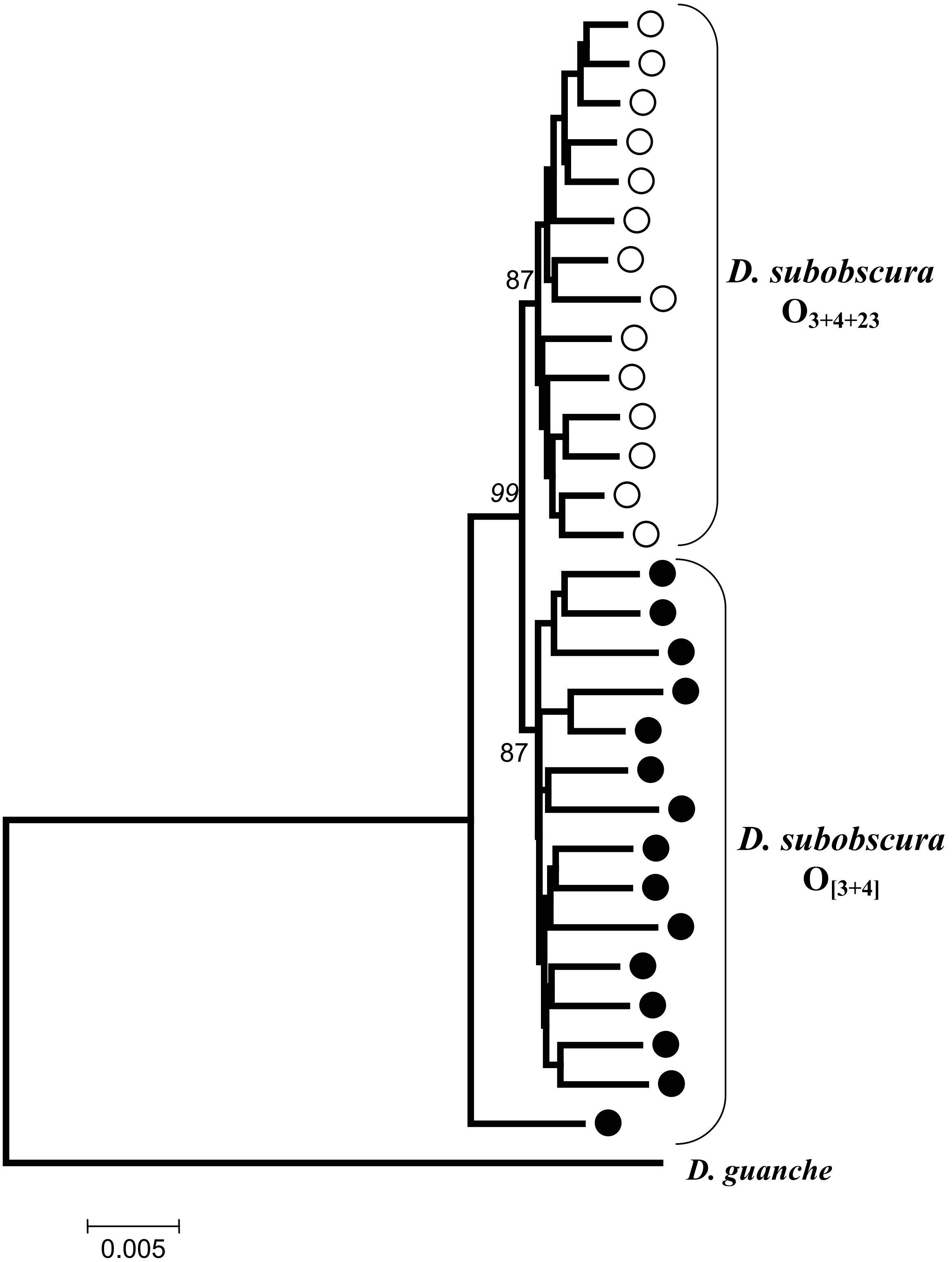
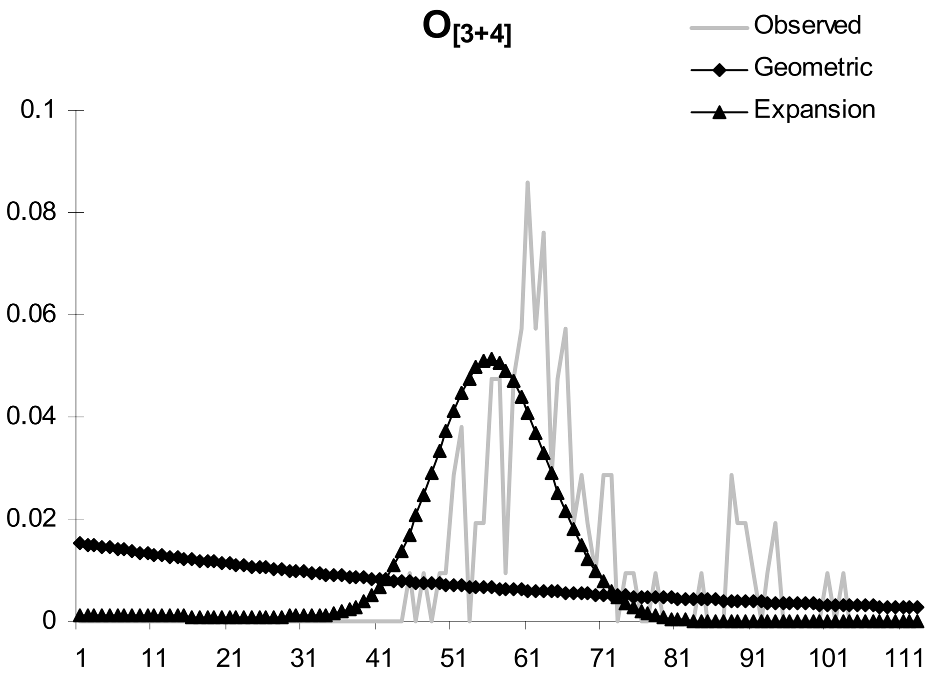
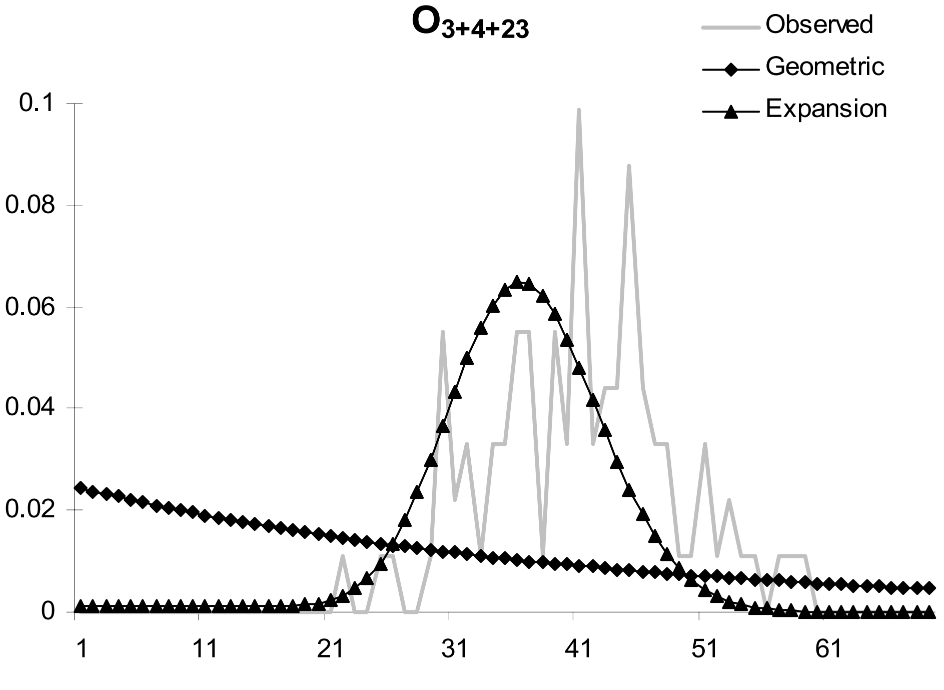


FIGURE 4


$O_{[3+4]}$
Observed
Geometric
Expansion
0.1
0.08
0.06
0.04
0.02
0
1
11
21
31
41
51
61
71
81
91
101
111

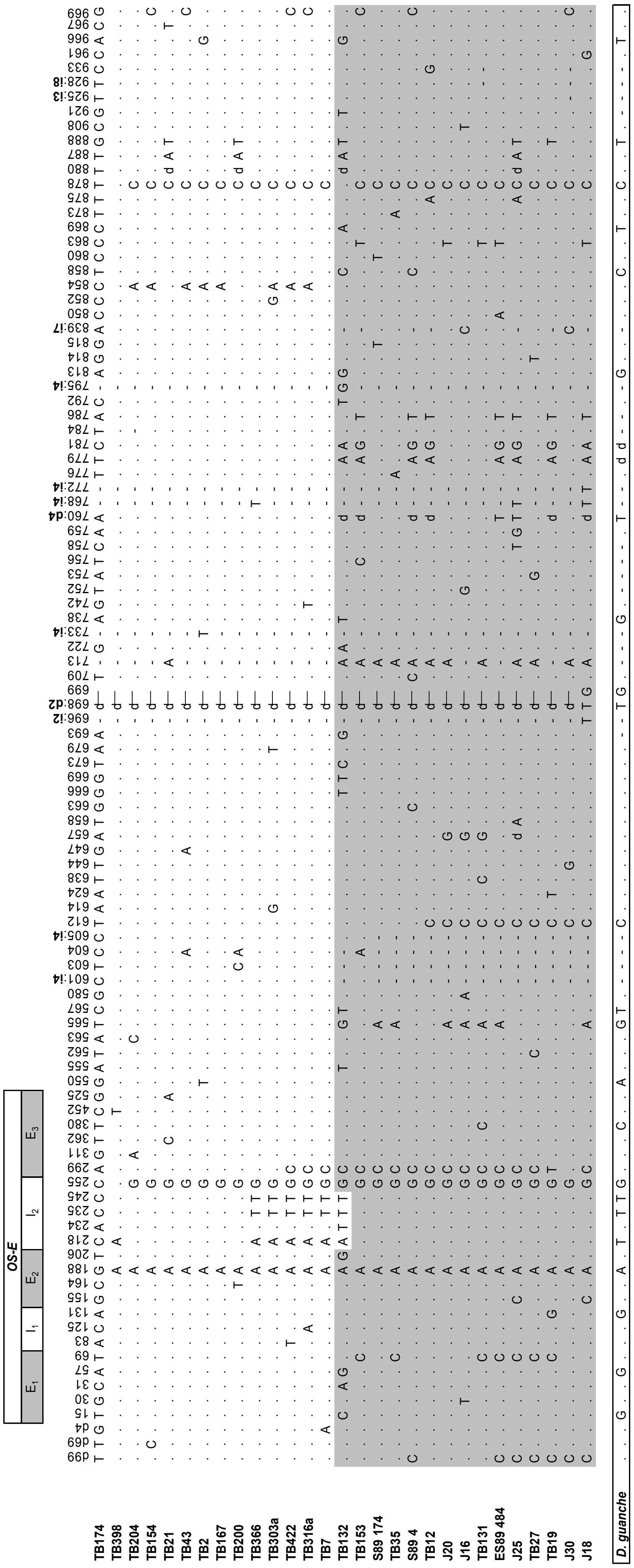

O_{3+4+23}
Observed
Geometric
Expansion
0.1
0.08
0.06
0.04
0.02
0
1
11
21
31
41
51
61

SUPPLEMENTARY MATERIAL

FIGURE LEGENDS

Figure S1. Nucleotide polymorphism at the *OS* region of *D. subobscura*. The last row shows the nucleotide variant present in *D. guanche* of these polymorphic sites. Nucleotides identical to the first sequence are indicated by a dot and deletions by a line. White and grey colors indicate information of O_{3+4+23} and $O_{[3+4]}$, respectively. Gene conversion tracts between gene arrangements are depicted with the color of the donor gene arrangement. d, deletion. i, insertion. E, exon. **I**, intron.

FIGURE S1


OS-E
E1
I1
E2
I2
E3
TB174
TB398
TB204
TB154
TB21
TB43
TB2
TB167
TB200
TB366
TB303a
TB422
TB316a
TB7
TB132
TB153
S89 174
TB35
S89 4
TB12
J20
J16
TB131
ES89 484
J25
TB27
TB19
J30
J18
D. guanche

	1032	1033	1035	1042	1071	1084	1086	1091	1109	1124:**i5**	1134	1144	1152	1217	1219	1231	1240	1248	1272	1275	1283	1291:**d2**	1292	1294	1307	1326	1327	1337	1339	1345	1350	1353	1357	1358	1361	1364	1370	1376	1380	1392	1403	1411	1415	1423:**d2**	1425:**i2**	1427:**i2**	1429:**i2**	1431:**i2**	1433:**i2**	1435:**i2**	1437:**i2**	1439:**i4**	1443:**i12**	1467	1472	1475	1481	1483	1489	1491	1534	1536	1550	1565	1586	1589	1596	1605	1610	1614	1622:**i2**	1624:**i5**	1635	1636:**i16**	1640	1643	1646	1647	1652	1653:**i12**	1665:**d7**	1672:**d8**	1684	1685	1686:**i2**	1688:**d9**	1692	1696	1710	1721	1756	1769	1795	1798	1803	1809	1829	1831	1832:**d5**	1838	1857
TB174	G	C	A	T	G	T	C	G	C	A	G	C	C	T	T	C	G	T	A	G	G	C	A	T	A	T	G	T	G	A	C	A	A	C	A	A	A	C	T	C	A	A	A	A	A	A	A	A	A	-	-	-	-	T	G	A	T	T	T	C	C	T	C	C	G	A	A	T	G	A	-	-	G	T	A	G	T	T	A	G	G	T	A	A	T	T	G	A	A	T	C	G	C	G	A	A	T	A	G	A	A
TB398	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB204	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	A	.	.	.	.	.	G	.	.	.	.	A	.	.	C	.	.	.	.	.	.	.	-	-	-	-	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	-	-	d—		-	-	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB154	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	A	C	.	.	.	.	.	.	.	-	-	-	-	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB21	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	.	.	.	.	.	.	A	A	-	-	.	.	.	G	.	.	.	.	.	.	.	.	T	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB43	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	C	C	.	.	.	.	.	.	.	A	-	-	-	.	.	.	C	.	.	.	.	.	A	.	.	.	.	.	.	.	-	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB2	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	-	-	-	-	-	-	-	-	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	G	.
TB167	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	.	.	.	.	-	-	-	-	-	-	.	.	.	G	.	.	.	.	.	.	G	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB200	.	.	.	.	A	.	T	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	-	-	-	-	-	-	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	C	A	T	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.
TB366	.	.	.	.	.	C	.	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	.	.	.	.	.	.	A	A	A	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB303a	.	.	.	.	.	.	.	.	.	-	d	.	A	G	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	.	.	.	.	.	-	-	-	-	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.
TB422	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	A	C	.	.	.	.	.	.	.	.	A	A	-	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB316a	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	A	.	.	.	.	.	G	.	.	.	.	A	.	.	C	.	.	.	.	.	.	.	-	-	-	-	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	-	-	d—		-	-	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB7	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.	A	.	.	C	.	.	.	.	.	.	.	.	-	-	-	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
TB132	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	T	T	A	.	.	.	.	.	A	.	C	A	G	.	G	.	.	.	.	.	.	.	.	A	.	C	.	.	d	-	-	-	-	-	-	-	-	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	T	-	.	.	T	.	A	A	.	-	d—		-	.	.	.	.	.	.	.	.	.	.	T	G	G	.	.	.	G	.
TB153	.	.	.	.	.	.	.	T	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	d	-	-	-	-	-	-	-	-	-	A	.	.	C	.	.	.	.	.	.	.	.	.	G	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	G	.	.	.	.	.	.	.	.	.	G	.
S89 174	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	-	-	-	-	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	G	.
TB35	A	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	A	.	.	.	T	.	.	.	.	.	.	.	.	.	C	.	.	d	-	-	-	-	-	-	-	-	-	A	.	.	C	.	.	.	.	C	.	.	.	.	.	.	.	.	-	-	.	.	.	T	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	d	.	G	.
S89 4	.	T	.	C	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	-	-	-	-	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.
TB12	.	.	.	.	.	.	.	.	T	-	d	A	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	T	G	.	.	.	C	.	C	.	.	.	.	-	-	-	-	-	-	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	A	A	.	.	.	.	.	.	G	.
J20	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	-	-	-	-	-	-	-	-	.	A	.	C	.	.	.	A	.	T	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	G	.
J16	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	T	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	A	A	A	A	.	.	T	C	.	.	T	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	.	.	.	G	.
TB131	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	A	A	-	-	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	G	.
ES89 484	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	T	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	-	-	-	-	-	-	.	.	.	C	.	.	.	.	.	T	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	d	-	-	-	d—			T	G	.	.	.	.	.	.	.	.	.	G	G
J25	.	.	.	.	.	.	T	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	-	-	-	-	-	-	-	-	-	.	.	.	C	.	.	.	.	.	T	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	G	.
TB27	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	d—		.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	d	-	-	-	-	-	-	-	-	-	A	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	A	.	.	G	.	.	.	G	.
TB19	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	d	-	-	-	-	-	-	-	-	-	.	.	.	G	.	.	.	A	.	T	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	G	.
J30	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	C	.	.	.	.	.	-	-	-	-	-	-	-	.	.	.	C	.	.	.	.	.	.	.	T	.	.	.	.	.	-	-	.	-	-	-	-	-	-	-	d	.	T	.	.	.	A	.	.	.	.	.	A	.	.	.	.	.	.	G	.
J18	.	.	T	.	.	.	T	.	.	-	d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	-	-	-	-	-	-	-	-	-	.	.	.	C	.	C	.	.	.	.	.	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	d—		G	.
D. guanche	.	.	.	.	.	.	A	.	.	-	.	.	.	.	.	T	T	G	C	A	.	A	T	.	.	-	-	-	-	.	.	.	.	.	.	.	.	.	A	.	C	.	.	.	-	-	-	-	-	-	-	-	-	.	A	.	-	-	.	.	.	C	.	.	.	.	.	.	.	.	-	-	-	-	-	-	-	-	-	-	C	.	-	-	-	.	.	.	-	-	.	A	.	.	G	G	.	.	.	G	.

	1859	1862	1866	1868	1874	1881	1887	1888	1889	1901	1909	1914	1915	1916	1917:**d3**	1919	1920:**d8**	1926	1928:**i8**	1936:**i4**	1937	1940	1941:**d3**	1943	1944	1945	1946	1947	1948	1949	1950	1951	1952	1953	1964	1969	1971	1986	2011	2032	2058	2063	2066	2067	2110	2114	2141	2145	2153	2164
TB174	T	G	C	T	G	C	C	C	A	T	G	T	A	-	d—	—	d—	—	T	T	A	T	C	A	-	-	T	A	T	G	T	G	T	G	C	A	T	T	A	A	G	C	A	C	C	G	T	C	G	G
TB398	.	.	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB204	.	T	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB154	.	T	.	.	.	.	.	.	.	.	.	-	-	-	d—	—	d—	—	-	-	-	-	d—	—	-	-	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB21	.	.	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB43	.	.	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	-	.	G	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB2	.	T	.	.	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	A	.	.	d	.	.	.	.	.	.	.	.	.	.	.	.
TB167	.	.	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	.	.	T	.
TB200	.	T	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB366	.	.	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.
TB303a	.	T	.	.	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	.	.	.	.
TB422	.	.	.	.	.	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	.	.	.	.
TB316a	.	.	.	.	.	.	.	.	T	.	.	.	.	C	A	A	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.
TB7	.	T	.	.	.	.	.	.	.	.	.	-	-	-	d—	—	d—	—	-	-	-	-	d—	—	-	-	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB132	.	T	.	.	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	T	G	.	G	.	.	.	T	.	.	.	.	.	.	.	G	.	T	.	.	.	.	.	d	.	.
TB153	.	T	.	.	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	G	.	A	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	C	.	.	T
S89 174	A	T	.	A	.	.	.	.	.	G	.	C	.	C	G	G	A	G	-	-	-	-	.	T	-	-	.	G	.	T	A	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB35	.	T	T	A	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	T	A	.	.	.	T	.	A	.	.	.	.	.	.	.	A	.	.	.	.	.
S89 4	.	T	.	.	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB12	A	T	.	A	.	.	.	.	.	.	T	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	.	.	.	d	.	.	T	.	.	.	.	.	.	.	.	.
J20	A	T	.	A	T	.	.	.	.	.	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	-	-	.	.	.	.	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	.	.	.	T
J16	.	T	.	C	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	.	.	.	d	T	.	.	.	.	.	.	.	.	.	.	.
TB131	.	T	.	A	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
ES89 484	.	T	.	.	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	G	.	A	.	.	.	A	.	.	d	.	.	.	.	.	.	.	.	.	.	.	T
J25	A	T	.	A	A	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.
TB27	A	T	.	A	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	T	A	.	d	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB19	.	T	.	A	.	.	.	.	.	.	.	C	.	C	G	G	A	G	-	-	-	-	.	.	-	-	.	G	.	.	A	.	d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J30	.	T	.	A	.	T	A	A	.	.	.	C	.	C	G	G	A	A	-	-	-	-	.	.	-	-	.	G	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J18	A	T	.	A	.	.	.	.	.	d	.	.	.	-	d—	—	d—	—	.	.	.	.	.	.	T	A	.	G	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
D. guanche	.	-	.	-	-	-	A	.	.	.	-	-	-	-	G	G	A	G	-	-	-	-	.	.	-	-	-	-	G	.	.	.	.	.	T	.	.	.	.	.	-	-	-	.	.	.	.	.	.	.

