Nucleotide and Copy-Number Polymorphism at the Odorant Receptor Genes Or22a and Or22b in Drosophila melanogaster

Montserrat Aguadé

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

In Drosophila, odorant receptors are encoded by an old and moderately sized multigene family. Or22a and Or22b are two tandemly arranged genes of this family that have proved to be the result of a rather young duplication. Nucleotide variation in the region spanning both duplicates was surveyed in four natural populations (two African and two non-African) of Drosophila melanogaster and also analyzed in species of the melanogaster subgroup. The intraspecific survey revealed a particular copy-number polymorphism in some of the studied populations, with the two genes (Or22a and Or22b) present in the long variant and a single chimeric gene (Or22ab) present in the short variant. Estimated nucleotide diversity was higher in the short than in the long variant, despite the ancestral character of the latter variant in D. melanogaster. The general skew toward low-frequency variants detected in the non-African long variant and its reduced level of silent polymorphism relative to divergence is consistent with the recent fixation of an advantageous mutation at, or nearby, the Or22 long variant region. The nonnegligible frequency of the short variant and the presence of a highly divergent haplotype in the East African sample would point to direct or indirect selection for its maintenance in the species. There was evidence for a generally more rapid evolution of the Or22b copy at both synonymous and nonsynonymous sites. However, an excess of nonsynonymous substitutions was only detected in the early history of this copy.

Introduction

In Drosophila, odorant receptors are encoded by a moderately sized multigene family (Robertson et al. 2003). Its members are distributed across the genome, which together with the relatively high average divergence between copies points to most of them being the result of old duplications. Indeed, over two-thirds of the Or genes present in Drosophila melanogaster are also present in the additional 11 Drosophila species with available whole-genome sequences (Drosophila 12 Genomes Consortium et al. 2007; Guo and Kim 2007), indicating that at least these members of the family predate the split of the Drosophila and Sophophora subgenera (i.e., they are over 40–60 My old).

In D. melanogaster, expression of Or genes is restricted to the two chemoreceptor organs of adult flies (i.e., antenna and maxillary palp) and the dorsal organ of larvae. In this species, the odorant profile of most Or genes has been characterized in adults and also in larvae (Hallem et al. 2004; Fishilevich et al. 2005; Hallem and Carlson 2006). The profiles obtained using pure odors at different concentrations are rather distinct among receptors. The differential profiles and the expression of different subsets of genes in different organs and life stages would support their functional divergence (Dobritsa et al. 2003; Fishilevich et al. 2005; Couto et al. 2005). Electrophysiological characterization showed, however, that particular odors could elicit a response from different ORs, which would suggest partial redundancy among receptors at least in laboratory conditions. It would seem therefore that both neofunctionalization and subfunctionalization have been important in the evolution of this gene family.

The recent availability of whole-genome sequences of 12 Drosophila species has fostered the evolutionary analysis of the Or multigene family (Drosophila 12 Genomes Consortium et al. 2007; Guo and Kim 2007; McBride et al. 2007; Nozawa and Nei 2007; Tunstall et al. 2007; Conceição and Aguadé 2008; Gardiner et al. 2008). These studies have focused generally on the family mode of evolution and also on the evolutionary rates of its members based on divergence estimates between species. Because most duplicates originated prior to the split of the Drosophila genus, this approach might in general fail to detect the forces contributing to their successful fixation and diversification given that any signal might have been obscured by the subsequent accumulation of mutations throughout the independent history of each duplicate (Lynch and Katju 2004).

Here, nucleotide variation was surveyed in natural populations of D. melanogaster in the region spanning two tandemly arranged and rather closely related Or duplicates, the Or22a and Or22b genes (Robertson et al. 2003), which have proved to be rather young duplicates (Drosophila 12 Genomes Consortium et al. 2007; Guo and Kim 2007; McBride et al. 2007; Tunstall et al. 2007; present study). Their initial functional characterization in D. melanogaster suggested that only the Or22a gene was functional in this species (Dobritsa et al. 2003; see however Couto et al. 2005; Ray et al. 2007). Expression-level studies would indicate that they are both expressed in the same neuron (ab3A) of basiconic type 3 sensilla (Hallem et al. 2004; Couto et al. 2005). The study of polymorphism in D. melanogaster and divergence relative to the closely related species Drosophila simulans might shed light on the possible action of adaptive selection in the recent history of these genes and also on the role of gene conversion between these rather young duplicates. Moreover, comparison with an outgroup predating the duplication might contribute to characterize the initial divergence between duplicates.