	2169	2179	2184	2200	2202	2220:**d24**	2223	2228	2246	2251	2265	2280	2298	2300	2306	2313	2315	2324	2333	2336	2345	2350	2351	2359	2360	2363	2366	2380	2382	2385	2405	2407	2415	2426	2442	2444	2454	2489	2490:**d8**	2511	2546	2548	2551	2558	2616	2618	2659	2667	2677	2722	2776
TB174	T	T	C	C	A	G	C	G	C	G	G	G	T	T	T	A	G	T	G	G	A	G	C	C	T	T	G	C	T	T	G	C	T	T	T	C	G	G	C	G	G	T	C	G	T	T	G	T	C	T	T
TB398	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	.	.	.	.
TB204	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	A	.	.	.	T	.	.
TB154	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	A
TB21	.	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB43	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB2	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	.
TB167	A	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	.
TB200	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	A	.	.	.	T	.	.
TB366	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	.
TB303a	.	.	.	.	C	.	.	.	.	.	.	C	C	.	.	.	.	.	.	.	.	A	T	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	A	.	.	.	T	.	.
TB422	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	T	.	.
TB316a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB7	.	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	A	T	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB132	.	.	.	T	C	d—	—	—	.	.	.	C	C	.	.	.	.	.	.	.	.	T	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB153	.	.	.	.	C	.	T	C	.	.	T	C	C	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
S89 174	A	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	.	.	.
TB35	.	d	T	.	C	.	T	.	.	.	.	C	C	.	.	.	C	.	.	.	.	.	T	G	C	.	.	.	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.
S89 4	.	.	.	.	C	.	T	.	.	.	.	C	C	.	.	G	.	.	.	.	.	.	T	.	C	.	.	.	A	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB12	.	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	A	.	.	.	.	C	.	.	.	.	.	.	C	.	.	.	.	.	T	.	.	.	.
J20	.	.	.	.	C	.	T	.	.	A	.	C	C	.	.	.	.	.	A	.	.	.	T	.	C	.	.	.	A	.	.	A	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J16	.	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	T	G	C	.	.	.	A	.	.	.	.	C	.	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB131	A	.	.	.	C	.	T	C	.	.	T	C	C	.	.	.	.	.	A	A	.	A	T	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	d	.	.	.	.	.
ES89 484	.	.	.	.	C	.	T	.	.	A	.	C	C	.	.	.	.	.	A	.	.	.	T	.	C	.	.	.	A	.	.	A	.	C	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.
J25	.	.	.	.	C	.	T	.	A	.	.	C	C	A	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	G	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB27	A	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	C	.	.	.	.	.	.	.	.	G	.
TB19	A	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	d	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J30	A	.	.	.	C	.	T	.	.	.	.	C	C	.	A	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
J18	A	.	.	.	C	.	T	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	T	.	C	.	.	.	A	.	T	.	.	.	.	.	T	A	d	.	.	.	.	.	.	.	.	.	.	.	.
D. guanche	.	.	.	.	C	.	T	.	.	.	-	C	C	.	.	.	.	.	.	.	.	C	G	.	.	-	-	-	-	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.

	2851	2855	2865	2869	2881	2883	2897	2899	2905	2915	2919	2920	2940:i2	2962:d2	2967	3018	3049	3082	3114	3130	3131	3133
TB174	A	G	C	C	A	T	C	C	T	A	A	C	-	G	G	A	C	C	T	d	C	C
TB398	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	d	.	.
TB204	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	T	.	.	d	.	.
TB154	.	.	.	.	.	.	.	T	.	.	.	.	-	.	.	.	.	.	.	d	.	.
TB21	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	d	d	.
TB43	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	T	.	d	.	.
TB2	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	T	.	.	T	.	.
TB167	.	.	.	.	.	C	.	.	.	.	.	.	-	d	.	.	.	.	.	d	.	.
TB200	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	d	.	.
TB366	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	T	.	.	d	.	.
TB303a	.	.	.	A	.	.	.	.	.	.	T	.	-	.	.	.	.	.	.	T	.	.
TB422	C	-	.	.	.	.	.	.	.	G	.	.	-	.	.	.	.	.	.	T	.	.
TB316a	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	T	.	.	d	.	.
TB7	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	d	.	.
TB132	C	-	.	.	.	.	.	.	.	G	.	.	-	d	.	.	.	.	.	T	.	.
TB153	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	T	.	.
S89 174	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	T	.	.
TB35	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	T	.	T
S89 4	C	-	.	.	.	.	.	.	.	G	.	.	-	d	.	.	.	.	.	T	.	.
TB12	.	.	.	.	.	.	.	.	C	.	.	.	-	.	T	G	.	.	.	T	.	.
J20	.	.	.	.	.	.	.	.	C	.	.	.	G	d	.	.	.	.	.	T	.	.
J16	C	-	T	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	T	.	.
TB131	.	.	.	.	T	.	.	.	.	.	.	A	-	.	.	.	.	.	.	T	.	.
ES89 484	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	G	T	.	.
J25	.	.	.	.	T	.	.	.	.	.	.	A	-	.	.	.	.	.	.	T	.	.
TB27	.	.	T	.	.	.	.	d	.	.	.	.	-	d	.	.	.	.	.	T	.	.
TB19	.	.	.	.	.	.	T	.	C	.	.	.	-	.	.	.	.	.	.	T	.	.
J30	.	.	.	.	.	.	.	.	.	.	.	.	-	d	.	.	.	.	.	T	.	.
J18	.	.	.	.	T	.	.	.	.	.	.	A	-	.	.	.	.	.	.	T	.	.
D. guanche	C	-	T	.	.	.	.	.	G	.	.	.	-	.	.	.	.	.	.	T	.	.

	3136:d2	3137	3146	3157	3160	3169	3187	3189	3203:d6	3204	3221	3227:d6	3233	3234	3245	3249	3252	3299	3365	3395	3400	3411	3416:d4
TB174	C	d	T	T	T	A	T	A	d—	—	A	d	T	G	G	A	G	A	G	C	C	G	A
TB398	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	.	.	T	.	.
TB204	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	.	.
TB154	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	.	.
TB21	.	C	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	.	.
TB43	.	d	.	.	.	.	.	G	C	T	.	d	.	.	.	.	.	.	C	.	T	.	.
TB2	.	C	.	C	C	.	.	G	C	T	.	d	.	.	.	.	.	.	.	.	T	.	.
TB167	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	.	.
TB200	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	.	.
TB366	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	.	.
TB303a	d—	—	.	C	C	.	.	.	C	T	.	d	.	.	.	.	.	.	C	.	T	.	.
TB422	d—	—	.	C	C	.	.	.	C	T	.	d	.	.	.	.	.	.	C	.	T	.	.
TB316a	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	.	.
TB7	.	d	.	.	.	.	.	.	d—	—	.	d	.	A	.	.	.	.	C	.	T	T	.
TB132	.	C	.	C	C	.	.	.	C	T	.	d	.	.	.	.	.	.	C	.	T	.	.
TB153	d—	—	.	.	C	T	.	.	C	T	.	d	.	A	.	.	.	.	C	.	T	.	d—
S89 174	.	C	.	C	C	T	.	.	C	T	T	A	.	A	.	G	A	.	C	.	T	.	.
TB35	.	d	.	.	C	T	C	.	C	T	.	d	.	.	.	.	.	.	C	.	T	.	.
S89 4	.	C	.	C	C	.	.	.	C	T	.	A	.	A	.	G	A	.	C	.	T	.	.
TB12	d—	—	.	.	C	T	.	.	C	T	.	d	.	.	.	.	.	.	.	.	T	.	.
J20	d—	—	.	C	C	.	.	.	C	T	.	d	.	.	A	.	.	.	C	.	T	.	.
J16	d—	—	.	C	C	.	.	.	C	T	.	d	.	.	.	.	.	.	C	.	T	.	.
TB131	d—	—	A	.	C	T	.	.	C	C	.	A	.	.	.	.	.	.	C	.	T	.	.
ES89 484	d—	—	.	.	C	T	.	.	C	C	.	A	.	.	.	.	.	.	C	.	T	.	.
J25	d—	—	.	.	C	T	.	.	C	T	.	d	A	.	.	.	.	.	C	.	T	.	.
TB27	d—	—	.	.	C	T	.	.	C	C	.	A	.	.	.	.	.	.	C	A	T	.	.
TB19	.	C	.	C	C	T	.	.	C	T	.	d	.	A	.	.	.	.	.	.	T	.	.
J30	d—	—	.	.	C	T	.	.	C	T	.	d	.	.	.	.	.	C	.	.	T	.	.
J18	d—	—	.	.	C	T	.	.	C	T	.	d	.	.	.	.	.	.	.	.	T	.	.
D. guanche	.	C	.	C	C	.	.	.	C	T	.	A	.	A	.	.	A	.	C	.	T	.	.

	3420:d2	3443	3462	3473	3479:d2	3480	3483	3484	3486	3492	3493	3507:d6	3508	3509	3513	3514	3515:d2	3517:d4	3521:i2	3523
TB174	d	G	T	T	d—	—	T	G	T	T	C	C	T	C	d—	—	—	—	-	-
TB398	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
TB204	d	.	.	d	G	d	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB154	d	.	.	d	G	d	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB21	d	.	.	d	G	d	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB43	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	C	T	d—	—	-	-
TB2	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
TB167	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	C				-	-
TB200	d	.	.	d	d—	—	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB366	d	.	.	d	d—	—	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB303a	d	.	.	d	G	d	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB422	d	.	C	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
TB316a	d	.	.	d	G	d	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB7	d	.	.	d	G	G	.	.	.	.	.	d—	—	—	—	—	—	—	-	-
TB132	d	.	C	.	d—	—	.	.	.	.	.	A	.	.	C	T	C	d	-	-
TB153	—	.	.	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
S89 174	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	T	T	C	C	C	-
TB35	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
S89 4	G	.	.	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
TB12	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	C	T	C	C	-	-
J20	d	.	.	.	d—	—	.	.	.	A	T	.	C	A	C	T	C	C	C	C
J16	d	.	C	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
TB131	d	.	C	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
ES89 484	d	.	C	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
J25	d	.	.	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
TB27	d	.	C	.	d—	—	.	.	.	.	.	.	.	.	C	T	d—	—	-	-
TB19	d	.	.	.	T	G	-	-	.	.	.	.	.	.	d—	—	—	—	-	-
J30	d	.	C	.	d—	—	.	.	.	.	.	.	.	.	d—	—	—	—	-	-
J18	d	T	.	.	T	G	.	-	d	.	.	.	.	.	d—	—	—	—	-	-
D. guanche	G	.	.	.	T	G	-	-	.	.	.	.	.	.	C	T	C	C	-	-

	3539	3540	3571	3610	3629	3647	3666	3693	3739	3743	3751	3756	3769	3770	3771	3772
TB174	d	G	G	G	G	T	T	C	C	T	G	G	C	C	G	-
TB398	d	.	.	.	.	.	.	T	.	A	.	.	A	T	C	-
TB204	d	.	.	A	.	.	.	.	.	.	.	.	A	T	C	-
TB154	d	.	.	.	.	.	.	.	.	.	.	.	.	T	.	-
TB21	d	.	.	.	.	.	G	.	.	.	.	.	A	T	C	-
TB43	d	.	.	.	.	.	.	.	.	.	.	.	A	T	C	-
TB2	d	.	.	.	.	.	.	.	.	.	.	.	.	T	.	-
TB167	G	.	.	.	.	.	.	.	.	.	.	.	.	T	.	-
TB200	d	.	.	.	.	.	.	.	.	.	.	.	.	T	.	-
TB366	d	.	.	.	.	.	.	.	.	.	.	.	.	T	.	-
TB303a	d	.	.	.	.	.	.	.	.	.	.	.	A	T	C	-
TB422	d	.	.	.	C	.	.	.	.	.	.	.	.	T	.	-
TB316a	d	.	.	C	.	.	.	.	.	.	.	.	.	T	.	-
TB7	d	.	.	.	.	.	.	.	.	.	.	.	A	T	C	-
TB132	d	.	.	.	.	.	.	.	.	.	.	.	.	T	C	-
TB153	d	.	.	.	.	.	.	.	.	.	.	.	.	T	C	-
S89 174	d	.	.	.	.	.	.	.	T	.	.	.	.	T	.	-
TB35	d	.	.	.	.	.	.	.	.	.	.	.	A	T	C	-
S89 4	d	.	.	.	.	C	.	.	.	.	.	.	.	T	.	-
TB12	d	.	.	.	.	.	.	.	.	.	.	.	.	T	d	-
J20	d—	—	.	.	.	.	.	.	.	.	.	.	.	T	.	G
J16	d	.	.	.	.	.	.	.	.	.	.	.	.	T	C	-
TB131	d	.	.	.	.	.	.	.	.	.	.	.	.	T	C	-
ES89 484	d	.	.	.	.	.	.	.	.	.	T	A	.	T	.	G
J25	d	.	.	.	.	.	.	.	.	.	.	.	.	T	C	-
TB27	d	.	.	.	.	.	.	.	.	.	T	A	.	T	.	G
TB19	d	.	C	.	.	.	.	.	.	.	.	.	.	T	C	-
J30	d	.	.	.	.	.	.	.	.	.	.	.	.	T	C	-
J18	d	.	.	.	.	.	.	.	.	.	.	.	.	T	C	-
D. guanche	C	.	-	.	.	.	.	.	A	.	T	.	.	G	C	-

	3774	3775	3782	3792	3798	3799	3803	3819	3848	3865	3879	3887	3892	3895	3896	3900	3901	3905	3907	3908
TB174	G	C	T	G	G	C	G	C	C	T	A	T	G	T	T	C	T	C	T	T
TB398	T	.	G	T	.	.	.	T	.	.	.	.	.	.	G	G	C	.	.	.
TB204	T	.	G	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.
TB154	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB21	T	.	G	T	.	.	.	T	.	A	.	.	.	.	.	.	.	.	.	.
TB43	T	.	G	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.
TB2	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.
TB167	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB200	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	G	C	.	.	.
TB366	.	.	.	.	.	.	.	T	.	.	.	.	.	.	G	G	C	.	.	.
TB303a	T	.	G	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	C
TB422	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB316a	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB7	T	.	G	T	.	.	.	T	.	.	.	.	.	.	G	G	C	.	.	.
TB132	.	.	G	T	.	.	.	T	.	.	.	.	.	.	.	.	.	.	C	.
TB153	.	.	G	.	.	.	.	T	.	.	G	.	.	.	.	.	.	.	.	.
S89 174	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	A	.	.	.
TB35	T	.	G	T	.	.	.	T	A	.	.	.	.	.	.	.	.	.	.	.
S89 4	.	.	.	C	.	.	.	T	.	.	.	C	T	.	.	G	C	.	.	.
TB12	T	.	.	.	.	.	.	T	.	.	.	.	.	.	G	G	C	.	.	.
J20	T	.	.	.	A	T	A	T	.	.	.	.	.	.	.	.	.	.	.	.
J16	.	.	G	C	.	.	.	T	.	.	.	C	T	.	.	G	C	.	.	.
TB131	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	G	C	.	.	.
ES89 484	T	.	.	.	A	T	A	T	.	.	.	.	.	.	G	G	C	.	.	.
J25	.	.	G	.	.	.	.	T	.	.	.	.	.	.	G	G	C	.	.	.
TB27	T	.	.	T	.	.	.	T	.	.	.	.	.	G	G	G	C	.	.	.
TB19	.	.	G	T	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.
J30	.	A	G	T	.	.	.	T	.	.	.	.	.	.	.	G	A	.	.	.
J18	.	.	G	T	.	.	.	T	.	.	.	.	.	.	.	.	.	T	.	.
D. guanche	T	.	.	-	A	T	A	T	.	.	.	.	.	.	.	.	.	.	.	.

OS-F

E_1 | I_1 | E_2 | I_2 | E_3 | I_3

	3910	3912	3913	3915	3916	3921	3923	3947	3953	3962	3963	3968	3969	3978:**d12**	3989	3990	4023	4029	4084	4088	4107	4123	4130:**d3**	4143	4184	4193	4224	4236	4245	4246	4247:**d2**	4251	4253	4274:**d5**	4279	4280:**d6**	4328	4346:**d2**	4361	4396	4404	4418	4420	4421	4422:**d5**	4430:**d3**	4433:**d6**	4439	4452	4462	4489	4500:**i4**	4504	4511	4524:**i9**	4539	4544	4551:**d10**	4576	4610	4626	4658:**i6**	4717	4726:**d3**	4750	4759	4816	4864	4865	4871:**d2**	4873	4874:**d3**	4877:**d3**	4879	4880:**d2**	4882:**i9**	4945	5040	5046	5105	5107	5209	5210	5213	5219	5222	5231	5238	5288	5327	5340	5341	5352	5358	5364	5369	5389	5421	5443	5464	5467
TB174	-	G	C	T	C	A	G	T	A	G	A	A	T	T	T	C	G	G	G	A	d	C	T	G	A	T	T	A	A	A	C	A	A	G	A	d	C	d	T	C	C	A	C	C	T	d	—	C	C	T	C	C	C	C	-	T	C	C	T	G	C	-	G	C	C	G	G	G	C	A	G	C	G	C	G	C	A	T	T	G	A	C	C	T	A	C	G	T	C	C	T	G	G	T	A	G	T	C	A	A	C
TB398	-	.	.	.	.	.	.	G	.	T	.	.	.	d	—	.	.	.	.	.	d	.	.	.	.	.	.	.	.	-	.	.	.	.	.	d	T	d	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB204	-	T	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	.	.	d	.	d	.	.	T	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	T	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB154	-	T	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	A	.	.	d	.	.	.	.	.	.	.	.	-	.	.	.	.	.	d	T	d	C	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	C	.	.	.	.	C	.	-	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB21	-	T	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	.	.	.	-	.	.	.	.	.	d	.	d	.	.	T	.	.	.	.	d	—	.	.	.	.	-	d	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB43	-	T	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	-	.	.	.	.	.	d	.	d	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB2	T	T	T	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	.	.	d	.	d	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB167	-	.	.	.	.	.	T	G	.	T	.	.	.	.	.	.	.	.	C	G	d	G	.	.	.	.	.	.	.	-	.	.	.	.	.	d	.	d	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB200	-	.	.	.	.	G	.	G	.	T	.	.	.	.	.	.	.	.	.	.	d	G	.	T	.	.	G	.	.	-	.	.	.	.	.	d	.	d	.	.	T	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	C	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB366	-	.	.	.	.	G	.	G	.	T	.	.	.	.	.	.	.	.	.	.	d	.	.	.	.	.	G	.	.	-	.	.	.	.	.	d	.	d	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	C	.	.	.	.	A	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB303a	-	T	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	-	.	.	.	.	.	d	.	d	.	.	T	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	A	A	.	.	.	.	.	.	.	.	.	C	.	.	C	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB422	-	T	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	.	.	.	.	.	.	.	.	-	.	.	.	.	.	d	T	d	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	T	.	.	.	.	.
TB316a	-	T	.	.	.	.	.	G	C	T	G	.	G	.	.	.	T	.	.	.	d	.	.	.	.	.	.	.	.	-	.	.	.	.	.	d	T	C	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB7	-	.	.	.	.	.	.	.	.	T	.	.	.	.	.	.	.	.	.	.	d	.	.	.	.	.	G	.	.	-	.	.	.	.	.	d	.	d	.	.	.	.	T	T	.	d	C	.	.	.	A	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.
TB132	-	.	.	.	.	.	.	G	C	T	.	.	G	.	C	G	.	.	.	.	d	G	.	.	.	G	G	.	.	-	.	.	.	.	.	d	.	d	.	.	T	.	.	.	.	d	—	.	.	.	.	.	A	.	-	.	.	.	.	.	.	-	.	d	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB153	-	.	.	G	T	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	.	.	d	.	d	.	.	T	.	.	.	.	d	C	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	C	.	.	.	.	.	.	.	.	.	.	.	C	C	.	.	.	.	.	.	.	.	.	.	A	G	A	T	.	C	.	.	.	.	T	.
S89 174	-	.	.	.	.	.	.	G	.	T	.	.	G	.	C	G	.	.	.	.	d	G	.	.	d	.	G	.	.	-	.	.	.	d	-	G	.	d	.	.	T	.	.	.	.	d	—	.	G	.	.	.	.	.	-	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	C	G	G	.	.	.	A	G	A	T	.	C	.	.	.	.	.	A
TB35	-	.	.	.	.	.	.	G	.	T	.	.	G	.	C	G	.	.	.	.	d	G	.	.	.	.	G	T	d	-	.	.	.	.	.	d	.	d	.	.	.	.	.	.	.	d	C	.	.	.	.	.	.	.	-	.	.	.	.	.	T	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	C	.	.	G	.	.	.	.	C	.	.	.	.	.	.	C	.	.	.	.	.	.
S89 4	-	.	.	.	.	.	.	G	.	T	.	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	.	.	d	.	d	.	.	.	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	.	.	d	—	—	—	—	—	-	.	C	.	.	.	.	.	C	G	G	.	.	.	A	G	A	T	.	C	.	.	.	.	.	.
TB12	-	.	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	d	-	G	.	d	.	.	T	.	.	.	.	d	C	.	.	C	.	.	.	.	-	.	T	.	C	.	.	-	.	.	T	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	A	.	A	.	.	C	.	.	.	.	.	.
J20	-	.	.	.	.	.	.	G	.	T	.	.	G	.	C	G	.	.	.	.	C	G	.	.	.	.	G	.	.	-	.	.	.	d	-	G	.	d	.	.	T	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	C	.	.	.	.	.	G	C	.	.	.	.	.	.
J16	-	.	.	.	.	.	.	G	.	T	.	.	.	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	d	-	G	.	d	.	.	T	.	.	.	.	d	C	.	.	.	.	.	.	.	-	.	.	d	.	.	.	-	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	G	G	.	.	.	.	.	A	.	.	C	.	C	T	.	.	T
TB131	-	.	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	d	-	G	.	d	.	T	.	G	.	.	.	T	C	.	.	.	.	.	A	.	-	.	.	.	.	.	.	-	.	.	.	.	.	.	.	G	C	d	.	G	A	G	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
ES89 484	-	.	.	.	.	.	.	G	C	T	.	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	.	.	.	.	d	-	G	.	d	.	.	T	.	.	.	.	d	C	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	C	G	G	.	.	T	A	.	.	.	G	C	.	.	.	T	T	.
J25	-	.	.	.	.	.	.	G	C	T	.	G	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	d	-	G	.	d	.	.	T	.	.	.	.	d	—	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	.	.	d	—	—	—	—	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	A	.	.	.	.	C	.	.	.	.	.	.
TB27	-	.	.	.	.	.	.	G	.	T	.	.	G	.	.	.	.	.	.	.	d	G	.	.	d	.	G	T	d	-	.	.	.	d	-	G	.	d	.	.	.	.	.	.	.	d	C	.	.	.	.	.	.	.	-	.	.	.	.	.	.	-	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	.
TB19	-	.	.	.	.	.	.	G	C	T	.	.	G	.	C	G	.	.	.	.	d	G	.	.	.	.	G	.	.	-	.	.	.	d	-	G	.	d	.	.	T	.	.	.	.	d	C	.	G	.	.	.	.	T	-	.	T	.	.	.	.	-	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	.	G	.	.	.	.	.	.	.	.	.	.	.	C	.	.	.	.	.	T
J30	-	.	.	.	.	.	.	G	C	T	G	.	G	.	.	.	.	.	.	.	d	G	.	.	.	.	G	.	.	-	d	T	T	d	-	G	.	d	.	.	T	.	.	.	.	d	C	.	.	C	.	.	.	.	-	C	.	.	.	.	.	-	.	.	.	.	.	.	.	G	C	d	.	G	A	G	.	.	.	.	.	.	.	.	.	.	.	.	.	A	.	.	.	.	C	.	.	.	.	.	.
J18	-	.	.	.	.	.	.	G	.	T	.	.	G	.	.	.	.	.	.	.	C	G	d	.	.	.	G	.	.	-	.	.	.	d	-	G	.	d	.	.	T	.	.	.	d	—	—	.	.	.	.	.	A	.	-	.	.	.	.	.	.	-	.	.	.	.	.	.	.	d	—	—	—	—	—	-	.	.	.	.	.	.	G	.	.	.	.	C	.	A	G	A	T	.	C	.	.	.	.	.	.
D. guanche	-	.	.	.	.	.	.	G	.	T	T	.	G	C	C	-	.	.	.	.	T	G	.	.	.	.	G	-	.	-	.	.	T	.	-	G	.	C	.	.	T	.	.	.	.	A	C	.	.	.	.	-	.	T	-	.	T	.	.	-	-	-	-	.	A	.	.	.	.	G	.	.	.	.	.	-	T	.	.	.	.	.	.	.	.	.	.	.	.	.	C	.	T	.	C	.	.	.	.	.	T