Materials and Methods

Drosophila Strains

A total of 63 strains randomly sampled from five natural populations of D. melanogaster were used for sequencing: 12 from Sant Sadurní d’Anoia, Catalonia (Spain), 14...
from North Carolina (United States), 14 from Lamto (Ivory Coast; West Africa), and 11 and 12 from two different collections in Malawi (Malawi-1 and Malawi-2; East Africa). For the Sant Sadurní d’Anoia sample, isofemale lines were established upon collection in September 2001 (Orengo and Aguadé 2004); isochromosomal lines for the second chromosome were extracted (through a series of crosses with the balancer stock y w; sco/SM6b) and kindly provided by E. Virágh and I. Kiss. Isochromosomal lines from North Carolina as well as those from Lamto and Malawi-1 constitute a subset of those previously studied for other regions (Aguadé et al. 1992; Bénassi et al. 1993; Bénassi and Veuille 1995). Lines from Malawi-2 were kindly provided by Charles H. Langley and David Begun as heterozygotes over the balancer CyO chromosome. For these lines, the Or22 wild locus was differentially amplified by the polymerase chain reaction (PCR) from individuals heterozygous over the Df(2L)frtz14 deficiency (Collier et al. 2004; results not shown). The presence or absence of the In(2L)it inversion was checked in most of the D. melanogaster strains used either cytologically (Aguadé et al. 1992; Bénassi et al. 1993; Bénassi and Veuille 1995) or molecularly (as in Andolfatto et al. [1999]). One highly inbred line of D. simulans from Montblanc, Catalonia, Spain (Aguadé 1999) and one highly inbred line of Drosophila yakuba kindly provided by Carmen Segarra were also used for sequencing. An additional set of 66 D. melanogaster lines was used to obtain a more reliable estimate of the frequency of the two Or22 region length variants (see Results).

DNA Extraction, PCR Amplification, and Sequencing

Genomic DNA was extracted by a modification of protocol 48 in Ashburner (1989) from 10 adult flies in the case of the D. melanogaster lines and from one single individual in the case of the highly inbred D. simulans and D. yakuba lines. Oligonucleotides for PCR amplification and sequencing were designed on the published sequence of the Or22 region (http://flybase.org) using the Oligo version 4.1 program (Rychlik 1992). Special attention was paid to design copy-specific oligonucleotides for sequencing. For each of the lines, the Or22 region was PCR amplified using one primer (o2F) designed on the 5’ flanking region of the Or22a gene and a second one (o2R) on the 3’ flanking region of the Or22b gene (fig. 1). Whenever the size of the amplified fragment revealed the presence of both genes, two fragments were generally amplified for sequencing, each spanning one gene and its flanking regions (o2F–o4R and o4F–o2R). A third fragment spanning the intergenic region and its flanking coding regions (o3F–o3R) was also obtained and completely sequenced in some strains. The PCR amplification products were purified either by treatment with Exo-SapIT (Amersham Biosciences, Piscataway, NJ) or as described in Dean et al. (2003). Purified products were used directly as templates for sequencing with the ABI PRISM version 3.1, or 3.2, kit (Applied Biosystems, Foster City, CA) according to the manufacturer’s conditions. After ethanol precipitation, sequencing reactions were separated on an ABI PRISM 3730 sequencer (Applied Biosystems). Chromatograms were in all cases visually inspected, and all polymorphic sites were checked both in each line and across lines. Sequences were obtained on both strands. The sequences newly reported in this article have been deposited in the EMBL sequence database library under GenBank accession numbers FM212142–FM212206.

Sequence Analysis

Sequences were assembled and multiply aligned using the DNAStar (Madison, WI) software package, and the multiple alignments were later edited with the MacClade version 3.06 program (Maddison WP and Maddison DR 2002). The DnaSP version 4.10.1 program (Rozas et al. 2003) was used for most intraspecific and some interspecific analyses. Genetic differentiation between copies was measured as either $D_s$ or $D_{ST}$ (Nei 1987), whereas genetic differentiation between populations and between chromosomal arrangements was measured as $F_{ST}$ (Nei 1987) and significance estimated through permutations (Hudson et al. 1992). The mcoalism program (Ramos-Onsins and Mitchell-Olds 199) was used to perform a haplotype-based test (Hudson et al. 1994). The HKA program distributed by Jody Hey (http://lifesci.rutgers.edu/heylab) was used to perform the Hudson–Kreitman–Aguadé (HKA; Hudson et al. 1987) test. Multilocus tests. The MEGA version 4 program (Tamura et al. 2007) and the PAML version 4 software package (Yang 2007) were used for phylogenetic reconstruction and for estimating the number of branch specific substitutions, respectively.