	5472	5474	5475	5491	5496	5498	5501	5531	5532	5578	5591	5612	5634:**i2**	5636	5646:**i2**	5655	5675	5676	5678	5679:**i4**	5683:**i22**	5697
TB174	A	A	A	C	T	C	A	T	G	A	C	C	-	A	T	A	A	C	G	A	G	A
TB398	.	.	.	.	.	.	.	A	T	.	.	.	-	.	-	.	.	.	-	-	-	-
TB204	.	.	.	.	.	.	.	A	T	.	.	.	-	.	-	.	.	.	.	.	.	.
TB154	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	.	.	.	.	.	.
TB21	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	.	.	.	.	.
TB43	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	.	G	.	.	.	.
TB2	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	T	.	.	.	.	.	T
TB167	.	.	.	.	.	.	.	A	T	.	.	.	C	.	.	.	.	.	.	.	.	.
TB200	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	G	.	.	.	.
TB366	.	.	.	.	.	.	.	A	T	.	.	.	-	.	-	.	.	.	-	-	-	-
TB303a	.	.	.	.	.	.	.	.	.	.	.	.	-	.	.	.	.	G	.	.	.	.
TB422	.	.	.	.	.	.	.	A	T	.	.	.	-	.	-	.	.	.	.	.	.	.
TB316a	.	.	.	.	.	.	.	A	T	.	.	.	-	.	-	.	.	.	.	.	.	.
TB7	.	.	.	.	.	.	.	A	T	.	.	.	-	.	-	.	.	.	.	.	.	.
TB132	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	.	.	.	.	.	.
TB153	.	.	.	.	.	.	.	A	T	G	.	.	-	.	.	.	G	G	.	.	.	.
S89 174	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	.	.	.	.	.	.
TB35	.	.	.	.	.	.	.	.	T	.	.	.	-	.	.	.	.	.	.	.	.	.
S89 4	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	.	.	.	.	.	.
TB12	.	.	.	A	G	.	.	.	T	.	.	.	-	.	.	.	.	.	.	.	.	.
J20	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	G	G	.	.	.	.
J16	.	G	T	.	.	.	.	A	T	.	.	A	-	.	.	.	.	.	.	.	.	.
TB131	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	.	.	.	.	.	.
ES89 484	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	G	G	.	.	.	.
J25	G	.	.	.	.	.	.	A	T	.	.	.	-	.	-	.	.	G	.	.	.	.
TB27	.	.	.	.	.	.	.	A	T	.	.	.	-	.	.	.	.	G	.	-	.	.
TB19	.	G	T	.	.	.	.	A	T	.	T	.	-	.	.	.	.	.	.	.	.	.
J30	.	.	.	.	.	.	.	.	T	.	.	.	-	.	-	.	G	G	.	.	.	.
J18	.	.	.	.	.	T	G	A	T	.	.	.	-	.	.	.	G	G	.	.	.	.
D. guanche	.	.	.	.	C	.	.	.	T	-	-	-	-	-	-	-	-	-	-	-	-	-

Capítulo 3.4. Molecular evolution of the *OS-E* and *OS-F* genes in Drosophila: evolutionary inferences from the sequence variation and protein structure.

3.4.1 Resumen

En este capitulo se analiza la evolución molecular de las secuencias codificadoras de los genes *OS-E* y *OS-F* en 14 especies del subgénero Sophophora de Drosophila, con objeto de determinar el impacto de la selección natural (positiva o negativa) en la evolución de estos genes duplicados. El trabajo se ha realizado combinando análisis por máxima verosimilitud de la variabilidad sinónima y no sinónima (mediante el uso de un modelo de evolutivo de codones), con la información de la putativa estructura tridimensional de la proteína (determinada en otras OBPs). A partir del número de substituciones nucleotídicas entre especies se ha estimado que la duplicación que originó los dos genes se originó hace unos 43-94 myr. Los resultados indican que, a pesar de que ambos genes presentan una gran y similar constricción funcional (ω ~0.06), el gen *OS-E* presenta una tasa evolutiva global significativamente más alta. Los resultados indican que la selección natural positiva pudo actuar sobre posiciones concretas de la zona codificadora de estos dos genes, y pudo tener un papel importante en la la preservación de los dos genes duplicados en el genoma.

Artículo IV

Molecular evolution of the *OS-E* and *OS-F* genes in Drosophila: evolutionary inferences from the sequence variation and protein structure

Alejandro Sánchez-Gracia and Julio Rozas

2005

(En preparación)

Departament de Genètica, Facultat de Biologia, Universitat de Barcelona,
Barcelona, Spain

INTRODUCTION

The detection of adaptive changes at the molecular level (i.e. molecular adaptation) is usually accomplished through the analysis of the number of nonsynonymous substitutions per nonsynonymous site (K_A) and the number of synonymous substitutions per synonymous site (K_S). Indeed, under strict neutrality, synonymous and nonsynonymous mutations will fix at identical rates, and their ratio ($\omega = K_A/K_S$) is expected to be 1. Purified selection against deleterious nonsynonymous mutations will cause the fixation of synonymous mutations at a faster rate (assuming that synonymous mutations are strictly neutral, but see AKASHI 1995) than nonsynonymous mutations, and therefore ω < 1. Advantageous nonsynonymous mutations, on the contrary, could be fixed in the population at a faster rate than synonymous mutations, and thus ω might be greater than 1.

There are a number of convincing examples where the action of positive natural selection has been detected using this approach (e.g. YANG and BIELAWSKI 2000; LIBERLES *et al*. 2001; LIBERLES and WAYNE 2002). In addition, recent experimental studies have confirmed the functional significance of some predicted positive selected changes (IVARRSON *et al*. 2003; SAWYER *et al*. 2005; BISHOP 2005). A critical point of the method lies in the estimation of the number of nonsynonymous and synonymous substitutions per site. There are three major groups of methods for detecting positive selection that use phylogenetic information on a site-by-site basis (i.e. allowing variation in the selective constraints across sites): i) counting methods, which involves the reconstruction of the ancestral sequences using parsimony (SUZUKI and GOJOBORI 1999), or by a likelihood-based approach (NIELSEN 2002; NIELSEN and HUELSENBECK 2002; SUZUKI 2004; KOSAKOVSKY POND and FROST 2005a), ii) random-effects models (NIELSEN and YANG 1998; YANG *et al.* 2000; HUELSENBECK and DYER 2004; KOSAKOVSKY POND and FROST 2005a), where the ω ratios are estimated assuming a given distribution, and ii) fixed-effects models (SUZUKI 2004; MASSINGHAM and GOLDMAN 2005; KOSAKOVSKY POND and FROST 2005a), where the nonsynonymous to synonymous substitutions ratio is directly estimated at each codon site.

Although random-effects models are powerful in detecting positive selection (ANISIMOVA *et al.* 2001, 2002, 2003; WONG *et al*. 2004), frequently (under certain conditions) generate false positive results (SUZUKI and NEI 2001, 2002, 2004; ZHANG 2004). Counting and fixed-effects models are, on the contrary, highly conservative (WONG *et al.* 2004; KOSAKOVSKY POND and FROST 2005a). Recently, YANG *et al*. (2005) developed a Bayes empirical Bayes (BEB) approach that takes into account sampling errors in the estimation of maximum likelihood parameters. These authors argued that the BEB approach would eliminate most false positive results incorporated in the random-effects models. Moreover, while random and fixed-effects models (NIELSEN and YANG 1998; YANG *et al.* 2000; SUZUKI 2004; MASSINGHAM and GOLDMAN 2005) assume a single (fixed) synonymous substitution rate for all sites, counting methods implicitly allow for variation across sites. Models incorporating a variable synonymous substitution rate across sites could likely provide more accurate estimates of the selective constraints when the synonymous substitution rate differs across sites (KOSAKOVSKY POND and FROST 2005a; see also YANG and SWANSON 2002).

We applied random and fixed-effect models on nucleotide variation at two genes belonging to the olfactory system. This system, allows individuals detecting enormously diverse information from the external environment, and in most animals the odor detection is a fundamental feature for their survival and reproduction. Therefore, olfactory-involved genes are good candidates to be evolved by positive natural selection. In fact, there are compiling evidences where positive selection has driven the evolution of these genes, both in insects and in vertebrates (e.g. KRIEGER and ROSS 2002; WILLETT 2000; EMES *et al.* 2004; GILAD *et al*. 2003; NGAI *et al*. 1993; CLARK *et al*. 2003; WATTS *et al*. 2004; see also GIMELBRANT *et al*. 2004; TAKAHASHI and TAKANO-SHIMIZU 2005).

In insects, the primary step in the olfactory perception is accomplished by the Odorant Binding Proteins (OBPs), which bind and solubilize odorants, including pheromones (HECKMAT-SCAFE *et al*. 2002). The insect OBPs are small hydrophilic proteins which bind odorant molecules at the pores of the olfactory sensilla, transporting them through the aqueous lymph, and delivering near the olfactory receptors (VOGT 2005). In spite the similar function of insect and vertebrate OBPs, they are evolutionarily unrelated (TEGONI *et al*. 2000). In Drosophila, two of these OBPs genes, *OS-E* and *OS-F*, arisen by an old gene duplication event (HECKMAT-SCAFE *et al*. 2000) and co-express in the same specific subset of olfactory sensilla (sensilla tricoidea), in the antennal segment 3 (MCKENNA *et al*. 1994). The *OS-E* and *OS-F* genes are located in tandem

(~1kb apart), with a similar gene structure and a high sequence similarity (70% of amino acid identity). Interestingly, the analysis of DNA polymorphism and divergence in some Drosophila species has showed that these genes evolved in a non-neutral fashion (Sánchez-Gracia *et al*. 2003, see chapters 3.2 and 3.3).

Here, we analyzed nucleotide variation at the *OS-E* and *OS-F* genes in a number of Drosophila species to understand the evolutionary forces underlying their molecular evolution. In particular, we surveyed nucleotide variation at the coding region of these genes in fourteen Drosophila species of the Sophophora subgenus. For the analysis, we applied random and fixed-effects models, integrating information of DNA polymorphism, amino acid and nucleotide-based divergence, and 3D structure data to determine the selective constraint levels, and the putative role of positive selection in the evolutionary history of these genes, and especially in the maintenance of these duplicated gene copies.

MATERIAL AND METHODS

Fly samples: We used highly inbreed lines (10 generations of sib mating) of *D. teissieri*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, *D. bifasciata*, *D. persimilis* and *D. miranda* (species kindly provided by F. LEMEUNIER, G. PERIQUET and R. C. LEWONTIN). In this study, we also include the DNA sequence of the *OS* region from *D. melanogaster*, *D. simulans*, *D. mauritiana* and *D. erecta* (SÁNCHEZ-GRACIA *et al*. 2003), and that of *D. subobscura, D. guanche* and *D. madeirensis* (chapter 3.3).

DNA extraction and sequencing: Total genomic DNA was extracted from live flies by using a modification of protocol 48 in ASHBURNER (1989). DNA fragments, including the complete coding region of the *OS-E* and *OS-F* genes, were amplified by using the Polymerase Chain Reaction (PCR, SAIKI *et al*. 1988). In addition to the primers previously used for the amplification of the *OS*-region in *D. melanogaster*, *D. simulans*, *D. mauritiana* and *D. erecta* (SÁNCHEZ-GRACIA *et al*. 2003; chapter 3.2), we also employed additional oligonucleotides for the amplification of the new species. Some of these primers were designed using information of conserved genomic regions between *D. pseudoobscura* and *D. melanogaster* (Berkley Drosophila Genome Project, Release 4; RICHARDS *et al*. 2005). Although the length of the amplified genomic regions varied among species, they always included the coding region of the two genes. PCR products were purified using the Qiaquick purification kit (QIAGEN, Chatsworth, CA), and cycle sequenced using primers separated at intervals of ~400 nucleotides. Occasionally, a genome walking strategy was also needed to complete the DNA sequence. Sequenced fragments were separated on the ABI 377 and 3700 sequencers. For all species, the DNA sequence on both stains was determined.

Data analysis: We used SeqMan version 5.53 (DNASTAR, Inc.) for assembling the DNA sequence of each line. The sequences of the 14 species were multiple aligned by using the Clustal W program (THOMPSON *et al*. 1994), and edited with MacClade version 3.05 program, (MADDISON and MADDISON 1992). For the analyses we used three different data sets: a multiple alignment for each individual gene (named as Al_OSE and Al_OSF for *OS-E* and *OS-F* genes, respectively), and another (26 sequences; named as Al_OS) including information of the mature protein, since the alignment of the signal peptide fragment between the gene copies was unreliable.

We calculated nucleotide variation estimates by using DnaSP version 4.10 (ROZAS *et al*. 2003), and MEGA version 3 (KUMAR *et al.* 2004) programs. The MEGA program was also used to estimate the phylogenetic relationships. Phylogenetic trees were constructed applying the neighbor-joining algorithm (SAITOU and NEI 1987), and their reliability was determined by the bootstrap method (based on 1000 replicates). We applied the MK test (MCDONALD and KREITMAN 1991) to assess for putative departures of the correlation between synonymous and nonsynonymous variation within and between species, expected under the neutral model.

For the maximum likelihood analysis, we applied the random-effects models from the *codeml* program, implemented in the PAML package (version 3.14, YANG 1997), to obtain the ML estimates under different codon substitution models:

1. *Branch models*: M0 (one ratio), FR (free-ratio) and branch-specific models (YANG 1998; YANG and NIELSEN 1998).
2. *Site models*: M1 (nearly neutral), M2 (positive selection), M3 (discrete), M7 (beta), M8 (beta&ω) (NIELSEN and YANG 1998; YANG *et al*. 2000).
3. *Branch-site models*: Models A and B (YANG and NIELSEN 2002).

We first estimated the branch lengths under model M0 since it allows obtaining good initial values to be used in further complex codon model analyses. To prevent incorrect parameter estimates caused by local optima, the *codeml* program was run, for the same model, multiple times specifying different initial values; these suboptimal estimates might cause low accuracy in Bayesian identification of positively selected codon sites. The likelihood ratio test (LRT) was used to compare the fit to the data of two nested models, assuming that twice the log likelihood ($2\Delta\ell$) difference between the two models follows a χ^2 distribution, with a number of degrees of freedom equal to the difference in the number of free parameters (WHELAN and GOLDMAN 1999). *P*-values were obtained using the *chi2* program, also included in the PAML package.

We also applied the fixed-effects likelihood model (FEL) described in KOSAKOVSKY POND and FROST (2005a) and included in the Datamonkey web page (http://www.datamonkey.org, KOSAKOVSKY POND and FROST 2005b). This model allows estimating synonymous and nonsynonymous substitution rates at individual codons but allowing variable rates for both types of substitutions. The statistical significance of the analysis was also assessed by the LRT.

To determine if the two paralogous genes evolve at different rates, we conducted the TAJIMA'S one-degree-of-freedom (1D) relative rate test (RRT; TAJIMA 1993). For the analysis we used all nucleotide substitutions present in the common sequences of the AI_OSE and AI_OSF data sets (see below).

The secondary structure of the *OS* proteins was predicted using the PHD program (ROST 1996, 2001). We also determined the putative 3D structure of these proteins by using the Predicting Protein 3D server (http://dove.embl-heidelberg.de/3D, HUYNEN *et al*. 1998). Such analysis was conducted using the amino acid sequence of the *OS-E* of *D. melanogaster* to search in the PDB (Protein Data Bank, BERMAN *et al*. 2000) for proteins with high amino acid sequence similarity and resolved 3D structure. The 3D structure of the sequence with the highest blast score was retrieved. The Cn3D version 4.1 program (available at the NCBI web site, http://www.ncbi.nlm.nih.gov/) was used to visualize the sequence alignment, the 3D structure, and the location of the relevant amino acid replacements identified in the codon evolution analysis.

Computer resources: Nearly all computations were performed in a 1GHz PowerPC G4 processor. Models FR, M7 and M8, which have high computation requirements, were computed using CESCA facilities (Centre de Supercomputació de Catalunya).

RESULTS

Amino acid and nucleotide divergence: Figure 1 shows the multiple alignment of the OS-E and OS-F mature proteins (conceptual translation). The six conserved cysteines, typical of the OBP protein family, are present in all proteins except in the OS-E of *D. erecta* (see SÁNCHEZ-GRACIA *et al.* 2003). We found 99 amino acid replacements, although none of them produced significant variations in the predicted secondary structure of the protein (results not shown). Estimates of the overall amino acid sequence divergence (using the JTT Matrix) were 0.204 (0.319 between OS-E and OS-F groups, and 0.090 and 0.069 within OS-E and OS-F proteins, respectively).

We determined, for each lineage, the number and direction of the amino acid replacements using the ancestral amino acid sequences estimated by the ML marginal reconstruction method (YANG *et al*. 1995). For such analysis we used the accepted phylogenetic tree topology for these species (figure 2). Twenty-tree replacements are placed in the internal branch separating the two paralogous groups, 43 in the OS-E and 33 in the OS-F branches. We analyzed separately radical and conservative replacements (in function of their physicochemical properties; see ZHANG 2000). Under this criterion, 33 amino acid replacements did not change any of their physicochemical properties. Since amino acid replacements generating changes in the net charge could play a critical role in molecular recognition (HUGHES *et al*. 1990), we studied in deep such replacements. We found that thirty of the substitutions resulting in physicochemical changes altered the net charge (10 of them are located in the internal branch separating the two duplicated genes). In addition, two K-V replacements (at positions 81 and 114), located in the branch separating *OS-E* and *OS-F* groups, involve the change of amino acids typically located in the inner part of the proteins to those located in the outer part and, therefore, might cause major changes in the physicochemical properties of the protein.

Figure S1 shows the multiple alignments of the coding regions (Al_OS, Al_OSE and Al_OSF data sets), and table S1 summarizes the estimates of the interspecífic nucleotide divergence. As expected by the gene duplication origin of the OBP genes, divergence levels are higher between paralogous copies. We estimated the divergence time of *OS-E* and *OS-F* gene duplication from information of the number of nucleotide substitutions per site. Assuming 25-55 myr (RUSSO *et al*. 1995; TAMURA *et al*. 2004) for the split of melanogaster and obscura groups this time would be 43-94 Myr.

In agreement with previous results (SÁNCHEZ-GRACIA *et al*. 2003) estimates of synonymous and nonsynonymous divergence at the *OS-E* coding region were higher than those at the *OS-F* (figure 3); in fact, the evolutionary rates of *OS-E* and *OS-F* coding regions are significantly different (RRT, *P* = 0.001). The distribution of nucleotide substitutions along the DNA sequence was not homogeneous: both genes exhibit high levels of nucleotide divergence in the signal-peptide fragment, in the so called heterogeneous region (HEKMAT-SCAFE *et al*. 2000; named here as *het1*), and in the area close to the N-terminal region of the mature protein (named here as *het2*). Figure 4 show the distribution of the average synonymous (K_S) and nonsynonymous (K_A) divergence along the coding region.

Phylogenetic trees (figure 5) show two clearly separated clusters, one for each duplicated gene-copy, suggesting that the two paralogs have evolved independently since the duplication event; in fact, we did not detect any evidence of gene conversion events between paralogous genes. Estimated and accepted (figure 2) topologies are, in fact, not completely concordant. Likely, the low number of substitutions in some lineages would account for this discrepancy; actually, nodes with discrepancies also have low bootstrap values.

We applied the MK test (McDonald and Kreitman 1991) to investigate putative departures of the correlation between polymorphic and fixed synonymous and nonsynonymous substitutions expected under the neutral model. For the analysis, we used polymorphism data from *D. melanogaster* (Sánchez-Gracia *et al*. 2003), *D. simulans* (chapter 3.2) and *D. subobscura* (chapter 3.3), and present estimates of the numbers of synonymous and nonsynonymous substitutions in each branch. For each polymorphism data set we used divergence data from the species belonging to same specific subgroup. We confirm that the significant departures previously detected at the *OS-E* gene using *D. simulans* and *D. subobscura* polymorphism data (chapters 3.2 and 3.3) were promoted by the divergence data of *D. melanogaster* and *D. guanche*, respectively. Since the rest of comparisons (using other outgroups) were not significant, present departures would be caused by an excess of nonsynonymous fixations in *D. melanogaster* and *D. guanche* lineages.

Maximum likelihood analysis at the *OS-E* and *OS-F* coding region:

Models for variable ω ratios among lineages

Table 1 shows the log likelihood (ℓ) values, and ML estimates of the relevant parameters, under all branch-specific models. Estimates of the transition to transversion ratio (κ) were similar among models, but different between the paralogous genes. To assess whether the data is compatible with a homogeneous selective pressure across branches we compared the M0 (one ratio) model, which assumes the same ω ratio for the complete tree, with the free-ratio model (FR), which allows for different ω ratios across tree-branches. Again, the two genes showed different behavior: while for the *OS-E* gene, FR model fits the data significantly better than M0 model ($2\Delta\ell =$ 36.06, d. f. = 22, *P* = 0.03), for the *OS-F* gene the LRT results are not significant; that is, the data is compatible with a single ω ratio for all branches. Estimates of the ω values, nevertheless, indicate that both genes have high and similar levels of functional constraint (ω ~0.06).

Since we had detected departures from the neutral equilibrium model in the *OS-E* gene of *D. melanogaster* and *D. guanche* (SÁNCHEZ-GRACIA *et al.* 2003; chapter 3.2; chapter 3.3), we examined whether the different ω ratios across branches might be attributed to a particular excess of nonsynonymous substitutions in these specific lineages. For such analysis, we applied models with two different ω ratios, one ratio (ω_1) for *D. melanogaster* or *D. guanche* branches (foreground branches), and another (ω_0) for the rest of species (background branches). These models (referred as M0*Dmel* and M0*Dgua*) were compared with the M0 and with the FR model (the models are nested). The simplest one-ratio model was rejected in *D. guanche* ($2\Delta\ell = 11.34$, d. f. = 1, *P* = 0.0008), being *D. melanogaster* not significant ($2\Delta\ell = 3.13$, d. f. = 1, *P* = 0.08). The FR model, however, was not significantly better than the M0*Dgua* model ($2\Delta\ell = 24.72$, d. f. = 21, *P* = 0.26). Consequently, *D. guanche* might explain most of the variation in K_A/K_S ratio detected among *OS-E* branches.

The FR model also fits the data better than the M0 model when all sequences are used (AL_OS data set) ($2\Delta\ell = 70.65$, d. f. = 48, *P* = 0.018). We repeated the *D. guanche* and *D. melanogaster* branch-specific analysis for this complete data set, and again the M0*Dgua* model matched significantly better the data than the FR model. Interestingly, using this data set the ML estimate of the ω_1 under M0*Dgua* model was slightly greater than one (ω_1 = 1.35). Therefore, the *OS-E* gene of *D. guanche* clearly has a distinctive selective behavior.

Models for heterogeneous selection pressure across amino acid sites

All previous analyses were conducted assuming a homogeneous ω ratio across the sequence. This criterion is, nevertheless, very conservative for detecting positive selection. In fact, positive selection will likely affect few amino acid sites in a protein, being the rest highly conserved (with very low ω values). Therefore, for most proteins the ω ratio, averaged for all sites, will likely be lower than 1. To avoid this conservative criterion we applied the random-effects models, implemented in the PAML package, which can account for variation in the selection pressure across codons (NIELSEN and YANG 1998; YANG *et al.* 2000). Table S2 shows the ML estimates separately for the three data sets.

Estimates of κ at the *OS-E* gene were similar among all random effect models (κ ~1.35). The analysis applying models M1 (nearly neutral), M2 (positive selection), and M3 (discrete) indicate that, in this gene, a high proportion of sites (70-90%) are under strong purifying selection (ω values ranging from 0.009 to 0.043), while selection at the rest of sites would be more relaxed (ω < 0.33, except in models M1 and M2 where there is one ω class fixed to 1). Moreover, ML estimates under models M7 (beta) and M8 (beta&ω) shows that ω values have a highly skewed L-shaped beta distribution, being most amino acids highly conserved, or almost invariable (*p* ~0.2 and *q* ~2).

To assess if there are sites evolving under positive selection (ω > 1), we compared models M2 and M1. The ML estimates under these models were identical; therefore, there is no evidence for positive selection at particular codon sites. Discrete models, on the contrary, (M3 with two K=2, or three K=3 ω classes) fitted the data better than model M0. In addition, the estimated parameters values under these models did not shown evidences for a strictly neutral class of sites (ω = 1); therefore, although discrete models can not be compared with M1 –the models are not nested- they likely fit the data better. In addition, results under model M3 (K=3) indicates that there are a small proportion of sites evolving under positive selection (p_2 = 0.012, ω = 1.45), and predicts that one site, a Leucine (L) in position 8 of the AL_OS-E alignment, evolved under positive selection (8L; 0.05 > *P* > 0.01). Nevertheless, this model is not significantly better than the more simple M3 (K=2) model ($2\Delta\ell$ = 2.82, d. f. = 2, *P* = 0.24), providing no evidences for sites evolving at $\omega > 1$ (figure 6; table S2). Model M8 also predicts that a small proportion of sites would also evolve under positive selection (p_1 = 0.008, ω = 1.66), and identifies the same target site (8L) than model M3 (*K*=3). Nevertheless, the posterior probability value at this site is too low, and the standard error of ω too high, for providing unambiguously evidence for positive selection at this site (results not shown).

At the *OS-F* gene, ML estimates of κ (κ ~2.14) were somewhat higher than those at *OS-E*. Nevertheless, for all other parameters, including the ω values across sites, estimates in *OS-E* and *OS-F* were very similar. Models incorporating more than one class of sites (i.e., M1, M2 and M3) also fit reasonably better the data than the one-ratio model [most sites are highly conserved (87-95%), with a very small ω ratio (ω = 0.02-0.04)]. Similarly, M3 models have higher ℓ values than M1 and M2, resulting in exactly the same parameter estimates for M3 (K=3) and M3 (K=2) models (table S2; figure 6). Results for models M7 and M8 (*OS-F* gene) also show the same L-shaped form of the ω distribution.

For the total data set, ML analysis give equivalent results than the previous gene-copy specific analyses, and again model M3 (but in this case with K=3 site classes) is the most likelihood model. There was no evidence for sites evolving under positive selection, and the distribution of the ω across sites also shows the pattern obtained at individual genes (figure 6).

Branch-site models

If the target of selection would affect only few amino acids at specific lineages, previous models would likely fail to detect the footprint of positive selection. We applied the branch-site models A and B (YANG and NIELSEN 2002), which were specifically developed to test for detecting positive selection in particular lineages, only to the complete data set (AL_OS alignment). Model A, which is an extension of the model M1, assumes that there are two classes of sites ($\omega_0 < 0$ and $\omega_1 = 1$) in all lineages, allowing that some sites evolve under positive selection (incorporating the $\omega_2 > 1$ class) along the lineages of interest (foreground branches). Model B, which is an extension of the model M3 (K=2), is similar than model A, but estimates ω_0 and ω_1 (as free parameters) from the data. For the analysis we choose 7 different foreground (figure 7): the two external branches leading to *D. guanche* and *D. melanogaster OS-E* (*guaE* and *melE* respectively), the melanogaster (*me* and *mf*) and obscura (*oe* and *of*) lineages, and the branch separating *OS-E* and *OS-F* clades (*d*).

The results under model A using *of* and *melE* as foreground branches did not show evidences for positive selection (table 2); in fact, model A fit the data worse than model M1 (these models can be compared by using the LRT). For the *oe* foreground branch, the ℓ value under model A was the same than under M1. Although the analysis infers positive selection at 5% of sites, none of them were significant under the Bayesian procedure. Fixing branches *me*, *mf*, *d* and *guaE* as foreground, however, the analysis predicts a number of positive selected sites with a high posterior probability, and fitted the data better than model M1 ($P < 0.006$).

The amino acids target of selection are, nevertheless, different across branches (5W, 45A and 77G for branch *me*; 119V for branch *mf*; 23P, 85A, 114V, 20I, 73N and 4E for branch *d*; 74A, 75I, 78L, 80N and 93E in branch *guaE*). To avoid the assignation of neutrally evolving sites as positively selected, we applied the so called test 2 (ZHANG *et al.* 2005) that compares the model A with a fixed $\omega_2 = 1$ class (named as null model A) and the standard model A. This modified model A fitted the data better than standard model A for all branches with positive results (models A vs. M1), except for *me* and *mf* branches (LRT, $P = 0.003$ and $P = 0.006$, respectively). Therefore, although branch-site models predict a number of positive selected sites in specific lineages, it can not discarded that the $\omega_2 > 1$ values in lineages *d* and *guaE* were, in fact, caused by a relaxation of the functional constraint ($\omega_2 \approx 1$), while positive selection might act in the *me* and *mf* branches.

The branch-site model B, on the other hand, fitted the data significantly better ($P < 0.01$) than discrete model M3 (K=2), for the same four foreground branches with significant results in model A. Indeed, the putative positive selected sites detected are essentially the same for the two branch-site models. Estimates of ω_0 and ω_1 (shared between foreground and background branches) are similar to that obtained under site-specific models.

Fixed-effects approach

The distribution of ω rate across sites in random-effects models might generate false predictions of positive selected sites (SUZUKI and NEI, 2004; ZHANG 2004). This problem could be overcome using the fixed-effects models, that directly estimate the nonsynonymous to synonymous substitution ratio at each site (SUZUKI 2004; KOSAKOVSKY POND and FROST 2005a; MASSINGHAM and GOLDMAN 2005). In addition, if the synonymous substitution rate was not homogeneous across sites, site-by-site methods (as FEL) –which allow variation in both synonymous and nonsynonymous substitution rates– will provide better ω estimates (KOSAKOVSKY POND and FROST 2005a).

Here, we applied the FEL model to our complete data set (Al_OS). Estimates of branch lengths and substitution rates were obtained by ML using the TrN93 (TAMURA and NEI 1993) nucleotide substitution model (the model that best fit the data), while the ω ML estimates at each site were obtained by the MG94 codon substitution model (MUSE and GAUT 1994). The average ω was 0.106. None LRT at individual site were significant for $\omega > 1$. Nevertheless, there were 62 significant sites with evidence of purified selection ($\omega < 1$); these sites are homogeneously distributed along the coding region (figure 8). Interestingly, three of such negative selected sites (23P, 74A and 93E) were identified as positive selected under the branch-site analysis.

Amino acid replacements and 3D structure:

To determine the putative functional role of the positive and negative selected amino acid replacements we studied their location in the 3D protein-structure. For the analysis we aligned the 26 amino acid sequences of this study with that of the pheromone binding protein of *Apis melifera* (Amel-ASP1); this protein gave the best blast score in the 3D Based Homologous Search (24% of identity) and has their 3D structure determined (LARTIGUE *et al.* 2003; PDB accession number 1R5R). The results shows that the *het2* region is located in the N-terminal of the protein (predicted as α-helix by the PHD program), and would correspond with the complete first helix (α–helix A), while the *het1* lies in the region forming the α–helices D and E in the honeybee PBP structure (figure 9).

Results under branch-sites models predicted a number of putative positive-selected evolving sites in different branches (figure 10). Eight of these changes are on (73N, 74A, 75I, 77G, 78L, 80N, 85A) or near (93E) the *het1* region, and thus between α–helices D and E), and include all changes with high posterior probability detected in branch *guaE*. Replacements at positions 4E, 5W, 20I and 23P are located on the *het2* region (corresponding to the α–helix A), positions 114V and 119V are on the C-terminal end of the protein, and position 45A is located at the beginning of the α–helix C in an external position (figure 10). The two replacements (positions 81 and 114), that involves changes from amino acid usually located in the outer part of the protein (K) to the inner part (V), are on the α–helix E and on the C-terminal part of the protein. Finally, sites predicted as negatively selected are homogeneously distributed along the protein structure, and did not show any noticeable pattern.

DISCUSSION

Orthologs of the *OS-E* and *OS-F* genes has been identified in all species studied (seven and six species of the melanogaster and obscura groups, respectively). Like HEKMAT-SCAFE *et al*. (2000), we did not found the ortholog of the *OS-E* gene in *D. virilis* (results not shown). Indeed, the analysis of the amino acid identity and of the intron/exon boundary pattern of two *Anopheles gambiae* genes suggested that this species has the orthologs of *OS-E* and *OS-F* genes (VOGT *et al*. 2002). Phylogenetic analysis, nevertheless, do not support this result (figure 11); instead, *Anopheles* sequences should be co-orthologs of the Drosophila *OS-E* and *OS-F* genes, i.e. they should originate by a gene duplication event after the split of Nematocera and Brachycera species (at least 250 myr). In addition, our estimate for the origin of the duplication event is clearly in agreement both with the phylogenetic analysis and with *D. virilis* results; that is, the duplication event likely occurred after the divergence of the Drosophila-Sophophora subgenera.

We found that the *OS-E* and *OS-F* genes in all Drosophila species are in the same orientation, have similar exon sizes, and show equivalent intron-exon boundary positions. Furthermore, the close physical distance between these genes (intergenic region) is also highly conserved. The maintenance of such paralogous copies for such long period of time is an unexpected feature, and suggests the existence of some mechanism of preservation. Several mechanisms have been proposed to explain the maintenance of duplicated copy members over long periods of time. The most striking difference among them refers to the relative role played by positive selection. OHNO'S (1970) classical view (i.e. the neofunctionalization model; see also CLARK 1994; WALSH 1995) postulates that functional diversification driven by positive selection would preserve the duplicated copies. Gene duplicates can also be stably maintained by the acquisition of independent sub-functions, being all copies needed for providing the original function. This functional subdivision can result from the action of positive selection (JENSEN 1976; ORGEL 1977; PIATIGORSKY and WISTOW 1991; HUGHES 1994), or from the accumulation of degenerative mutations, causing complementary loss of function -without invoking positive selection- (subfunctionalization model, FORCE *et al*. 1999; LYNCH and FORCE 2000).

If changes on the coding region promote functional diversification it could be expected that the two duplicated copies evolve with asymmetric evolutionary rates (e.g. different nonsynonymous substitution rates). Here, we have found that *OS-E* and *OS-F* genes evolved at significantly different rates; nevertheless, the functional constraint levels are very strong and nearly identical in the two copies. Similar functional constraint levels, however, are expected in ancient duplicates (LYNCH and CONERY 2000, 2003). Thus, these genes could undergo differences in their selective pressure on a short period of time after the duplication event, which promoted their functional diversification.

In the present study we have identified 23 replacements in the branch separating *OS-E* and *OS-F* groups (branch *d*, figure 7). Ten are radical in terms of charge and 14 in terms of polarity (one of them fitted both criteria).

Interestingly, five of these replacements (4E, 20I, 23P, 73N and 114V) could have evolved by positive selection (inferred by the branch-site analysis, models A and B). Although, the null model A was not rejected (i.e., we can not completely discard that they correspond to a relaxation of the functional constraint), the test is highly conservative (see below), mainly for weak selection (i.e. ω close to 1). Yet, 114V is one of the two replacements that might produce relevant changes in the function of the molecule. It is suggestive, therefore, that some of these 23 radical changes fixed between the *OS-E* and *OS-F* genes, might have evolved promoted by positive selection, and might have an important role in the functional diversification of the duplicated members after the gene duplication event.

Three of the radical replacements predicted as positive selected (4E, 20I and 23P) are located in a region (*het2* region), likely involved in the formation of the first α-helix of the OBP structure (figures 9 and 10); replacements in this region can alter the position of the first disulfide bridge modifying the size of the binding cavity (LARTIGUE *et al*. 2003; TEGONI *et al*. 2004). For instance, in LUSH (an OBP of *D. melanogaster* where the 3D structure has been determined), the first α-helix is in a more internal position than in Amel-ASP1 and has a small binding cavity. These changes can affect significantly the binding proprieties of these molecules. In addition, another putative positive selected replacement (5W) would place in this first α-helix, consistent with the hypothesis that the N-terminal part of the protein could be a major target for adaptive changes. Therefore, amino acid replacements in this region (as the four putative positive selected sites) might have been relevant for the functional diversification between OS-E and OS-F proteins.

The *het1* region lies in the α-helices D and E, which in other OBPs includes hydrophobic residues covering the binding cavity. Therefore, changes in this region might also have a major functional significance. Replacement 73N might be located close to the α-helix D end (figures 9 and 10; chapter 3.3), nevertheless, it is difficult to know whether this specific residue is in the internal part of the structure and thus exposed to the ligand.

We have found a distinctive evolutionary pattern in *D. guanche*: all substitutions increasing the average ω value in this lineage (branch models and branch-site models, tables 1 and 2), are placed in the *het1* region. Several authors suggested (LLOPART *et al* 1999; PÉREZ *et al*. 2003) that the reduction in the effective population size in this insular species would have increased the fixation probability of slightly deleterious mutations (as “unpreferent” synonymous mutations, or amino acid replacements). Under this scenario, the less constrained part of the protein would accumulate a higher number of substitutions. Consistent with this prediction, we have found that sites with the greatest ω ratio were clustered in the signal peptide, and in the *het1* region (figure 12). Hence, although in this species we have found ω values higher than 1, we should be careful with the evolutionary meaning, since it is difficult to discriminate the relative role of positive selection and relaxed purified selection. Nonetheless, the number of nonsynonymous substitutions in the *het1* and *het2* (*het*1/*het*2) regions are uneven distributed among branches (branch *d*, 8/10;

rest of the branches, 41/16; $\chi^2 = 4.563$, P = 0.03) suggesting that the evolutionary pattern of these two regions were dissimilar at different periods of time.

.

The amino acid replacement at position 114 lies in the C-terminal part of the protein. In Amel-ASP1, this domain folds inside the protein forming one binding cavity wall and contains residues that directly interact with the ligand (LARTIGUE *et al*. 2003). Conformational changes of this part trigger the release of the ligand close to the odorant receptor (HORST *et al*. 2001; LEAL *et al*. 2005). Interestingly, this part of the protein is the most divergent among the 4 OBPs with resolved structures: it presents different lengths, different secondary structure, or even it is missing (TEGONI *et al*. 2004). Replacement at site 119, identified as positive selected in the branch *mf* (figure 7), is also located in this C-terminal part. Current information suggests that replacements in this region could have evolved by positive selection because of their effect in changing the OBP-ligand interaction between OS-E and OS-F.

In spite that the reported radical changes are the most likelihood to produce important functional changes between the OS-E and OS-F proteins, we can not discard that conservative replacements also have an important role. In fact, IVARRSON *et al.* (2003) showed that a biochemical conservative replacement, identified as positively selected using codon-evolution models, caused a measurable increase in the activity of an enzyme. Therefore, conservative changes could cause weakly but important functional changes in the OBP binding activity, or in the ligand specificity, that in turn might be adaptive.

We should notice that only the ML branch-site model detected positive selected sites. It has been demonstrated (ZHANG 2004; SUZUKI and NEI 2004) that violations of the underlined assumptions (a few number of ω classes, an equal κ value for synonymous and nonsynonymous substitutions, an equal proportion of sites belonging to certain classes in foreground and background branches, or the phylogenetic tree performance) could generate frequent false positives results. Therefore, we can not discard that sites predicted as positive selected in the branch-site analysis were spurious. Model B, however, performs reasonably well in all cases, but seems to be unable to distinguishing between positive and relaxed purifying selection in the foreground branch (ZHANG 2004). To avoid this problem we used the new model A, which modifies the older version of model A (YANG *et al*. 2005) to be more realistic, and it could be compared with the null model A (fixing $\omega_2 = 1$). This test (named as test 2 in YANG *et al*. 2005), nevertheless, is highly conservative and appears to be a direct test of positive selection. This test rejected the null hypothesis of complete relaxation in *me* and *mf* branches, and therefore we can conclude that positive selection likely acted in these lineages. In *D. guanche* and branch *d*, the test 2 was not significant; therefore predicted positive selected substitutions in these lineages could be selectively neutral. One possibility would be that these substitutions were complementary degenerative mutations causing some form of subfunctionalization (Force et al. 1999; Lynch and Force 2000). Since, branch *d* includes substitutions occurred after duplication event, in both *OS-E* and *OS-F* lineages (two different branches), the putative positive selected substitutions in one branch can be compensated by the action of purifying selection in the other. To distinguish these factors we polarize the nonsynonymous substitutions of branch *d* using as outgroup the OBP nucleotide sequence of the mosquito *Culex pipiens* (GenBank Accession number AF468212). In this new analysis, branch *d* is splitted into two internal branches, one leading to the *OS-E* and the other to the *OS-F* (figure 13). The test 2 gave a significant result (*P* = 0.01) fixing the *OS-E* internal branch as the foreground. Furthermore, the amino acids predicted as positive selected are nearly the same than those inferred using branch *d* as foreground. Therefore, positive selection might be involved in the functional diversification of the OS-E protein, which in turn would be the responsible of the current differences in the evolutionary rate between *OS-E* and *OS-F* coding regions.

Although present analysis allows inferring the footprint of positive selection in the coding region of the *OS-E* and *OS-F*, the evidences are weak. The putative positive-selected sites promoting the genetic differentiation between duplicates might have occurred in untranslated or noncoding regions, and therefore, invisible to present investigation. In *D. melanogaster*, the two *OS* genes have not spatial differences in gene expression (McKenna *et al* 1994; Hekmat-Scafe *et al*. 1997), avoiding any possibility of functional diversification at this level. On the other hand, since the relative gene expression levels of the *OS-E* and *OS-F* are unknown we can not determine whether there exist differences in the amount of product. It has been suggested, however, that differences in intron sizes, especially in the first intron, could lead to differences in expression levels and in the evolutionary rate (Marais *et al.* 2005); the presence of some regulatory sequences in these introns could explain this effect. The *OS-F* gene of *D. melanogaster* has a large first intron (~1.7kb), which separates the two first exons (the 5’ untranslated and the first translated exon), that is absent in the *OS-E* gene. Putative regulatory sequences in this intron could cause differences in the gene expression pattern between genes. In fact, we identified two highly conserved fragments in the first intron of *OS-F* gene (chapter 3.3), that might represent regulatory regions.

Although the preservation of the *OS-E* and *OS-F* gene duplication structure might be promoted by positive selection, at least just after the duplication event, further functional experiments should be necessary to demonstrate the adaptive character of the amino acid inferred positive selected by the ML analysis. Furthermore, it will be also indispensable the analysis of flanking non coding regions, such the putative regulatory sequences of these genes, and to investigate for specific differences in the gene regulation. All these studies and experiments will certainly contribute to the understanding of the precise role of natural selection in the molecular evolution of *OS-E* and *OS-F* gene duplication.