Results

Extent of the Or22a/Or22b Duplication

Figure 1 shows schematically the structure of the Or22 region in the publicly available sequence of Drosophila melanogaster (http://flybase.org) with the Or22a and...
Or22b genes separated by an approximately 600-bp intergenic region. The dot plot analysis of an ~4.5-kb region indicates that the duplicated fragment includes the coding Or22 region as well as ~200 and ~100 bp of its 5’ and 3’ flanking regions, respectively. In D. simulans and Drosophila sechellia, two copies were also found in this region, with Or22b having become a pseudogene in the latter species; in contrast, a single copy was detected in D. yakuba and Drosophila erecta (Dobritsa et al. 2003; Drosophila 12 Genomes Consortium et al. 2007; Guo and Kim 2007; Nozawa and Nei 2007; Tunstall et al. 2007; McBride et al. 2007; Gardiner et al. 2008; present work). Therefore, the presence of two copies can be considered the derived state resulting from a duplication event that occurred after the split of the melanogaster species complex and the D. yakuba–D. erecta lineages (~12.6–12.8 Ma; Tamura et al. 2004) and prior to the split of the D. melanogaster and D. simulans lineages (~5.4 Ma; Tamura et al. 2004).

Presence of a Chimeric Gene in Some of the Studied Populations

PCR amplification of the Or22 gene region using oligonucleotides complementary to the nonduplicated flanking parts of the Or22a (5’) and Or22b (3’) genes (fig. 1) revealed that some population samples segregated for two length variants, henceforth named long and short variants. Similar results were obtained when more external amplification oligonucleotides (o1F and o1R) were used for PCR amplification. Sequencing revealed that the long variant contained both genes, whereas the short variant contained a single Or22 copy (see below).

The multiple alignment of the duplicated fragment of all sequenced lines allowed a comparison of the two copies (Or22a and Or22b) present in the long variant and the single copy present in the short variant. This comparison revealed that the degree of genetic differentiation (measured as D; Nei 1987) between the short variant and each the Or22a and Or22b copies varied along the gene: only at the first exon was differentiation lower to Or22a than to Or22b, whereas the reverse was true at the rest of the gene (fig. 2). The short variant could be considered a priori a chimeric natural allele (henceforth named Or22ab) that had originated by unequal crossing-over between the two highly differentiated copies. Southern blot analysis using four of the sequenced isochromosomal strains from the North Carolina sample (two with the short variant and two with the long variant) allowed establishing unambiguously the natural character of the chimeric allele (results not shown) and therefore discarding any possible artificial origin of the short variant by PCR-mediated recombination.

Table 1 shows the frequencies of the two length variants among the 63 sequenced alleles as well as in the 66 additional lines characterized by PCR amplification. Frequencies in each geographical area did not differ significantly between the sequenced and PCR-amplified alleles. In the extended samples (i.e., sequenced plus PCR amplified), the short variant frequency was highest in the East African sample and lowest in the European sample. No significant difference in the two length variant frequencies was detected either between the two East African samples or between the European and West African samples. Frequencies differed however significantly in all other comparisons.

Nucleotide Polymorphism at the Or22 Region

Genetic differentiation between populations (measured as FST; Nei 1987) was estimated separately for the

### Table 1

<table>
<thead>
<tr>
<th>Geographic Area</th>
<th>Population</th>
<th>Long</th>
<th>Short</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cataloni (0.00)</td>
<td>CN_seq</td>
<td>12 (1)</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>MO_amp</td>
<td>17 (2)</td>
<td>0</td>
</tr>
<tr>
<td>North America (0.28)</td>
<td>NC_seq</td>
<td>9</td>
<td>5</td>
</tr>
<tr>
<td></td>
<td>NC_amp</td>
<td>10 (1)</td>
<td>3</td>
</tr>
<tr>
<td>West Africa (0.14)</td>
<td>LA_seq</td>
<td>12* (5)</td>
<td>2 (f)</td>
</tr>
<tr>
<td></td>
<td>LA_amp</td>
<td>18 (13)</td>
<td>3</td>
</tr>
</tbody>
</table>
| East Africa (0.84) | MA_seq   | 4 (1) | 7* (1)
|                | MA_amp     | 2 (1) | 13 |
|                | MW_seq     | 0     | 12 (4) |

* Frequency of short variant in parentheses.
* CN and MO, Sant Sadurní d’Anoia and Montblanc (Spain); NC, North Carolina; LA, Lamto; and MA and MW, Malawi-1 and Malawi-2.
* seq, sequenced; amp, PCR amplified.
* Number of In(2L)t chromosomes in parentheses.
* Information for one line is missing.
* In(2L)NS with breakpoints at bands 23E2-3 and 35F1-F2.
**FIG. 3.**—Nucleotide and length polymorphisms at the Or22 region in four natural populations of *Drosophila melanogaster*. Polymorphisms are numbered from the translation initiation codon of the *Or22a* gene. Double bars indicate the duplicated fragment. Dots represent the same nucleotide as in the first sequence. Lines with the long variant are presented first and then lines with the short variant. NC, North Carolina; CN, Catalonia; LA, Lamto; MA, Malawi-1; and MW, Malawi-2. Boxed sets of polymorphism indicate gene conversion tracts.

<table>
<thead>
<tr>
<th>Exon</th>
<th>Length (bp)</th>
<th>NC</th>
<th>CN</th>
<th>LA</th>
<th>MA</th>
<th>MW</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>150</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>100</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>3</td>
<td>120</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>4</td>
<td>90</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>5</td>
<td>80</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>6</td>
<td>110</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>7</td>
<td>130</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>8</td>
<td>140</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
</tr>
</tbody>
</table>

- **a**, exon of the *Or22a* gene
- **b**, exon of the *Or22b* gene
- **d**, deletion of # bp
- **i**, insertion of # bp
- **n**, nonsynonymous
- **d**, deletion
- **i**, insertion
- **x**, unavailable sequence
- **t**, lines with the In(2L)t inversion
- **NS**, line with the In(2L)NS inversion
- **?**, lines with missing information on the chromosomal arrangement
long and short variants (supplementary table S1, Supplementary Material online). Significant genetic differentiation was detected in all population pairwise comparisons except those between the European and North American samples for the long variant and between the two East African samples for the short variant. When the joint European and North American sample (i.e., the non-African sample) and the joint East African sample were used in within-variant comparisons, a highly significant genetic differentiation was detected in all cases. Nucleotide variation analyses were thus performed with the non-African, West African, and East African samples.

The region studied is located at band 22A2, 0.7 Mb apart from the telomere-proximal breakpoint of the In(2L)t inversion at band 22D3-E1 (Andolfatto et al. 1999). In three of the studied populations, some of the sequenced lines carried this inversion (table 1 and figure 3), with the African samples exhibiting much higher inversion frequencies (0.42 and 0.27 in West Africa and East Africa, respectively) than the non-African sample (0.04). Among the sequenced lines, no association between chromosomal arrangement and length variant was detected either in the complete sample or in the different populations. Given the unique origin of the inversion, this would imply that recombination between the breakpoint and the Or22 region had been high enough over evolutionary time to erode the original associations.

For long variants, no significant genetic differentiation was detected between sequences from inverted and noninverted chromosomes, whereas differentiation was highly significant for short variants in the East African sample ($P = 0.002$). There was indeed a strong association between a highly divergent short variant haplotype and the In(2L)t arrangement in this sample (fig. 3; see below). For long variants, all nucleotide variation analyses were performed jointly for inverted and noninverted chromosomes, whereas for short variants, analyses were performed separately for inverted and noninverted chromosomes as well as for the joint sample (with and without the highly divergent haplotype).

Table 2 and figure 3 give a summary of polymorphism in the long and short variants. A single frameshift mutation was detected at the Or22b gene (fig. 3; supplementary fig. S1, Supplementary Material online). The East African sample harbored the most variation for the long variant (both for the complete region and for each the Or22a and Or22b duplicated fragments) and for the short variant (both when all chromosomes and only standard chromosomes were considered). The higher level of short-variant variation in the East African sample was only partly due to the presence of five nearly identical sequences with a very divergent haplotype, present mostly in chromosomes with the In(2L)t inversion. Indeed, when these five lines were excluded from the analysis, variation in the short variant was still higher in the Malawi than in the North Carolina sample, both when comparing all chromosomes from both samples and only Standard chromosomes (results not shown). Computer simulations revealed that the high frequency of identical sequences (five lines) for an ~2-kb fragment in the highly variable East African sample (with 19 sequences harboring 110 mutations) did not conform to neutral predictions even under the conservative assumption of no recombination ($P < 0.05$ for the one-tailed test).