LITERATURE CITED

AKASHI, H., 1995 Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. Genetics 139: 1067-1076.

ANISIMOVA, M., J. P. BIELAWSKI and Z. YANG, 2001 Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18: 1585-1592.

ANISIMOVA, M., J. P. BIELAWSKI and Z. YANG, 2002 Accuracy and power of bayes prediction of amino acid sites under positive selection. Mol Biol Evol 19: 950-958.

ANISIMOVA, M., R. NIELSEN and Z. YANG, 2003 Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. Genetics 164: 1229-1236.

ASHBURNER, M., 1989 *Drosophila: A laboratory handbook*. Cold Spring Harbor Laboratory Press, New York.

BERMAN, H. M., T. N. BHAT, P. E. BOURNE, Z. FENG, G. GILLILAND et al., 2000 The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7 Suppl: 957-959.

BISHOP, J. G., 2005 Directed mutagenesis confirms the functional importance of positively selected sites in Polygalacturonase Inhibitor protein. Mol Biol Evol 22: 1531-1534.

CLARK, A. G., 1994 Invasion and maintenance of a gene duplication. Proc Natl Acad Sci U S A 91: 2950-2954.

CLARK, A. G., S. GLANOWSKI, R. NIELSEN, P. D. THOMAS, A. KEJARIWAL et al., 2003 Inferring nonneutral evolution from Human-Chimp-Mouse orthologous gene trios. Science 302: 1960-1963.

EMES, R. D., S. A. BEATSON, C. P. PONTING and L. GOODSTADT, 2004 Evolution and comparative genomics of odorant- and pheromone-associated genes in rodents. Genome Res 14: 591-602.

FORCE, A., M. LYNCH, F. B. PICKETT, A. AMORES, Y. L. YAN et al., 1999 Preservation of duplicate genes by complementary, degenerative mutations. Genetics 151: 1531-1545.

GILAD, Y., C. D. BUSTAMANTE, D. LANCET and S. PAABO, 2003 Natural selection on the olfactory receptor gene family in humans and chimpanzees. Am J Hum Genet 73: 489-501.

GIMELBRANT, A. A., H. SKALETSKY and A. CHESS, 2004 Selective pressures on the olfactory receptor repertoire since the human-chimpanzee divergence. Proc Natl Acad Sci U S A 101: 9019-9022.

HEKMAT-SCAFE, D. S., R. A. STEINBRECHT and J. R. CARLSON, 1997 Coexpression of two odorant-binding protein homologs in Drosophila: implications for olfactory coding. J Neurosci 17: 1616-1624.

HEKMAT-SCAFE, D. S., R. L. DORIT and J. R. CARLSON, 2000 Molecular evolution of odorant-binding protein genes *OS-E* and *OS-F* in Drosophila. Genetics 155: 117-127.

HEKMAT-SCAFE, D. S., C. R. SCAFE, A. J. MCKINNEY and M. A. TANOUYE, 2002 Genome-wide analysis of the odorant-binding protein gene family in *Drosophila melanogaster*. Genome Res 12: 1357-1369.

HORST, R., F. DAMBERGER, P. LUGINBUHL, P. GUNTERT, G. PENG et al., 2001 NMR structure reveals intramolecular regulation mechanism for pheromone binding and release. Proc Natl Acad Sci U S A 98: 14374-14379.

HUELSENBECK, J. P., and K. A. DYER, 2004 Bayesian estimation of positively selected sites. J Mol Evol 58: 661-672.

HUGHES, A. L., T. OTA and M. NEI, 1990 Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol Biol Evol 7: 515-524.

HUGHES, A. L., 1994 The evolution of functionally novel proteins after gene duplication. Proc Biol Sci 256: 119-124.

HUYNEN, M., T. DOERKS, F. EISENHABER, C. ORENGO, S. SUNYAEV et al., 1998 Homology-based fold predictions for *Mycoplasma genitalium* proteins. J Mol Biol 280: 323-326.

IVARSSON, Y., A. J. MACKEY, M. EDALAT, W. R. PEARSON and B. MANNERVIK, 2003 Identification of residues in glutathione transferase capable of driving functional diversification in evolution. A novel approach to protein redesign. J Biol Chem 278: 8733-8738.

JENSEN, R. A., 1976 Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30: 409-425.

KOSAKOVSKY POND, S. L., and S. D. W. FROST, 2005a Not so different after all: A comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22: 1208-1222.

KOSAKOVSKY POND, S. L., and S. D. W. FROST, 2005b Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21: 2531-2533.

KRIEGER, M. J., and K. G. ROSS, 2002 Identification of a major gene regulating complex social behavior. Science 295: 328-332.

KUMAR, S., K. TAMURA and M. NEI, 2004 MEGA3: Integrated software for Molecular Evolutionary Genetics Analysis and sequence alignment. Brief Bioinform 5: 150-163.

LARTIGUE, A., A. GRUEZ, L. BRIAND, F. BLON, V. BEZIRARD et al., 2003 Sulfur-SAD crystal structure of a pheromone binding protein from the honeybee *Apis mellifera* L. J Biol Chem 279: 4459-4464.

LEAL, W. S., A. M. CHEN, Y. ISHIDA, V. P. CHIANG, M. L. ERICKSON et al., 2005 Kinetics and molecular properties of pheromone binding and release. Proc Natl Acad Sci U S A 102: 5386-5391.

LIBERLES, D. A., D. R. SCHREIBER, S. GOVINDARAJAN, S. G. CHAMBERLIN and S. A. BENNER, 2001 The adaptive evolution database (TAED). Genome Biol 2: Research0028.1–Research0028.6.

LIBERLES, D. A., and M. L. WAYNE, 2002 Tracking adaptive evolutionary events in genomic sequences. Genome Biol 3: Reviews1018.1-1018.4.

LLOPART, A., and M. AGUADÉ, 1999 Synonymous rates at the *RpII215* gene of Drosophila: variation among species and across the coding region. Genetics 152: 269-280.

LYNCH, M., and J. S. CONERY, 2000 The evolutionary fate and consequences of duplicate genes. Science 290: 1151-1155.

LYNCH, M., and A. FORCE, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics 154: 459-473.

LYNCH, M., and J. S. CONERY, 2003 The evolutionary demography of duplicate genes. J Struct Funct Genomics 3: 35-44.

MADDISON, W. P., and D. R. MADDISON, 1992 *MacClade: Analysis of Phylogeny and Character Evolution*. Version 3.0. Sinauer, Sunderland, MA.

MARAIS, G., P. NOUVELLET, P. D. KEIGHTLEY and B. CHARLESWORTH, 2005 Intron size and exon evolution in Drosophila. Genetics 170: 481-485.

MASSINGHAM, T., and N. GOLDMAN, 2005 Detecting amino acid sites under positive selection and purifying selection. Genetics 169: 1753-1762.

MCDONALD, J. H., and M. KREITMAN, 1991 Adaptive protein evolution at the *Adh* locus in Drosophila. Nature 351: 652-654.

MCKENNA, M. P., D. S. HEKMAT-SCAFE, P. GAINES and J. R. CARLSON, 1994 Putative Drosophila pheromone-binding proteins expressed in a subregion of the olfactory system. J Biol Chem 269: 16340-16347.

MUSE, S. V., and B. S. GAUT, 1994 A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11: 715-724.

NGAI, J., M. M. DOWLING, L. BUCK, R. AXEL and A. CHESS, 1993 The family of genes encoding odorant receptors in the channel catfish. Cell 72: 657-666.

NIELSEN, R., and Z. YANG, 1998 Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936.

NIELSEN, R., and J. P. HUELSENBECK, 2002 Detecting positively selected amino acid sites using posterior predictive *P*-values. Pac Symp Biocomput: 576-588.

NIELSEN, R., 2002 Mapping mutations on phylogenies. Syst Biol 51: 729-739.

OHNO, S., 1970 *Evolution by gene duplication*. Springer, Berlin.

ORGEL, L. E., 1977 Gene-duplication and the origin of proteins with novel functions. J Theor Biol 67: 773.

PÉREZ, J. A., A. MUNTÉ, J. ROZAS, C. SEGARRA and M. AGUADÉ, 2003 Nucleotide polymorphism in the *RpII215* gene region of the insular species *Drosophila guanche*: Reduced efficacy of weak selection on synonymous variation. Mol Biol Evol 20: 1867-1875.

PIATIGORSKY, J., and G. WISTOW, 1991 The recruitment of crystallins: new functions precede gene duplication. Science 252: 1078-1079.

RICHARDS, S., Y. LIU, B. R. BETTENCOURT, P. HRADECKY, S. LETOVSKY et al., 2005 Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and cis-element evolution. Genome Res 15: 1-18.

ROST, B., 1996 PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266: 525-539.

ROST, B., 2001 Review: protein secondary structure prediction continues to rise. J Struct Biol 134: 204-218.

ROZAS, J., J. C. SÁNCHEZ-DELBARRIO, X. MESSEGUER and R. ROZAS, 2003 DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics 19: 2496-2497.

RUSSO, C., N. TAKEZAKI and M. NEI, 1995 Molecular phylogeny and divergence times of drosophilid species. Mol Biol Evol 12: 391-404.

SAIKI, R. K., D. H. GELFAND, S. STOFFEL, S. J. SCHARF, R. HIGUCHI et al., 1988 Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239: 487-491.

SAITOU, N., and M. NEI, 1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4: 406-425.

SÁNCHEZ-GRACIA, A., M. AGUADÉ and J. ROZAS, 2003 Patterns of nucleotide polymorphism and divergence in the Odorant-Binding Protein genes *OS-E* and *OS-F*: Analysis in the Melanogaster Species Subgroup of Drosophila. Genetics 165: 1279-1288.

SAWYER, S. L., L. I. WU, M. EMERMAN and H. S. MALIK, 2005 Positive selection of primate *TRIM5alpha* identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A 102: 2832-2837.

STOLTZFUS, A., 1999 On the possibility of constructive neutral evolution. J Mol Evol 49: 169-181.

SUZUKI, Y., and T. GOJOBORI, 1999 A method for detecting positive selection at single amino acid sites. Mol Biol Evol 16: 1315-1328.

SUZUKI, Y., and M. NEI, 2001 Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. Mol Biol Evol 18: 2179-2185.

SUZUKI, Y., and M. NEI, 2002 Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino acid sites. Mol Biol Evol 19: 1865-1869.

SUZUKI, Y., 2004 New methods for detecting positive selection at single amino acid sites. J Mol Evol 59: 11-19.

SUZUKI, Y., and M. NEI, 2004 False-positive selection identified by ML-based methods: examples from the *Sig1* gene of the diatom *Thalassiosira weissflogii* and the *tax* gene of a human T-cell lymphotropic virus. Mol Biol Evol 21: 914-921.

TAJIMA, F., 1993 Simple methods for testing the molecular evolutionary clock hypothesis. Genetics 135: 599-607.

TAKAHASHI, A., and T. TAKANO-SHIMIZU, 2005 A High-frequency null mutant of an Odorant-Binding Protein gene, *Obp57e*, in *Drosophila melanogaster*. Genetics 170: 709-718.

TAMURA, K., and M. NEI, 1993 Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol Biol Evol 10: 512-526.

TAMURA, K., S. SUBRAMANIAN and S. KUMAR, 2004 Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks. Mol Biol Evol 21: 36-44.

TEGONI, M., P. PELOSI, F. VINCENT, S. SPINELLI, V. CAMPANACCI, S. GROLLI, R. RAMONI, and C. CAMBILLAU, 2000 Mammalian odorant-binding proteins. Biochim Biophys Acta 1482: 229–240.

TEGONI, M., V. CAMPANACCI and C. CAMBILLAU, 2004 Structural aspects of sexual attraction and chemical communication in insects. Trends Biochem Sci 29: 257-264.

THOMPSON, J. D., D. G. HIGGINS and T. J. GIBSON, 1994 CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22: 4673-4680.

VOGT, R. G., 2002 Odorant binding protein homologues of the malaria mosquito *Anopheles gambiae*; possible orthologues of the OS-E and OS-F OBPs of *Drosophila melanogaster*. J Chem Ecol 28: 2371-2376.

VOGT, R. G., 2005 Molecular basis of pheromone detection in insects. In *Comprehensive Insect Physiology, Biochemistry, Pharmacology and Molecular*

Biology. Volume 3. Endocrinology. (L.I. GILBERT, K. IATRO and S. GILL eds). pp. 753-804. Elsevier, London.

WALSH, J. B., 1995 How often do duplicated genes evolve new functions? Genetics 139: 421-428.

WATTS, R. A., C. A. PALMER, R. C. FELDHOFF, P. W. FELDHOFF, L. D. HOUCK et al., 2004 Stabilizing selection on behavior and morphology masks positive selection on the signal in a salamander pheromone signaling complex. Mol Biol Evol 21: 1032-1041.

WHELAN, S., and N. GOLDMAN, 1999 Distributions of statistics used for the comparison of models of sequence evolution in phylogenetics. Mol Biol Evol 16: 1292-1299.

WILLETT, C. S., 2000 Evidence for directional selection acting on pheromone-binding proteins in the genus Choristoneura. Mol Biol Evol 17: 553-562.

WONG, W. S. W., Z. YANG, N. GOLDMAN and R. NIELSEN, 2004 Accuracy and power of statistical methods for detecting adaptive evolution in protein coding sequences and for identifying positively selected sites. Genetics 168: 1041-1051.

YANG, Z., S. KUMAR and M. NEI, 1995 A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141: 1641-1650.

YANG, Z., 1997 PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13: 555-556.

YANG, Z., 1998 Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15: 568-573.

YANG, Z., and R. NIELSEN, 1998 Synonymous and nonsynonymous rate variation in nuclear genes of mammals. J Mol Evol 46: 409-418.

YANG, Z., and J. P. BIELAWSKI, 2000 Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496-503.

YANG, Z., R. NIELSEN, N. GOLDMAN and A. M. PEDERSEN, 2000 Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics 155: 431-449.

YANG, Z., and W. J. SWANSON, 2002 Codon-substitution models to detect adaptive evolution that account for heterogeneous selective pressures among site classes. Mol Biol Evol 19: 49-57.

YANG, Z., and R. NIELSEN, 2002 Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908-917.

YANG, Z., W. S. WONG and R. NIELSEN, 2005 Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107-1118.

ZHANG, J., 2000 Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. J Mol Evol 50: 56-68.

ZHANG, J., 2004 Frequent false detection of positive selection by the likelihood method with branch-site models. Mol Biol Evol 21: 1332-1339.

TABLE 1

Parameter estimates for the branch models

Data	Model	*f*	*ℓ*	Estimates of parameters	Positive selection
OS-E	M0	25	-1733.97	ω = 0.072, κ=1.31	None
	FR	47	-1715.94	κ=1.36	None
	M0*Dmel*	26	-1732.41	ϖ =0.068, ω_l = 0.205, κ=1.32	None
	M0*Dgua*	26	-1728.30	ϖ =0.063, ω_l = 0.508, κ=1.33	None
OS-F	M0	25	-1398.27	ω = 0.058, κ=2.12	None
	FR	47	-1390.12	κ=2.12	None
mature	M0	51	-2245,69	ω = 0.045, κ=1.23	None
	FR	99	-2210,36	κ=1.24 *D. guanche* ω=**1.63**	Yes
	M0*Dgua*	52	-2238,89	ϖ =0.042, ω_l = **1.35**, κ=1.25	Yes

f, number of free parameters

TABLE 2

Parameter estimates in branch-site model A[a]

Foreground branch	ℓ	Estimates of parameters	Positive selection
me	-2226.50	p_0=0.94, p_1=0.03 (p_2+p_3)=0.03, ω_2=998.97	5W**, 45A**, 77G*
oe	-2236.72	p_0=0.92, p_1=0.03 (p_2+p_3)=0.05, ω_2=2.61	none
mf	-2230.98	p_0=0.96, p_1=0.03 (p_2+p_3)=0.01, ω_2=999	119V*
of	-2235.40	p_0=0.96, p_1=0.04 (p_2+p_3)=0.00, ω_2=1.00	11L, 78L, 96T, 99H: (P > 0.72)
guaE	-2231.57	p_0=0.58, p_1=0.02 (p_2+p_3)=0.40, ω_2=1.97	74A, 75I, 78L, 80N, 93E: (P > 0.92)
melE	-2234.59	p_0=0.80, p_1=0.03 (p_2+p_3)=0.17, ω_2=1.00	14G, 71V, 72L, 82M, 90I: (P > 0.66)
d	-2227.34	p_0=0.83, p_1=0.03 (p_2+p_3)=0.14, ω_2=1.37	23P*, 85A*, 114V* 20I, 73N: (P > 0.93) 4E: (P > 0.84) 3G, 9A: (P > 0.54)

p, proportion of sites for each site class. Posterior probabilities: * *P* > 0.95, ** *P* > 0.99. [a] branch-site model A applied to Al_OS data contains 54 free parameters

FIGURE LEGENDS

Figure 1. Multiple alignment of the OS-E and OS-F mature proteins. Black shades indicate identical amino acids. Grey shades show amino acids with a similar biochemical properties. mel, *D. melanogaster*. sim, *D. simulans*. mau, *D. mauritiana*. tei, *D. teissieri*. yak, *D. yakuba*. ana, *D. ananassae*. gua, *D. guanche*. mad, *D. madeirensis*. sub, *D. subobscura*. mir, *D. miranda*. per, *D. persimilis*. pse, *D. pseudoobscura*. bif, *D. bifasciata.*

Figure 2. Accepted topology used for the phylogenetic codon analyses. mel, *D. melanogaster*. sim, *D. simulans*. mau, *D. mauritiana*. tei, *D. teissieri*. yak, *D. yakuba*. ana, *D. ananassae*. gua, *D. guanche*. mad, *D. madeirensis*. sub, *D. subobscura*. mir, *D. miranda*. per, *D. persimilis*. pse, *D. pseudoobscura*. bif, *D. bifasciata.*

Figure 3. Estimated pairwise synonymous (K_S) and nonsynonymous (K_A) divergence for the *OS-E* and *OS-F* coding regions.

Figure 4. Sliding window of the average nucleotide divergence along the coding region (Al_OS data set). K_A, nonsynonymous divergence. K_S, synonymous divergence. Window length, 50 sites; Step size, 10 sites.

Figure 5. Phylogenetic trees for the *OS-E* and *OS-F* coding regions. Trees were built using information of (A) Al_OS, (B) Al_OSE and (C) Al_OSF data sets. Numbers indicate the percentage of bootstrap values supporting the nodes. mel, *D. melanogaster*. sim, *D. simulans*. mau, *D. mauritiana*. tei, *D. teissieri*. yak, *D. yakuba*. ana, *D. ananassae*. gua, *D. guanche*. mad, *D. madeirensis*. sub, *D. subobscura*. mir, *D. miranda*. per, *D. persimilis*. pse, *D. pseudoobscura*. bif, *D. bifasciata*.

Figure 6. ML estimates of the proportion of sites (p) with a particular K_A/K_S ratio (ω) under the best fitted models.

Figure 7. Foreground branches used in the branch-site analyses.

Figure 8. Distribution of *P*-values (H_0: K_A-K_S = 0) of the LRT at individual sites across the coding region.

Figure 9. Predicted location of the two heterogeneous regions (in yellow): (A) *het2* and (B) *het1*), on the 3D structure of AmelASP1. α, α-helix. C, C-terminal. N, N-terminal.

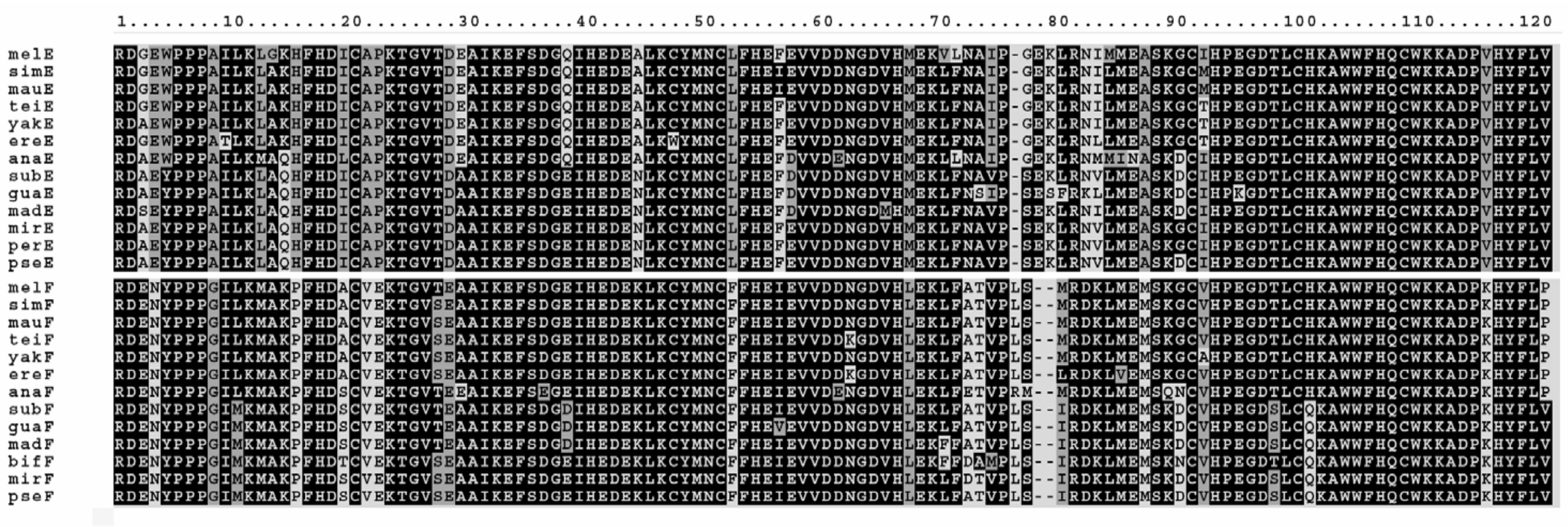
Figure 10. Putative location of the amino acid sites (in yellow) predicted to be under positive natural selection in the branch-site analysis.

Figure 11. Phylogenetic tree for the *OS-E* and *OS-F* coding regions in different Diptera species. The tree was built using the total number of nucleotide substitutions. Numbers indicate the percentage of bootstrap values supporting the nodes. AF437886-AF437884, predicted orthologs of the *OS-E* and *OS-F* genes in *Anopheles gambiae* (VOGT *et al*. 2002).