Tajima’s $D$ statistic (Tajima 1989) was used to summarize the pattern of polymorphism in the long and short variants (for intermediate-sized samples, $n > 5$). Tajima’s $D$ values were negative in both the non-African and West African long variant samples (supplementary fig. S2,
Supplementary Material online), indicating an excess of low-frequency variants that was only significant in the former sample. The estimated $D$ value for the short variant was positive as a result of the five highly divergent sequences. Indeed, the value was negative ($-0.71$) when these sequences were not considered in the analysis.

Recombination and Gene Conversion

There was evidence of recombination in the history of both the long and short variants, with a minimum number of 4 and 12 recombination events, respectively. Moreover, there was evidence of gene conversion between length variants. Indeed, several gene conversion tracts between length variants were detected (Betran et al. 1997) both in the 5’ nonduplicated region and in the duplicated region (fig. 3). Some of the tracts detected in the short variant encompassed the inferred limit between the two parts of the chimeric short variant (figs. 2 and 3). Moreover, the presence of two short variants in the West African sample (LA6_s and LA116_s) with extended sequence similarity to the Or22a gene (fig. 3) would point to additional undetected gene conversion events between short and long variants or more unlikely to their independent origin by unequal crossing-over between long variants.

Genealogical Relations between Or22 Copies

Although different parts of a recombining region such as the Or22 region may have different evolutionary histories, the Neighbor-Joining tree allows describing overall relations among sequenced alleles. The duplicated region (present in Or22a, Or22b, and Or22ab) was used for phylogenetic reconstruction. As shown in figure 4, sequences

Table 3

<table>
<thead>
<tr>
<th>Region</th>
<th>Comparison</th>
<th>mel/sim</th>
<th>Syn</th>
<th>Nsyn</th>
<th>Sign</th>
<th>mel lineage</th>
<th>Syn</th>
<th>Nsyn</th>
<th>Sign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dup-down</td>
<td>mOr22ab/Or22a</td>
<td>mel</td>
<td>4</td>
<td>10</td>
<td></td>
<td>mel</td>
<td>2</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Or22b</td>
<td>43</td>
<td>24</td>
<td>**</td>
<td>F</td>
<td>16</td>
<td>10</td>
<td>**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Or22ab</td>
<td>11</td>
<td>6</td>
<td></td>
<td>P</td>
<td>8</td>
<td>4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Or22ab</td>
<td>44</td>
<td>52</td>
<td>n.s.</td>
<td>F</td>
<td>15</td>
<td>18</td>
<td>n.s.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Or22ab</td>
<td>46</td>
<td>31</td>
<td>n.s.</td>
<td>P</td>
<td>32</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Or22ab</td>
<td>34</td>
<td>41</td>
<td>n.s.</td>
<td>F</td>
<td>12</td>
<td>11</td>
<td>n.s.</td>
<td></td>
</tr>
</tbody>
</table>

* dup-down, dup-downstream region.

** m. melanogaster; s. simulans; Or22b_ab, joint sample of Or22b and Or22ab sequences; Or22b_ab*, joint sample of Or22b and Or22ab sequences excluding the five highly differentiated sequences.

*** syn, synonymous; nsyn, nonsynonymous; P, polymorphic (all populations); F, fixed.

**** significance; *, 0.01 < $P$ < 0.05; **, 0.001 < $P$ < 0.01; n.s., nonsignificant.
from the duplicated fragment clustered according to copy, resulting in two main clusters: one with all the Or22a copies from *D. melanogaster* and also from *D. simulans* and *D. sechellia* and another with all the Or22b and Or22ab copies from *D. melanogaster* and the Or22b copy of *D. simulans*.

Both the Or22a and Or22b copies present in the long variant had a shallow genealogy that reflected their low level of variation. In the cluster formed by the Or22b and Or22ab copies, there was a deep split that separated the five Or22ab sequences with the highly divergent haplotype and a cluster including the rest of the short variant sequences and all the Or22b sequences that formed a derived group. The divergent haplotype might be indeed representative of the ancestral unequal crossing-over event, with the other Or22ab sequences representing a more recent conversion event(s) from a further differentiated Or22b allele (fig. 4).