Figure 12. Distribution of the estimated ω values at individual sites along the coding region of the *OS-E* and *OS-F* genes under the M3 (K=3) model (Al_*OS* data set).

Figure 13. Foreground branches for the ML branch-site analysis using the *Culex pipiens* OBP coding sequence as the outgroup.

FIGURE 1



```
      1.........10........20........30........40........50........60........70........80........90.......100.......110.......120
melE  RDGEWPPPAILKLGKHFHDICAPKTGVTDEAIKEFSDGQIHEDEALKCYMNCLFHEFEVVDDNGDVHMEKVLNAIP-GEKLRNIMMEASKGCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV
simE  RDGEWPPPAILKLAKHFHDICAPKTGVTDEAIKEFSDGQIHEDEALKCYMNCLFHEIEVVDDNGDVHMEKLFNAIP-GEKLRNILMEASKGCMHPEGDTLCHKAWWFHQCWKKADPVHYFLV
mauE  RDGEWPPPAILKLAKHFHDICAPKTGVTDEAIKEFSDGQIHEDEALKCYMNCLFHEIEVVDDNGDVHMEKLFNAIP-GEKLRNILMEASKGCMHPEGDTLCHKAWWFHQCWKKADPVHYFLV
teiE  RDGEWPPPAILKLAKHFHDICAPKTGVTDEAIKEFSDGQIHEDEALKCYMNCLFHEFEVVDDNGDVHMEKLFNAIP-GEKLRNILMEASKGCTHPEGDTLCHKAWWFHQCWKKADPVHYFLV
yakE  RDAEWPPPAILKLAKHFHDICAPKTGVTDEAIKEFSDGQIHEDEALKCYMNCLFHEFEVVDDNGDVHMEKLFNAIP-GEKLRNILMEASKGCTHPEGDTLCHKAWWFHQCWKKADPVHYFLV
ereE  RDGEWPPPATLKLAKHFHDICAPKTGVTDEAIKEFSDGQIHEDEALKWYMNCLFHEFEVVDDNGDVHMEKLFNAIP-GEKLRNLLMEASKGCTHPEGDTLCHKAWWFHQCWKKADPVHYFLV
anaE  RDAEWPPPAILKMAQHFHDLCAPKTGVTDEAIKEFSDGQIHEDEALKCYMNCLFHEFDVVDENGDVHMEKLLNAIP-GEKLRNMMINASKDCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV
subE  RDAEYPPPAILKLAQHFHDICAPKTGVTDAAIKEFSDGEIHEDENLKCYMNCLFHEFDVVDDNGDVHMEKLFNAVP-SEKLRNVLMEASKDCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV
guaE  RDAEYPPPAILKLAQHFHDICAPKTGVTDAAIKEFSDGEIHEDENLKCYMNCLFHEFDVVDDNGDVHMEKLFNSIP-SESFRKLLMEASKDCIHPKGDTLCHKAWWFHQCWKKADPVHYFLV
madE  RDSEYPPPAILKLAQHFHDICAPKTGVTDAAIKEFSDGEIHEDENLKCYMNCLFHEFDVVDDNGDMHMEKLFNAVP-SEKLRNILMEASKDCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV
mirE  RDAEYPPPAILKLAQHFHDICAPKTGVTDAAIKEFSDGEIHEDENLKCYMNCLFHEFEVVDDNGDVHMEKLFNAVP-SEKLRNVLMEASKGCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV
perE  RDAEYPPPAILKLAQHFHDICAPKTGVTDAAIKEFSDGEIHEDENLKCYMNCLFHEFEVVDDNGDVHMEKLFNAVP-SEKLRNVLMEASKGCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV
pseE  RDAEYPPPAILKLAQHFHDICAPKTGVTDAAIKEFSDGEIHEDENLKCYMNCLFHEFEVVDDNGDVHMEKLFNAVP-SEKLRNVLMEASKDCIHPEGDTLCHKAWWFHQCWKKADPVHYFLV

melF  RDENYPPPGILKMAKPFHDACVEKTGVTEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKLFATVPLS--MRDKLMEMSKGCVHPEGDTLCHKAWWFHQCWKKADPKHYFLP
simF  RDENYPPPGILKMAKPFHDACVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKLFATVPLS--MRDKLMEMSKGCVHPEGDTLCHKAWWFHQCWKKADPKHYFLP
mauF  RDENYPPPGILKMAKPFHDACVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKLFATVPLS--MRDKLMEMSKGCVHPEGDTLCHKAWWFHQCWKKADPKHYFLP
teiF  RDENYPPPGILKMAKPFHDACVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDKGDVHLEKLFATVPLS--MRDKLMEMSKGCVHPEGDTLCHKAWWFHQCWKKADPKHYFLP
yakF  RDENYPPPGILKMAKPFHDACVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKLFATVPLS--MRDKLMEMSKGCAHPEGDTLCHKAWWFHQCWKKADPKHYFLP
ereF  RDENYPPPGILKMAKPFHDACVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDKGDVHLEKLFATVPLS--LRDKLVEMSKGCVHPEGDTLCHKAWWFHQCWKKADPKHYFLP
anaF  RDENYPPPGILKMAKPFHDSCVEKTGVTEEAIKEFSEGEIHEDEKLKCYMNCFFHEIEVVDENGDVHLEKLFETVPRM--MRDKLMEMSQNCVHPEGDTLCHKAWWFHQCWKKADPKHYFLP
subF  RDENYPPPGIMKMAKPFHDSCVEKTGVTEAAIKEFSDGDIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKLFATVPLS--IRDKLMEMSKDCVHPEGDSLCQKAWWFHQCWKKADPKHYFLV
guaF  RDENYPPPGIMKMAKPFHDSCVEKTGVTEAAIKEFSDGDIHEDEKLKCYMNCFFHEVEVVDDNGDVHLEKLFATVPLS--IRDKLMEMSKDCVHPEGDSLCQKAWWFHQCWKKADPKHYFLV
madF  RDENYPPPGIMKMAKPFHDSCVEKTGVTEAAIKEFSDGDIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKFFATVPLS--IRDKLMEMSKDCVHPEGDSLCQKAWWFHQCWKKADPKHYFLV
bifF  RDENYPPPGIMKMAKPFHDTCVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKFFDAMPLS--IRDKLMEMSKNCVHPEGDTLCQKAWWFHQCWKKADPKHYFLV
mirF  RDENYPPPGIMKMAKPFHDSCVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKLFDTVPLS--IRDKLMEMSKDCVHPEGDSLCQKAWWFHQCWKKADPKHYFLV
pseF  RDENYPPPGIMKMAKPFHDSCVEKTGVSEAAIKEFSDGEIHEDEKLKCYMNCFFHEIEVVDDNGDVHLEKLFATVPLS--IRDKLMEMSKDCVHPEGDSLCQKAWWFHQCWKKADPKHYFLV
```

FIGURE 2

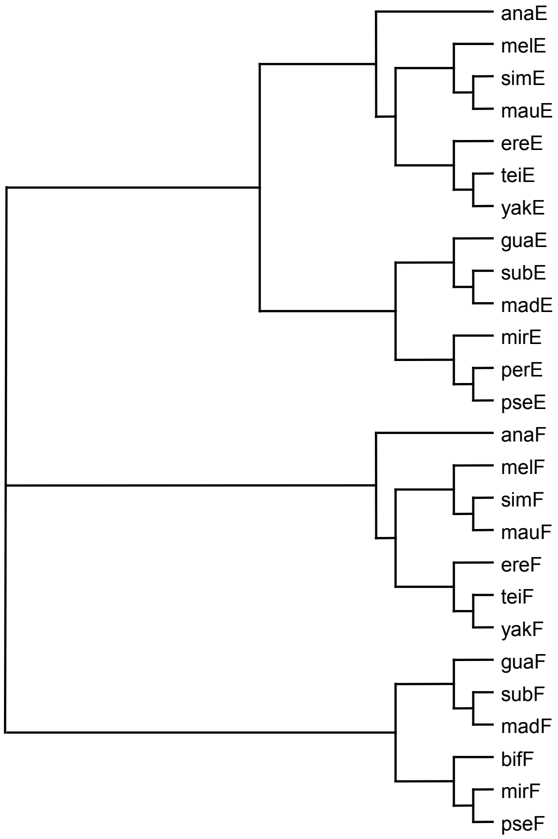

anaE
melE
simE
mauE
ereE
teiE
yakE
guaE
subE
madE
mirE
perE
pseE
anaF
melF
simF
mauF
ereF
teiF
yakF
guaF
subF
madF
bifF
mirF
pseF

FIGURE 3

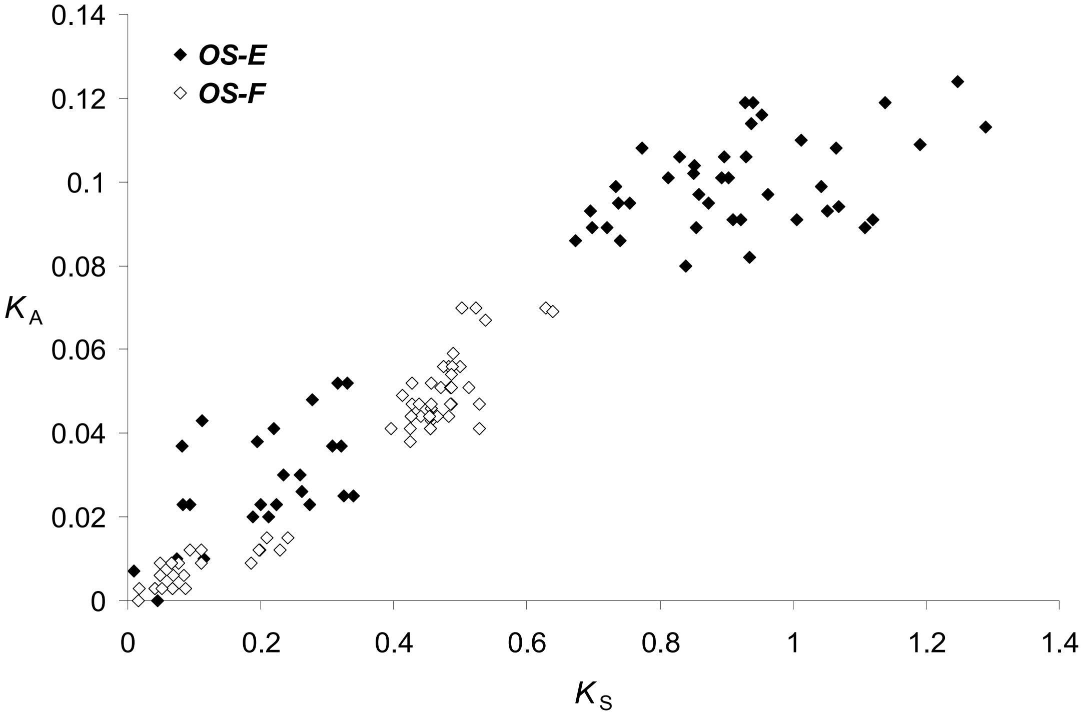
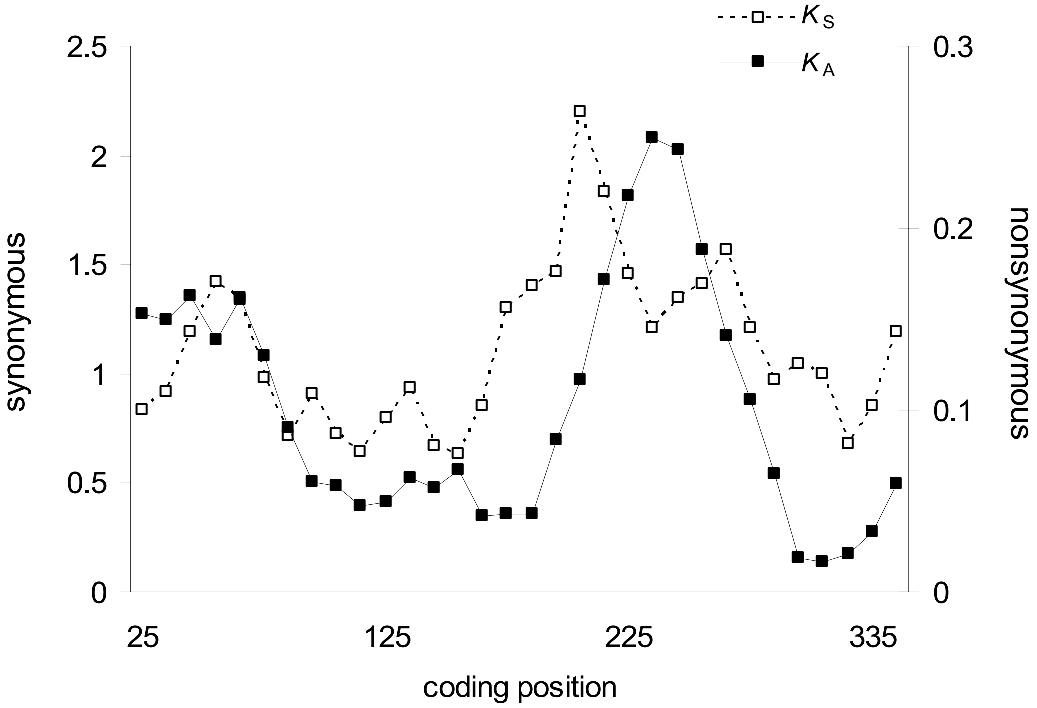


FIGURE 4



K_S
K_A
2.5
2
1.5
1
0.5
0
synonymous
0.3
0.2
0.1
0
nonsynonymous
25
125
225
335
coding position

FIGURE 5

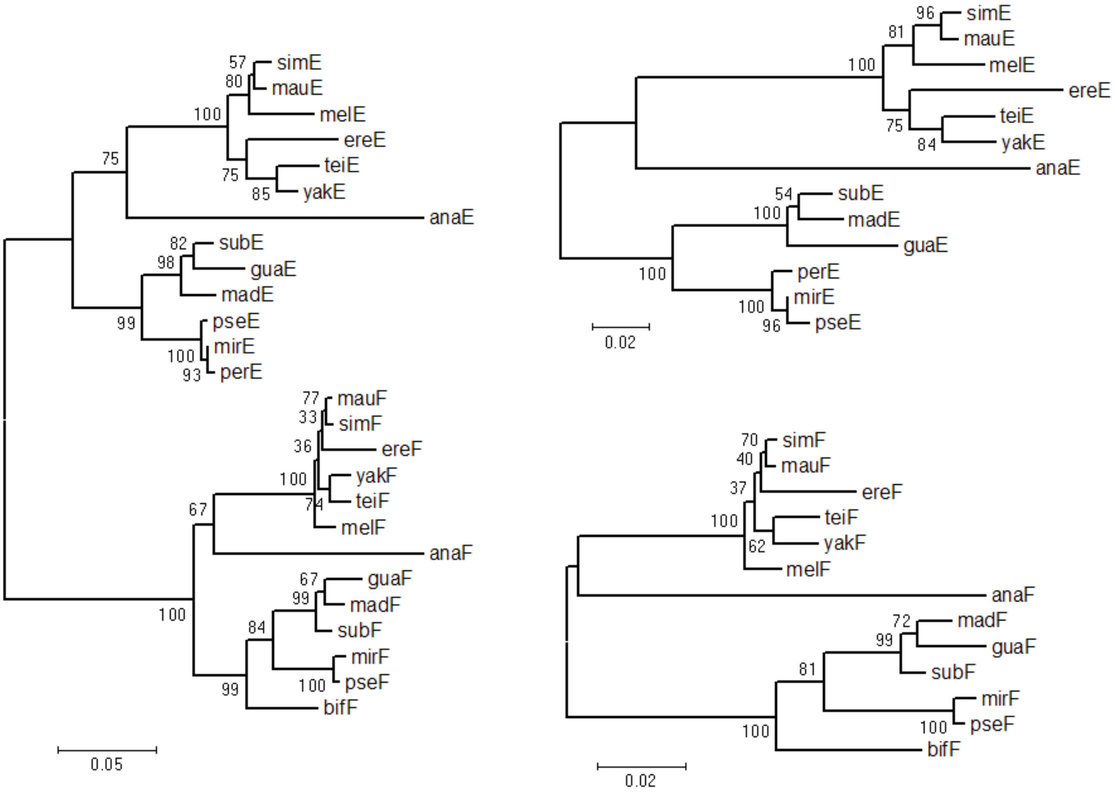

57
simE
80
mauE
100
melE
ereE
75
75
teiE
85
yakE
anaE
82
subE
98
guaE
madE
99
pseE
100
mirE
93
perE
77
mauF
33
simF
36
ereF
100
yakF
67
74
teiF
melF
anaF
67
guaF
99
madF
100
84
subF
mirF
99
100
pseF
bifF
0.05
96
simE
81
mauE
100
melE
ereE
75
teiE
84
yakE
anaE
54
subE
100
madE
guaE
100
perE
100
mirE
96
pseE
0.02
70
simF
40
mauF
37
ereF
100
teiF
62
yakF
melF
anaF
72
madF
99
guaF
81
subF
mirF
100
100
pseF
bifF
0.02

FIGURE 6

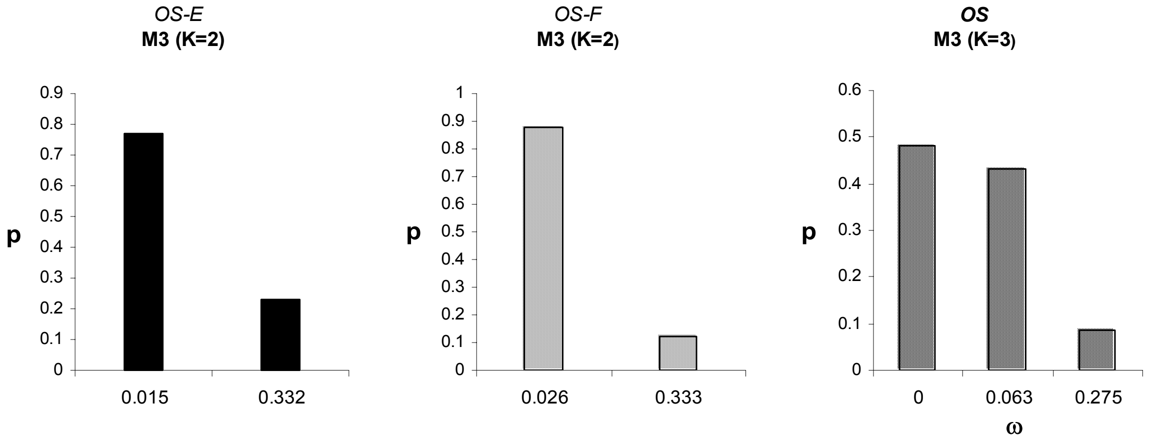

OS-E
M3 (K=2)
p
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.015
0.332
OS-F
M3 (K=2)
p
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.026
0.333
OS
M3 (K=3)
p
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.063
0.275
ω

FIGURE 7

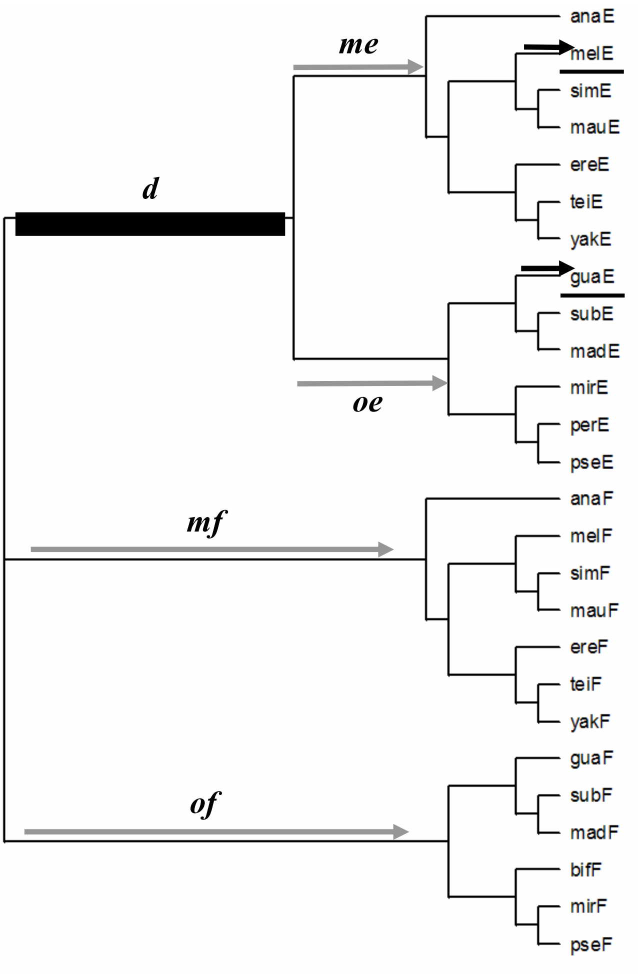

me
d
oe
mf
of
anaE
melE
simE
mauE
ereE
teiE
yakE
guaE
subE
madE
mirE
perE
pseE
anaF
melF
simF
mauF
ereF
teiF
yakF
guaF
subF
madF
bifF
mirF
pseF

FIGURE 8

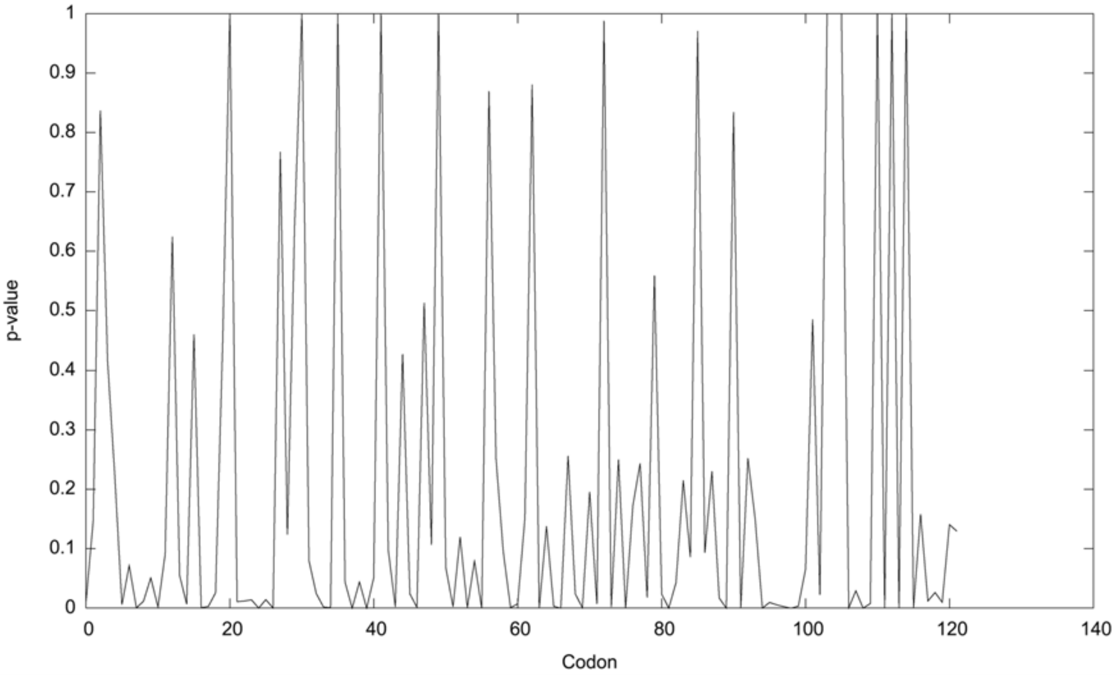


FIGURE 9

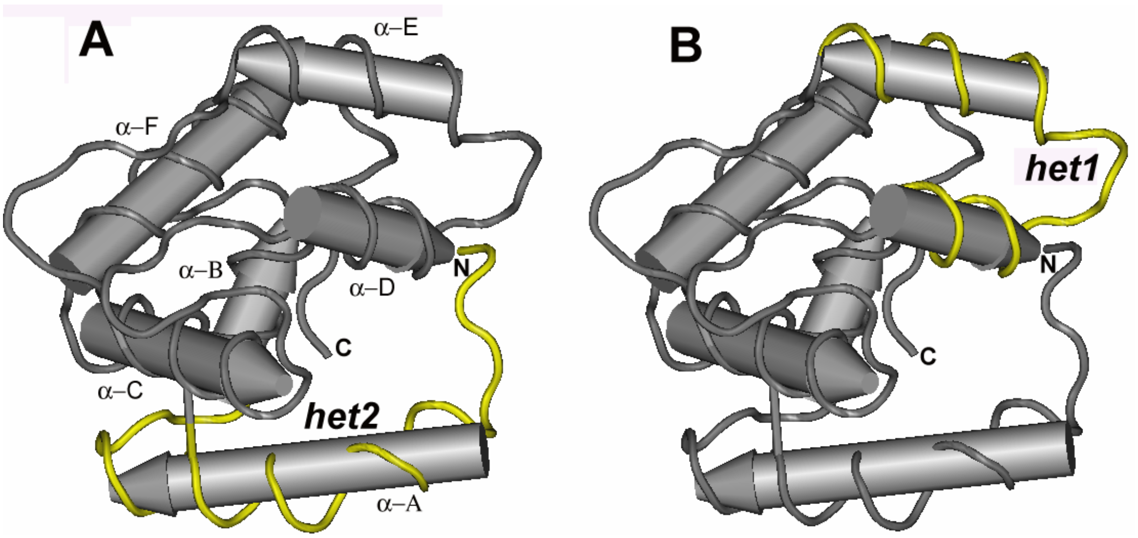


FIGURE 10

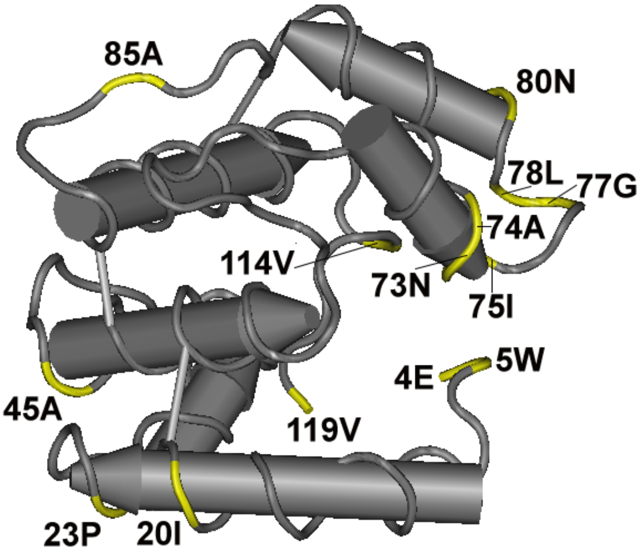


FIGURE 11

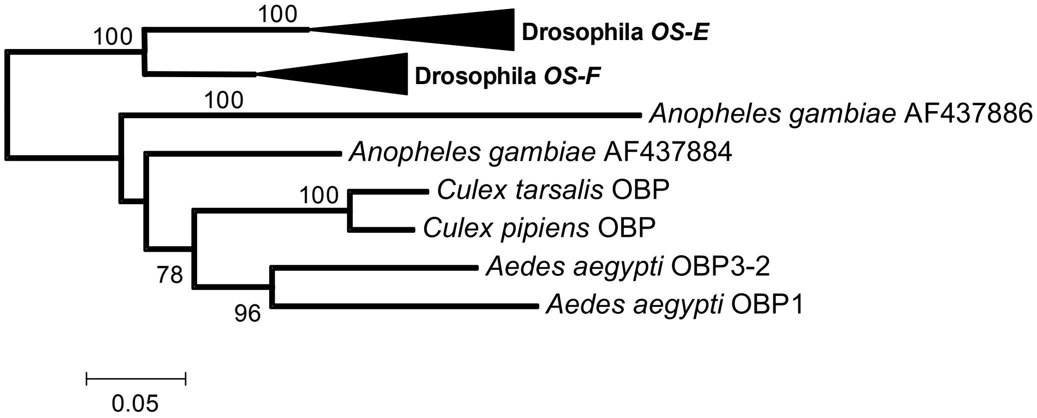

100
100
Drosophila *OS-E*
Drosophila *OS-F*
100
Anopheles gambiae AF437886
Anopheles gambiae AF437884
100
Culex tarsalis OBP
Culex pipiens OBP
78
Aedes aegypti OBP3-2
96
Aedes aegypti OBP1
0.05

FIGURE 12

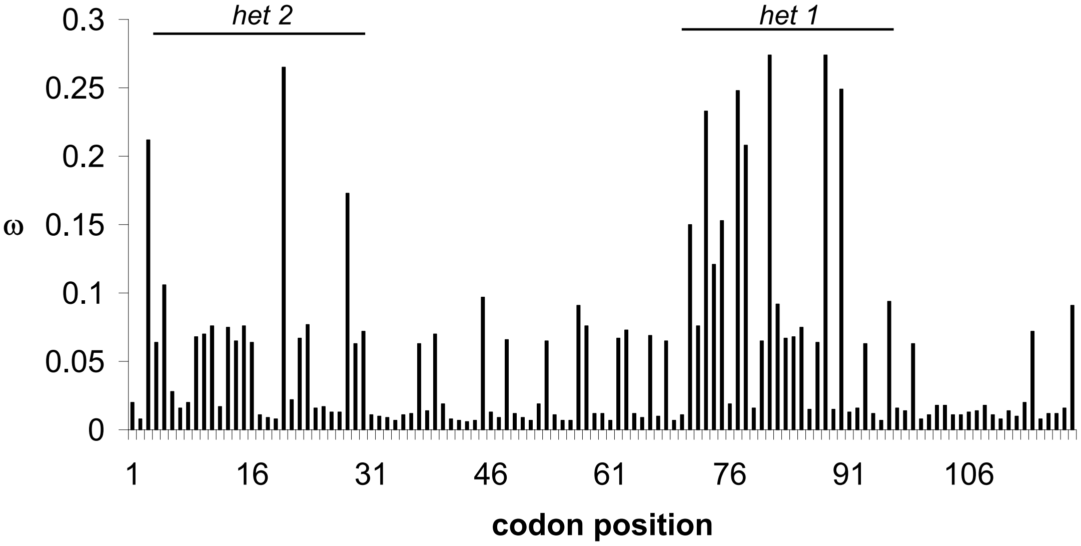
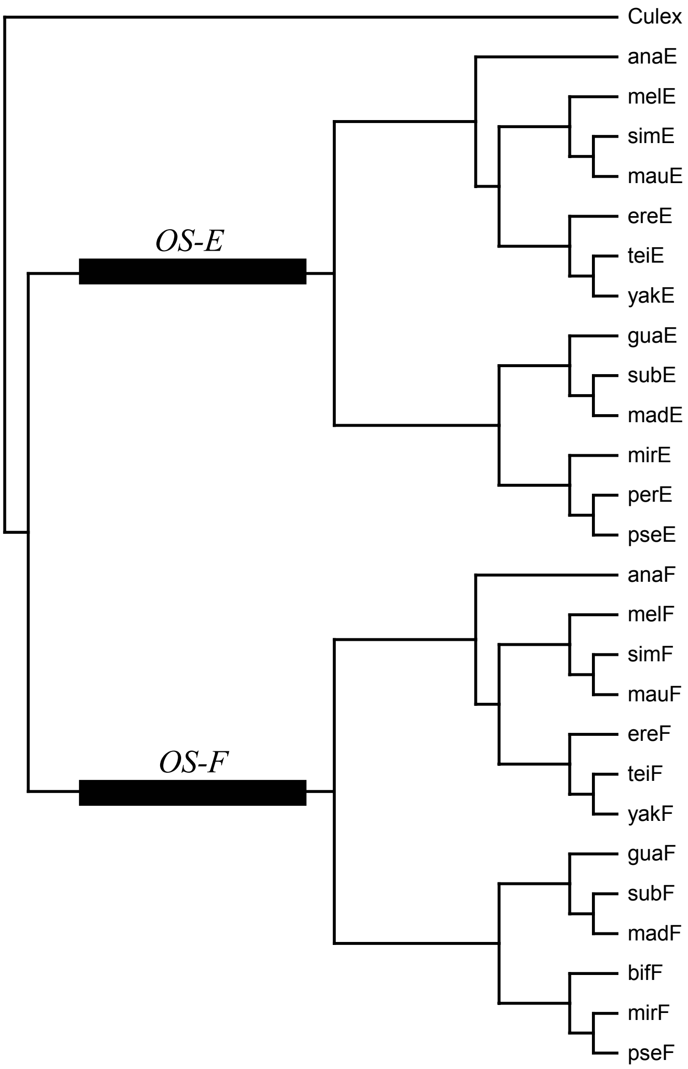

het 2
het 1
0.3
0.25
0.2
ω 0.15
0.1
0.05
0
1
16
31
46
61
76
91
106
codon position

FIGURE 13


Culex
anaE
melE
simE
mauE
ereE
teiE
yakE
guaE
subE
madE
mirE
perE
pseE
anaF
melF
simF
mauF
ereF
teiF
yakF
guaF
subF
madF
bifF
mirF
pseF
OS-E
OS-F

SUPPLEMENTARY MATERIAL

TABLE S1

Pairwise K_S

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
melE																									
simE	0.102																								
mauE	0.090	0.055																							
teiE	0.227	0.219	0.248																						
yakE	0.256	0.233	0.263	0.114																					
ereE	0.303	0.248	0.278	0.279	0.293																				
anaE	1.129	1.169	1.144	1.255	1.195	1.482																			
subE	0.790	0.785	0.800	0.834	0.589	1.052	0.894																		
guaE	0.766	0.822	0.806	0.873	0.617	1.062	0.901	0.065																	
madE	0.723	0.776	0.733	0.891	0.629	1.000	0.834	0.088	0.101																
mirE	0.665	0.714	0.637	0.954	0.815	1.119	0.860	0.258	0.230	0.237															
perE	0.691	0.742	0.662	0.993	0.847	1.074	0.894	0.272	0.244	0.252	0.010														
pseE	0.680	0.730	0.663	0.939	0.834	1.147	0.897	0.273	0.245	0.252	0.010	0.021													
melF	0.711	0.593	0.593	0.838	0.744	0.759	1.359	0.823	0.951	0.804	0.839	0.872	0.842												
simF	0.658	0.548	0.548	0.776	0.716	0.731	1.588	0.734	0.846	0.716	0.748	0.777	0.750	0.055											
mauF	0.684	0.570	0.570	0.776	0.716	0.731	1.504	0.707	0.814	0.689	0.720	0.748	0.722	0.055	0.022										
teiF	0.633	0.505	0.505	0.718	0.716	0.677	1.683	0.734	0.846	0.716	0.808	0.839	0.810	0.103	0.067	0.067									
yakF	0.697	0.563	0.563	0.796	0.735	0.694	1.634	0.763	0.862	0.689	0.808	0.839	0.810	0.103	0.079	0.079	0.055								
ereF	0.624	0.586	0.541	0.668	0.666	0.667	1.541	0.795	0.832	0.733	0.766	0.795	0.768	0.126	0.090	0.090	0.152	0.139							
anaF	1.436	1.321	1.433	1.317	1.608	1.628	1.065	0.906	0.914	0.990	1.000	1.042	1.004	0.679	0.679	0.679	0.679	0.580	0.723						
subF	0.913	0.822	0.806	0.947	0.907	0.856	1.095	0.717	0.708	0.699	0.821	0.853	0.823	0.548	0.548	0.504	0.570	0.570	0.552	0.707					
guaF	0.984	0.920	0.902	1.110	0.977	0.998	0.923	0.660	0.652	0.669	0.756	0.786	0.758	0.720	0.640	0.615	0.693	0.693	0.669	0.650	0.118				
madF	0.923	0.863	0.846	0.997	0.881	0.975	0.977	0.696	0.687	0.652	0.767	0.797	0.769	0.599	0.575	0.552	0.623	0.623	0.627	0.686	0.069	0.093			
bifF	0.845	0.823	0.781	0.951	0.840	0.782	1.302	0.755	0.703	0.689	0.684	0.712	0.686	0.423	0.423	0.386	0.463	0.433	0.428	0.683	0.271	0.303	0.273		
mirF	1.143	1.020	0.999	1.368	1.086	1.111	1.373	0.621	0.638	0.655	0.633	0.659	0.635	0.577	0.577	0.532	0.626	0.601	0.630	0.858	0.261	0.325	0.279	0.242	
pseF	1.090	0.975	0.956	1.297	1.037	1.061	1.242	0.619	0.636	0.653	0.632	0.657	0.633	0.576	0.576	0.530	0.624	0.600	0.628	0.838	0.245	0.308	0.262	0.242	0.023

Pairwise K_A

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25
melE																									
simE	0.023																								
mauE	0.023	0.000																							
teiE	0.025	0.008	0.008																						
yakE	0.029	0.012	0.012	0.004																					
ereE	0.037	0.019	0.019	0.012	0.015																				
anaE	0.057	0.065	0.063	0.067	0.063	0.078																			
subE	0.078	0.070	0.068	0.068	0.064	0.076	0.079																		
guaE	0.091	0.083	0.081	0.080	0.076	0.085	0.087	0.023																	
madE	0.085	0.076	0.074	0.074	0.070	0.087	0.083	0.012	0.031																
mirE	0.068	0.060	0.060	0.058	0.053	0.066	0.083	0.008	0.031	0.021															
perE	0.068	0.060	0.060	0.058	0.053	0.066	0.083	0.008	0.031	0.021	0.000														
pseE	0.074	0.066	0.064	0.064	0.060	0.072	0.079	0.004	0.027	0.017	0.004	0.004													
melF	0.198	0.177	0.177	0.187	0.187	0.194	0.211	0.154	0.170	0.158	0.138	0.138	0.142												
simF	0.203	0.181	0.181	0.192	0.192	0.199	0.216	0.159	0.175	0.163	0.143	0.143	0.147	0.004											
mauF	0.203	0.181	0.181	0.192	0.192	0.199	0.216	0.159	0.175	0.163	0.143	0.143	0.147	0.004	0.000										
teiF	0.208	0.186	0.186	0.197	0.197	0.204	0.221	0.164	0.180	0.167	0.147	0.147	0.152	0.008	0.004	0.004									
yakF	0.211	0.188	0.188	0.189	0.189	0.196	0.224	0.164	0.182	0.167	0.147	0.147	0.152	0.008	0.004	0.004	0.008								
ereF	0.214	0.187	0.187	0.195	0.195	0.205	0.227	0.162	0.181	0.168	0.148	0.148	0.152	0.015	0.012	0.012	0.008	0.015							
anaF	0.216	0.194	0.190	0.204	0.204	0.214	0.215	0.182	0.201	0.187	0.175	0.175	0.170	0.045	0.049	0.049	0.053	0.053	0.061						
subF	0.213	0.186	0.184	0.193	0.193	0.206	0.212	0.159	0.170	0.167	0.157	0.157	0.152	0.035	0.039	0.039	0.043	0.043	0.053	0.063					
guaF	0.213	0.191	0.189	0.194	0.194	0.206	0.213	0.159	0.171	0.167	0.157	0.157	0.152	0.039	0.043	0.043	0.047	0.047	0.057	0.067	0.004				
madF	0.213	0.191	0.188	0.198	0.198	0.210	0.217	0.163	0.175	0.172	0.161	0.161	0.156	0.039	0.043	0.043	0.047	0.047	0.057	0.067	0.004	0.008			
bifF	0.200	0.179	0.175	0.186	0.186	0.200	0.211	0.155	0.164	0.162	0.152	0.152	0.148	0.053	0.057	0.057	0.061	0.059	0.071	0.069	0.036	0.040	0.032		
mirF	0.212	0.186	0.183	0.193	0.193	0.205	0.209	0.154	0.165	0.162	0.151	0.151	0.147	0.039	0.043	0.043	0.047	0.047	0.057	0.061	0.011	0.015	0.015	0.025	
pseF	0.218	0.191	0.188	0.198	0.198	0.210	0.214	0.158	0.170	0.167	0.156	0.156	0.151	0.035	0.039	0.039	0.043	0.043	0.053	0.063	0.008	0.011	0.011	0.029	0.004

TABLE S2

Parameter estimates for the site-models

Data	Model	f	ℓ	Estimates of parameters	Positive selection
OS-E	M0	25	-1733.97	ω = 0.07	none
	M1	26	-1713.95	p_0 = 0.904, (p_1 = 0.096) ω_0 = 0.04	not allowed
	M2	28	-1713.95	p_0 = 0.904, p_1 = 0.096 ω_0 = 0.04 (p_2 = 0.000), ω_2 = 31.32	none
	M3 (K=2)	27	-1705.09	p_0 = 0.770, (p_1 = 0.230) ω_0 = 0.02, ω_1 = 0.33	none
	M3 (K=3)	29	-1703.68	p_0 = 0.709, p_1 = 0.278 (p_2 = 0.012), ω_0 = 0.01 ω_1 =0.24, ω_2 = 1.45	yes 8L*
	M7	26	-1704.41	p = 0.21, q = 1.97	not allowed
	M8	28	-1703.89	p = 0.23, q = 2.55 p_0 =0.992, (p_1= 0.008) ω_1 =1.66	yes 8L
OS-F	M0	25	-1398.27	ω = 0.06	none

	M1	26	-1394.30	$p_0 = 0.958$, ($p_1 = 0.042$) $\omega_0 = 0.04$	not allowed
	M2	28	-1394.30	$p_0 = 0.958$, $p_1 = 0.042$ $\omega_0 = 0.04$ ($p_2 = 0.000$), $\omega_2 = 30.00$	none
	M3 (K=2)	27	-1391.36	$p_0 = 0.878$, ($p_1 = 0.122$) $\omega_0 = 0.03$, $\omega_1 = 0.33$	none
	M3 (K=3)	29	-1391.36	$p_0 = 0.220$, $p_1 = 0.657$ ($p_2 = 0.123$), $\omega_0 = 0.03$ $\omega_1 =0.03$, $\omega_2 = 0.33$	none
	M7	26	-1391.69	$p = 0.29$, $q = 4.15$	not allowed
	M8	28	-1391.69	$p = 0.29$, $q = 4.15$ $p_0 =1.000$, ($p_1= 0.000$) $\omega_1 =1.00$	none
mature	M0	25	-2245.69	$\omega = 0.05$	none
	M1	26	-2236.72	$p_0 = 0.965$, ($p_1 = 0.035$) $\omega_0 = 0.04$	not allowed
	M2	28	-2236.72	$p_0 = 0.965$, $p_1 = 0.035$ $\omega_0 = 0.04$ ($p_2 = 0.000$), $\omega_2 = 8.47$	none

M3 (K=2)	27	-2212.41	$p_0 = 0.638$, ($p_1 = 0.362$) $\omega_0 = 0.01$, $\omega_1 = 0.13$	none
M3 (K=3)	29	-2207.77	$p_0 = 0.408$, $p_1 = 0.432$ ($p_2 = 0.088$), $\omega_0 = 0.00$ $\omega_1 =0.06$, $\omega_2 = 0.28$	none
M7	26	-2209.45	$p = 0.33$, $q = 5.73$	not allowed
M8	28	-2209.45	$p = 0.33$, $q = 5.74$ $p_0 =1.000$, ($p_1= 0.000$) $\omega_1 = 2.03$	none

f, number of free parameters

* *P* > *0*.95 (Bayesian analysis)

FIGURE LEGENDS

Figure S1. (A) Multiple alignment of the *OS-E* and *OS-F* coding regions corresponding to the mature protein portion (Al_OS data set). (B) Multiple alignment of the complete *OS-E* coding region (Al_OS-E data set). (C) Multiple alignment of the complete *OS-F* coding region (Al_OS-F data set). Dots indicate the same nucleotide that the reference sequence (melE).