The position in the reconstructed trees of the single Or22 copy present in *D. yakuba* and *D. erecta* was not always the same. For synonymous sites, they both had an outgroup position in agreement with the species history (fig. 4). In contrast, for nonsynonymous sites, the *D. yakuba* and *D. erecta* sequences clustered with the Or22a copies (fig. 4). Likewise, in the phylogeny reconstructed from amino acid sequences, the OR22 sequence of *D. yakuba* and *D. erecta* also clustered with the OR22a copy (result not shown). The differential position of the sequences from these species in the trees reconstructed from synonymous and nonsynonymous sites would support a scenario where purifying selection would have acted on the OR22a copy to preserve its ancestral function, whereas the evolution of the OR22b copy would have been either driven by adaptive selection or by the relaxation of purifying selection resulting from its putative dispensability.

Variation within and between Species and Duplicates

The MK test (McDonald and Kreitman 1991) was performed using polymorphism in *D. melanogaster* and divergence from *D. simulans* (and also divergence since the species split). General and lineage-specific tests were performed for the duplicate downstream region (i.e., that part shared by the Or22b and Or22ab copies; fig. 2). For the Or22a region, a significant departure from the expected proportionality between nonsynonymous and synonymous polymorphic and fixed changes was detected, pointing to an excess of nonsynonymous polymorphism (table 3). For the Or22b region, no significant departure from neutral predictions was detected except for the general test performed excluding the five highly divergent Or22ab sequences (table 3). The MK test results would provide, therefore, no clear evidence for adaptive nonsynonymous fixations in the recent history of either duplicate.

In order to ascertain whether the low level of variation detected at the long variant in the non-African samples might reflect recent selection, a multilocus HKA (Hudson et al. 1987) test was performed. The test compared silent polymorphism (at the European sample and also at the joint non-African sample) and divergence from *D. simulans* at the complete Or22 region with polymorphism and divergence estimates at a subset of 30 noncoding fragments from a previous genome scan from the same European population (Orengo and Aguade 2004), for which there was no evidence for any decoupling between polymorphism and divergence (\( \chi^2 = 20.26, P = 0.86 \) from simulations). The tests performed after adding the Or22 long variant region yielded significant results (table 4), with a very important contribution of polymorphism and divergence at the Or22 region to the test statistic. In contrast, no significant decoupling between levels of silent polymorphism and divergence was detected when the short variant (Or22ab) was used in the analysis. Silent polymorphism at the Or22 region exhibited, therefore, a significant reduction relative to divergence only at the long variant.

The number of fixed and shared differences between copies as well as the distribution of between-copy genetic differentiation across the duplicates are shown in supplementary figure S3 (Supplementary Material online). As expected from the putative age of the duplication, the number of fixed differences detected in all comparisons greatly exceeded the number of both copy exclusive and shared polymorphisms. Divergence between copies was much higher in the second-exon region, with 63 and 72 synonymous and nonsynonymous differences fixed between the long variant copies (and 52 and 64 between Or22a and Or22b, ab).

Maximum likelihood was used to infer the number of synonymous and nonsynonymous substitutions in the early history of the duplicates (i.e., in the branch from their origin to the *D. melanogaster*D. simulans split). In addition to the sequences of *D. yakuba* (Or22), *D. erecta* (Or22), *D. simulans* (Or22a and Or22b) and *D. sechellia* (Or22a), a single randomly chosen sequence from each gene of *D. melanogaster* (Or22a and Or22b) was used for that purpose. The ratio of nonsynonymous to synonymous substitutions differed significantly between branches: 15 synonymous versus 6 nonsynonymous at the Or22a branch and 45 synonymous versus 59 nonsynonymous at the Or22b branch (G-test with Williams correction, \( P = 0.019 \)). Indeed, a relative excess of nonsynonymous substitutions was detected at the Or22b branch. Similar results were obtained when only substitutions in the second exon were considered (18 synonymous vs. 7 nonsynonymous and 30 synonymous vs. 47 nonsynonymous at the Or22a and Or22b branches, respectively; G-test with Williams correction, \( P = 0.004 \)) and also when parsimony was used to infer substitutions, although in the latter case the total number of substitutions compared was lower as a result of ambiguities (results not shown).