FIGURE S1

A

```
[            10        20        30        40        50        60        70        80        90       100       110       120 ]
[             *         *         *         *         *         *         *         *         *         *         *         * ]
melE CGCGATGGAGAGTGGCCTCCGCCAGCGATTTTAAAACTGGGCAAGCACTTCCATGACATTTGTGCTCCCAAAACTGGCGTTACTGATGAGGCCATCAAGGAGTTCAGCGATGGGCAAATTCA
simE .......................C......C.........C.................................C.....G.........................................
mauE .......................G......C.........C.................................C.....G.........................................
teiE .................A.....C......C.C.......CA........T.......................C............................................A..
yakE .......C.........A.....C......C.C.......CA........T..C...........A........C.....A......................................A..
ereE ....................C..C....C.C.C..G....CA.................C..............C...........C................................A..
anaE ..A....CT........G........T.....G...A...CTC....T..T.....TC.C.....C........A..A..A..A..C..............A...........C........
subE ..G....CT...AT..A............C.G...T...C.C.A.....T..C..T........A...........T..C.......C...............T.....C...G....A..
guaE ..G....CT...AT..A........A...C.G...T...C.C.A.....T..C..T........A..T........T..C.......C...............T.....C...G....A..
madE ..G...TCT...AT..A............C.G...T...C.C.A.....T..C..T........A..T........T..C.......C...............T.....C..CG....A..
mirE ..G....CT...AT..A........A...C.....T...C.C.A........C..T........C..A........T..C.......C......................C..AG.G..A..
perE ..G....CT...AT..A........A...C.....T...C.C.A........C..T........C..A........T..C.......C......................C..AG.G..A..
pseE ..G....CT...AT..A........A...C.....T...C.C.A........C..T........C..A........T..C.......C......................C..AG.G..A..
melF ..T..C.AGA.C.AT..A.....G.GC..CC.G...A...C.....C......C...GCG....TGGAG..G..G.....A..C..G.CT........................G.G.....
simF ..T..C.AGA.C.AT..A.....G.GC..CC.G...A...C.....C......C...GCG....TGGAG..G..G.....CT.C..G.CT........................G.G.....
mauF ..T..C.AGA.C.AT..A.....G.GC..CC.G...A...C.....C......C...GCG....TGGAG..G..G.....CT.C..G.CT........................G.G.....
teiF ..T..C.AGA.C.AT..A.....G.GC..CC.G...A...C.....C......C...GCG....TGGAG..G..C.....CT.C..G.CT..............T.........G.......
yakF ..T..C.AGA.C.AT..A.....G.GC..CC.G...A...C.....C......C...GCG....TGGAG..G..C.....CT.C..G.CT..............T........CG.G.....
ereF ..T..C.AGA.C.AT..A.....G.GC..CC.C...A...C.....C......C...GCG....TGGAG..G..G.....CT.C..G.CT........................G.G..C..
anaF ..T..C.AGA.C.AT..A.....G.GC..CC.G...A...C...A.C..........TCG....TGGAG.....A........C..G...........A..A..T.....A..AG.G..C..
subF ..T..C.AGA.C.AT..A.....G.GC..AA.G...A...CG..A.C......C..TTCC....TGGAG.....G.....C..C..G.C.....................C...G.C..A..
guaF ..T..C.AGA.C.AT..A.....G.GC...A.G...A...CG..A.C......C..TTCC....TGGAG.....G.....C..C..G.C.....................C...G.C..C..
madF ..T..C.AGA.C.AT..A.....G.GC..AA.G...A...CG..A.C......C..TTCC....TGGAG.....G.....C..C..G.C.....................C...G.C..A..
bifF ..T..C.AGA.C.AT..A.....G.GC..AA.G...A...C.....C......C..T.CC....TGGAG..G..G.....C.GC..G.C.........................G.G..A..
mirF ..T..C.AGA.C.AT..A.....G.GC..AA.G...A...C.....C......C..TTCC....TGGAG.....G.....C.GC..G.C......T...............C...G.G..A..
pseF ..T..C.AGA.C.AT..A.....G.GC..AA.G...A...C.....C......C..TTCC....TGGAG.....G.....C.GC..G.C.....................C...G.G..A..
```

```
[         130       140       150       160       170       180       190       200       210       220       230       240    ]
[           *         *         *         *         *         *         *         *         *         *         *         *    ]
melE TGAGGACGAGGCCCTCAAGTGCTATATGAACTGCCTCTTCCACGAGTTCGAGGTGGTCGACGACAATGGGGATGTCCACATGGAGAAGGTCTTGAACGCCATTCCG---GGAGAAAAGCTGA
simE ..............................................A....A....................................C....C............---.............
mauE ..............................................A....A....................................C....C............---.............
teiE C..............T.............................A.......................T..................C....T............---...........A.
yakE C...................................G................................T..................C....C............---.....G.....A.
ereE C..............A.....G.............................A....................................C....C............---.....G.......
anaE C........A..T.....A..............T.....T.....A.....T..C..T..T..A..C..C..................C.G.....T..T.....C---..T........CC
subE C.........AAT...........C........T..G..............T.....G..T..T.....C.....G..T........AC.A..C..T...G....A---TCG......T.AC
guaE C.........AAT...........C........T..G.....T........T........T..T.....C.....G...........AC.A..C..TT.......A---TCG....GCT.TC
madE C.........AAT...........C........T..G..............T.....G..T..T.....C...A.G...........AC.A..C..T...G.....---TCG.........C
mirE C.........AAT...........C........T.................A........T..T.....C.....A..T.........C.A..C..T...G....C---TCG.........C
perE C.........AAT...........C........T.................A........T..T.....C.....A..T.........C.A..C..T...G....C---TCG.........C
pseE C.........AAT...........C........T.................A........T..T.....C.....A..T.........C.A..C..T...G....C---TCG.........C
melF C.........AAG.....A.....C.........T...........A....A.....G..............C..G..TC........C....CGC.A.GG.A...CTCTCC------A..C
simF C.........AAG...........C.........T...........A....A.....G..............C..G..TC........C....CGC.A.GG.C...CTCTCC------A..C
mauF C.........AAG...........C.........T...........A....A.....G...........C..C..G..TC........C....CGC.A.GG.C...CTCTCC------A..C
teiF C.........AAG...........C.........T...........A....A.....G........A.....C..G..TC........C....CGC.A.GG.C...CTCTCC------A..C
yakF C.........AAG...........C.........T...........A....A.....G..............C..G..TC........C....CGC.A.AG.G...CTTTCC------A..C
ereF C.........AAG...........C.........T..........AA....A.....G........G.....C..G...C........C....CGC.A..G.C...CTCTCC------T..C
anaF C.....T..AAAG...........C........TT....T.....AA.T..A.....G..T..A.....A..C..T..TC.......AC....TG.GA.AG.G..TCGCATG------A..C
subF C.........AAG...........C.....T...T...........A....A.....T.....T.....C.....T...T.......AC....TGC.A..G....ACTCTCC------A.TC
guaF C.........AAG...........C.....T..TT....T......G....A.....T.....T.....C.....T...T.......AC....TGC.A.TG....ACTCTCC------A.TC
madF C.........AAG...........C.....T...T...........A....A.....T.....T.....C.....T...T.......AT....TGC.A.TG.....CTCTCC------A.TC
bifF C.........AAG...........C.....T..TT...........A....A.................C.....A...T.......AT....TG.......G...CTTTCC------A.TC
mirF C.........AAA...........C.....T..TT...........A....A...........T.....C.....A..TT....A..AC.T..CG..A..G....ACTCTCC------A.TC
pseF C.........AAA...........C.....T..TT...........A....A...........T.....C.....A..TT....A..AC.T..CGC.A..G....ACTCTCC------A.TC
```

```
[      250       260       270       280       290       300       310       320       330       340       350       360        ]
[         *         *         *         *         *         *         *         *         *         *         *         *       ]
melE  GGAACATTATGATGGAGGCTTCCAAGGGATGCATTCATCCTGAGGGCGACACCCTGTGCCACAAAGCCTGGTGGTTCCACCAATGCTGGAAGAAGGCTGATCCTGTCCACTACTTTTTGGTC
simE  ........C................A........G..............T..G...........G.........................................................
mauE  ........C...................T.....G.................A...........G.........................................................
teiE  ........C................A.......CG........A........T...........G................................A.......................G
yakE  ........T................A.......CG........A........T...........G.................G..............A........................
ereE  .....C..C................A.......CG.....C...........T...........G...........T.....G..T...........C.....C..G...........A...
anaE  .T..T..G.....CA.T..C..G....AT..T.....C..G........T..TT..........G..............T..G........A..............T..T..T..C......
subE  .C...G..T..........C.....A.AT...........G...........GT..........G..............T..G..............C..............T..C......
guaE  .C..GC..T..........C.....A.AT...........GA...........T..........G..............T..G..............C..............T..C......
madE  .C......T..........C.......AT...........G...........GT..........G..............T..G..............C.....G........T..C......
mirE  .C...G.AT..........C........T.....C.....G............T..........G..............T................................T..C......
perE  .C...G.AT..........C........T.....C.....G............T..........G..............T.....T..........................T..C......
pseE  .C...G.AT..........C.......AT.....C.....G............T..........G..............T................................T..C.....A
melF  .CG...AGC........ATG........C...G.C.....G........T..G...........G.................G........A.....C.....CAAG........C...CCG
simF  .CG...AGC........ATG........C...G.C.....G........T..G...........G.................G..............C.....AAAG........C...CCG
mauF  .CG...AGC........ATG........C...G.C.....G........T..G...........G.................G..............C.....CAAG........C...CCG
teiF  .CG...AGC........ATG........C...G.C.....C........T..G...........G................................C.....CAAG........C...CCG
yakF  .CG...AGC........ATG........C...GCC.....C........T..G...........G.................G..............C.....CAAG........C...CCG
ereF  .CG...AGC..G.....ATG........C...G.C.....G...........G...........G.................G..............G.....CAAG........C...CCG
anaF  .AG.T.AAC.......AATG..GC.AAAC..TG.......C..A.....T...T..........G..............T..G........A.....C......AAG........C...CCT
subF  ..G.T.AGC.......AATG.......AT..TG.G.....C.....A..TT.G........G..G..............T..G........A.....C.....CAAG..T...........T
guaF  ..G.T.AGT.......AATG.......AT..TG.G.....C.....A..TT.GT.......G..G...........T..T..G..............C......AAG..T...........T
madF  ..G.T.AGT.......AATG.......AT..TG.G.....C.....T..TT.GT.......G..G..............T..G........A.....C......AAG..T...........T
bifF  .CG...AAT.......AATG......AAC...G.G.....C.....A..T..AT.......G..G...........T..T..G........A.....C.....CAAG........C.....T
mirF  .CG.T.AGT.......AATG.......AT...G.G.....C.....G..TT.GT.A..T..G..G..............T..G........A.....C.....CAAG........C.....T
pseF  .TG.T.AGT.......AATG.......AT...G.G.....C.....G..TT.GT.A..T..G..G..............T..G........A.....C.....CAAG........C.....T
```

B

```
[            10        20        30        40        50        60        70        80        90       100       110]
[             *         *         *         *         *         *         *         *         *         *         *]
melE ATGGTCAAATAC---CCACTGATACTACTTTTGATTGGCTGTGCCGCTGCCCAGGAACCAAGGCGCGATGGAGAGTGGCCTCCGCCAGCGATTTTAAAACTGGGCAAGCA
simE ............---............T..C.......................................................C......C.........C......
mauE ............---............T..C.......................................................G......C.........C......
teiE ............CTGG.G...G.....T..A..C..............................................A.....C......C.C.......CA.....
yakE ...T........CTGG...........T..A..C....................................C.........A.....C......C.C.......CA.....
ereE ....C.......CTG............T..AC.C..........A.................A....................C..C....C.C.C..G....CA.....
anaE ...A.T..GC..CTGG.TG..C.GA..T.....C....T.C...G..........CT..CC.T..A...CT.........G........T.....G...A...CTC....
subE ...A.G....TTGCTG.T...C.CACC..GC..C.C..T.T..GG..C..T........CC....G...CT.....AT..A............C.G...T...C.C.A..
guaE ...A.G....TTGTGG.T...C.CGCC..GC..C.C..T..C.TG..C..T.....G..CC.C..G...CT.....AT..A........A...C.G...T...C.C.A..
madE ...A.G....CTGTTG.T..TC.CA.G..GC.AC.........TA..........G...C..A..G...CT.....AT..A........A...C.....T...C.C.A..
pseE ...A.G....CTGTTGGT..TC.CA.G..GC.AC.........TA..........G...C..A..G...CT.....AT..A........A...C.....T...C.C.A..
perE ...A.G....CTGTTG.T..CC.C.CGT.GC.AC.........TA..........G...C..A..G...CT.....AT..A........A...C.....T...C.C.A..

[           120       130       140       150       160       170       180       190       200       210       220]
[             *         *         *         *         *         *         *         *         *         *         *]
melE CTTCCATGACATTTGTGCTCCCAAAACTGGCGTTACTGATGAGGCCATCAAGGAGTTCAGCGATGGGCAAATTCATGAGGACGAGGCCCTCAAGTGCTATATGAACTGCC
simE ...........................C.....G............................................................................
maue ...........................C.....G............................................................................
teiE ...T.......................C............................................A..C..............T...................
yakE ...T..C...........A........C.....A......................................A..C..................................
ereE ............C..............C...........C................................A..C..............A.....G.............
anaE T..T.....TC.C.....C........A..A..A..A..C..............A...........C........C........A..T.....A..............T.
subE ...T..C..T........A...........T..C.......C...............T.....C...G....A..C.........AAT...........C........T.
guaE ...T..C..T........A..T........T..C.......C...............T.....C...G....A..C.........AAT...........C........T.
madE ...T..C..T........A..T........T..C.......C...............T.....C..CG....A..C.........AAT...........C........T.
mirE ......C..T........C..A........T..C.......C.....................C..AG.G..A..C.........AAT...........C........T.
pseE ......C..T........C..A........T..C.......C.....................C..AG.G..A..C.........AAT...........C........T.
perE ......C..T........C..A........T..C.......C.....................C..AG.G..A..C.........AAT...........C........T.
```

```
[           230       240       250       260       270       280       290       300       310       320       330]
[             *         *         *         *         *         *         *         *         *         *         *]
melE TCTTCCACGAGTTCGAGGTGGTCGACGACAATGGGGATGTCCACATGGAGAAGGTCTTGAACGCCATTCCGGGAGAAAAGCTGAGGAACATTATGATGGAGGCTTCCAAG
simE ...........A....A....................................C....C.................................C................A
mauE ...........A....A....................................C....C.................................C.................
teiE ..........A.......................T..................C....T.......................A.........C................A
yakE .G................................T..................C....C.................G.....A.........T................A
ereE ................A....................................C....C.................G............C..C................A
anaE ....T.....A.....T..C..T..T..A..C..C..................C.G.....T..T.....C..T........CC.T..T..G.....CA.T..C..G...
subE .G..............T.....G..T..T.....C.....G..T........AC.A..C..T...G....ATCG......T.AC.C...G..T..........C.....A
guaE .G.....T........T........T..T.....C.....G...........AC.A..C..TT.......ATCG....GCT.TC.C..GC..T..........C.....A
madE .G..............T.....G..T..T.....C...A.G...........AC.A..C..T...G....TCG.........C.C......T..........C......
mirE ................A........T..T.....C.....A..T.........C.A..C..T...G....CTCG.........C.C...G.AT..........C......
pseE ................A........T..T.....C.....A..T.........C.A..C..T...G....CTCG.........C.C...G.AT..........C......
perE ................A........T..T.....C.....A..T.........C.A..C..T...G....CTCG.........C.C...G.AT..........C......

[          340       350       360       370       380       390       400       410       420      ]
[            *         *         *         *         *         *         *         *         *      ]
melE GGATGCATTCATCCTGAGGGCGACACCCTGTGCCACAAAGCCTGGTGGTTCCACCAATGCTGGAAGAAGGCTGATCCTGTCCACTACTTTTTGGTC
simE ........G..............T..G...........G.........................................................
mauE ..T.....G.................A...........G.........................................................
teiE .......CG........A........T...........G................................A.......................G
yakE .......CG........A........T...........G.................G..............A........................
ereE .......CG.....C...........T...........G...........T.....G..T...........C.....C..G...........A...
anaE .AT..T.....C..G........T..TT..........G..............T..G........A..............T..T..T..C......
subE .AT...........G...........GT..........G..............T..G..............C..............T..C......
guaE .AT...........GA...........T..........G..............T..G..............C..............T..C......
madE .AT...........G...........GT..........G..............T..G..............C.....G........T..C......
mirE ..T.....C.....G............T..........G..............T................................T..C......
pseE .AT.....C.....G............T..........G..............T................................T..C.....A
perE ..T.....C.....G............T..........G..............T.....T..........................T..C......
```

C

```
[            10        20        30        40        50        60        70        80        90       100       110]
[             *         *         *         *         *         *         *         *         *         *         *]
melF ATGGCTTTGAATGGCTTTGGTCGG---------------------------------------------------CGTGTCAGTGCGTCTGTCCTTTTAATCGCCTTGTC
simF ........................---------------------------------------------------...................................
mauF ........................---------------------------------------------------...................................
ereF ........................---------------------------------------------------...................................
teiF ........................---------------------------------------------------...........................A.......
yakF ........................---------------------------------------------------...................................
anaF ........................---------------------------------------------------------------.......................
guaF ........................CAG------------------CAGCTTGGAC------GCCGCCTTCGTCGTAC.........AG.C.....C..............
madF ........................CAGGAGCAGCAGGCGGAGCAGCAGCTTGGAC------GCCGCCTTCGTCGT.C.........AG.C.....C..............
subF ........................CAGCAGGAGGAGCAGCAGC------TTGGAC------GCCGCCTTCGTCGT.C.........AG.C.....C..............
bifF ............A...........CGGCAGC------------------TTGGAC------ACCGCCTTCGCCGCGC.........AG.C.....C..............
pseF .......................ACAGCAGCAGCAGCAGCAGCAGC---TTGGACTTGGACGCCGCCTTCGTCGT.C.........AG.C.....C..............
mirF .......................ACAGCAGCACCAGCGGCAGCAGC---TTGGACTTGGACGCCGCCTTCGTCGT.C.........AG.C.....C..............

[           120       130       140       150       160       170       180       190       200       210       220]
[             *         *         *         *         *         *         *         *         *         *         *]
melF GCTGCTCAGCGGAGCGCTGATCCTGCCGCCGGCTGCGGCGCAGCGTGACGAGAACTATCCACCGCCGGGCATCCTGAAAATGGCCAAGCCCTTCCACGACGCGTGTGTGG
simF ..............................................................................................................
mauF ..............................................................................................................
ereF ............................................................................C.................................
teiF ..............................................................................................................
yakF ..............................................................................................................
anaF ....................................TC..................................................A........T...T........
guaF ............AG........G.....A............................................TA..........G..A...........TT.C......
madF ............AG........G.....A............................................AA..........G..A...........TT.C......
subF ............AG........G.....A............................................AA..........G..A...........TT.C......
bifF ............AG........G.....A............................................AA.........................TA.C......
pseF ............AGC.......G.....A............................................AA.........................TT.C......
mirF ............AGC.......G.....A............................................AA.........................TT.C......
```

```
[          230       240       250       260       270       280       290       300       310       320       330]
[            *         *         *         *         *         *         *         *         *         *         *]
melF AGAAGACGGGCGTAACCGAGGCTGCCATCAAGGAGTTCAGCGATGGGGAGATTCACGAGGACGAGAAGCTCAAATGCTACATGAACTGCTTCTTCCACGAGATCGAAGTG
simF .............CT..........................................................G....................................
mauF .............CT..........................................................G....................................
ereF .............CT.....................................C....................G..........................A.........
teiF .......C.....CT......................T...........A.......................G....................................
yakF .......C.....CT......................T........C..........................G....................................
anaF ....A..A.....T.......AG........A..A..T.....A..A.....C........T..A........G..............T.....T.....A..T......
guaF ....A........C........G....................C.....C..C....................G...........T..T.....T......G........
madF ....A........C........G....................C.....C..A....................G...........T........................
subF ....A........C........G....................C.....C..A....................G...........T........................
bifF .............C.G......G.............................A....................G...........T..T.....................
pseF ....A........C.G......G....................C........A..............A.....G...........T..T.....................
mirF ....A........C.G......G.....T..............C........A..............A.....G...........T..T.....................

[         340       350       360       370       380       390       400       410       420       430       440]
[           *         *         *         *         *         *         *         *         *         *         *]
melF GTGGACGACAATGGGGACGTGCATCTGGAGAAGCTCTTCGCCACGGTACCGCTCTCCATGCGCGACAAGCTGATGGAGATGTCCAAGGGCTGCGTCCATCCGGAGGGCGA
simF ...............................................C..............................................................
mauF ..............C................................C..............................................................
ereF ...........G...........C....................C..C.........T..............G.....................................
teiF ...........A...................................C.....................................................C........
yakF ............................................A..G.....T........................................C......C........
anaF .....T..A.....A.....T...........A.....T.AG..A..G..T.G.ATG.....A..T..A........A.....GC.AAA...T..T.....C..A.....
guaF ..T.....T.....C..T..T..CT.......A.....T.....T..T..A........T..G..T...T.......A..........AT..T..G.....C.....A..
madF ..T.....T.....C..T..T..CT.......AT....T.....T..T...........T..G..T...T.......A..........AT..T..G.....C.....T..
subF ..T.....T.....C..T..T..CT.......A.....T.....C..T..A........T..G..T...........A..........AT..T..G.....C.....A..
bifF ..C...........C..T..A..CT.......AT....T.A.G.CA.G.....T.....T........AT.......A.........AA......G.....C.....A..
pseF ..C.....T.....C..T..A...T....A..A..T........C..T..A........T..T..T...T.......A..........AT.....G.....C.....G..
mirF ..C.....T.....C..T..A...T....A..A..T....A...C..T..A........T.....T...T.......A..........AT.....G.....C.....G..
```

```
[              450       460       470       480       490       500       510   ]
[                *         *         *         *         *         *         *   ]
melF TACGCTGTGCCACAAGGCCTGGTGGTTCCACCAGTGCTGGAAAAAGGCCGATCCCAAGCACTACTTCTTGCCG
simF ..........................................G...........A..................
mauF ..........................................G..............................
ereF C.........................................G.....G........................
teiF .................................A........G..............................
yakF ..........................................G..............................
anaF ...CT.........................T.......................T.................T
guaF .T..T.......G..............T..T...........G...........T.....T.....T...GTT
madF. T..T.......G.................T.......................T.....T.....T...GTT
subF .T..........G.................T.............................T.....T...GTT
bifF ...AT.......G..............T..T.......................................GTT
pseF .T..T.A..T..G.................T.......................................GTT
mirF .T..T.A..T..G.................T.......................................GTT
```