### Table 4

<table>
<thead>
<tr>
<th>Test Components</th>
<th>EU,</th>
<th>nAF,</th>
<th>all,</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \chi^2 )</td>
<td>46.15</td>
<td>52.44</td>
<td>20.27</td>
</tr>
<tr>
<td>d.f.</td>
<td>30</td>
<td>30</td>
<td>30</td>
</tr>
<tr>
<td>( P ) value (sim.)</td>
<td>0.037</td>
<td>0.007</td>
<td>0.888</td>
</tr>
<tr>
<td>Pol contribution</td>
<td>3.899</td>
<td>5.069</td>
<td>0.248</td>
</tr>
<tr>
<td>Div contribution</td>
<td>4.821</td>
<td>6.308</td>
<td>0.356</td>
</tr>
</tbody>
</table>

* d.f., degrees of freedom; \( P \) value (sim.), probability obtained from simulations; pol contribution, contribution from polymorphism to the test statistic; and div contribution, same from divergence.

EU, European long variant sample; nAF, non-African long variant joint sample; and all, short variant.
Variation within and between Species and Copies at the Protein Level

The OR22a and OR22b proteins exhibited 9 and 6 different haplotypes, respectively (supplementary fig. S1, Supplementary Material online), without considering that resulting from the single frameshift mutation at the OR22 gene. The most frequent haplotype of each protein was present in all 4 sampled populations. The chimeric OR22ab protein is essentially an OR22b protein, and the chimeric gene could therefore be considered an OR22a deletion. As expected from the analysis of nonsynonymous substitutions, the OR22b protein exhibited a higher number of amino acid replacements than the OR22a protein both at its distant past and more recently (i.e., soon after its origin and after the D. melanogaster and D. simulans split). Moreover, most amino acid replacements between D. melanogaster and D. simulans were in the region encoded by exon 2. Similarly, most amino acid replacements in the early history of the OR22b duplicate occurred in those structural domains encoded by exon 2. Indeed, the relative number of amino acid replacements in the first 256 residues of the protein, which include these domains as well as the N-terminal domain, was significantly higher than in the remaining 141 residues: 42 and 8 replacements, respectively (G-test with Williams correction, $P = 0.001$).

Discussion
Gene Duplicates, Chimeric Gene, and Nucleotide Polymorphism

The survey of variation at the region spanning the tandemly arranged OR22a and OR22b genes (long variant) revealed the presence of a chimeric short variant (OR22ab; fig. 2) in three of the four populations sampled. Its frequency differed however between these three populations, with the East African and West African samples exhibiting the highest and lowest frequencies, respectively. Turner et al. (2008) recently reported the presence of the short variant at latitudinally varying frequency in Australia.

The high level of nucleotide diversity detected at the chimeric OR22ab gene in the East African sample and the presence of a highly divergent haplotype would support its rather old origin in the D. melanogaster lineage (figs. 3 and 4). Indeed, the estimated average pairwise distance between the dup-downstream region of this highly divergent haplotype and the rest of chimeric alleles was rather high ($D_{XY} = 0.038$) although lower than the estimated divergence between D. melanogaster and D. simulans at the comparable region of the OR22b gene ($K = 0.112$; fig. 2).

The highly divergent and old OR22ab haplotype harbors little variation despite its relatively high frequency in East Africa (5 out of 19; fig. 3). This reduced variation in the Malawi population and its partial association with the In(2)Lt inversion would suggest the rather recent arrival in this population of an inverted chromosome carrying this haplotype and its subsequent increase in frequency, likely driven by selection. The time elapsed in this population would have been sufficient for breaking up its initial complete association with the inversion but not long enough for it to have recovered variation either by mutation or by recombination with the remaining haplotypes.

The differential frequency of the chimeric Or22ab gene in the North American and European samples (table 1) seems at odds with the nonsignificant genetic differentiation detected between populations not only for the long variant of the Or22 region but also at other genomic regions previously surveyed in these or nearby populations (Aguadé et al. 1992; Martín-Campos et al. 1992; Aguadé 1998). The detected disparity among regions and the rather low variation of the short variant in the North American sample, together with the latitudinal variation in its frequency recently detected in Australia (Turner et al. 2008), would point to the locus-specific action of selection. Despite the recently reported abundance of copy-number deletion variation affecting coding regions in D. melanogaster (Emerson et al. 2008), the particular copy-number deletion variation detected at the Or22 region and the deletion variation previously detected at the Obp57e gene (Takahashi and Takano-Shimizu 2005) would be among the few cases with some evidence for the polymorphism being adaptive.

As compared with the short variant, nucleotide variation in the long variant was low in all sampled populations. Nevertheless, the East African population exhibited more variation than the other populations. In the Or22b gene region, however, gene conversion events from the short variant would have greatly contributed to extant long variant diversity in the small East African sample and also in the larger West African sample.

In the non-African sample, and more specifically in the European sample, estimated silent nucleotide diversity in the long variant ($\pi = 0.001$) was well below the average level of variation detected at a large number of noncoding autosomal fragments in this geographic area ($\pi_{sil} = 0.0064$ in Hutter et al. [2007]). In contrast, the estimated silent divergence from D. simulans at the complete Or22 region ($K_{sil} = 0.146$) was at the uppermost part of the estimated divergence range for those fragments and well above its average value ($K_{sil} = 0.05$). Nucleotide diversity in the non-African sample was low across the complete region, that is, both at the Or22a and Or22b duplicated fragments (table 2 and fig. 3). A significant reduction of variation was indeed detected at this region when compared with a subset of 30 noncoding neutrally evolving X chromosome fragments from a previous genome scan (Orengo and Aguadé 2004) by the multilocus HKA test. Given the ancestral character of the long variant in D. melanogaster as well as in D. simulans, its lower level of variation relative to the short variant and its consequently reduced variation relative to interspecific divergence would support the recent fixation in this variant (or at a nearby region) of an advantageous mutation. The skew in the frequency spectrum toward low-frequency variants (as revealed by Tajima’s $D$ negative values) and the star-like genealogy of long variant Or22 copies would be consistent with this hypothesis (fig. 4). The most parsimonious scenario would be for the within-variant fixation to have occurred prior to the out-of-Africa expansion of the species, with subsequent gene conversion events having blurred its effect in the East African population. Moreover, the recombinational distance of the Or22 region to the In(2)Lt inversion proximal breakpoint would explain the presence of the newly selected variant in both inverted and noninverted chromosomes.
The Or22a and Or22b Duplicates History

The most common fate of gene duplicates, similarly to that of point mutations, is the loss of the new copy soon after its origin. The Or22a/Or22b genes constitute one successful duplicate pair that originated by tandem duplication of the complete single-copy Or22 gene present in the D. yakuba-D. melanogaster species complex ancestor and attained fixation soon after its split, as reflected by both copies being present in all species of the D. melanogaster complex (with the Or22b copy having become a pseudogene in the specialist species D. sechellia).

Comparison of both genes with the single-copy outgroup revealed some asymmetry in the rate of evolution of both copies. The Or22b copy evolved more rapidly than the Or22a copy, especially at nonsynonymous sites (fig. 4). The number of both synonymous and nonsynonymous substitutions in their early history (i.e., prior to the split of the D. melanogaster and D. simulans lineages) was higher at the Or22b branch, with a relative excess of nonsynonymous substitutions. This excess could reflect the action of positive selection driving some nonsynonymous mutations to fixation and therefore contributing to functional divergence between copies. It could also reflect the differential relaxation of purifying selection at synonymous and nonsynonymous sites of the Or22b copy, under the assumption that in the single-copy ancestor (and also in the Or22a copy) both kinds of sites were subject to purifying selection.

No decoupling between levels of polymorphism and divergence at synonymous and nonsynonymous sites of the Or22b gene was detected in the D. melanogaster lineage, whereas an excess of nonsynonymous polymorphism was detected at the Or22a gene (table 3). These results do not support the adaptive evolution of the encoded proteins since the D. melanogaster/D. simulans split. The relative low ratios of nonsynonymous to synonymous divergence between these species for both genes (0.18 and 0.33, respectively) would support that they have both been subject to purifying selection in their recent past, even if its strength has been lower at Or22b. This together with the nonhomo- geneous distribution of substitutions across the gene (see below), the conservation of regulatory motifs and gene structure, and the presence of the chimeric Or22ab gene segregating at high frequency in a putatively ancestral population would not favor a nonfunctionalization scenario either in the early history of the Or22b gene or soon after the D. melanogaster/D. simulans split. Indeed, pseudogenes are rather short lived in Drosophila given the detected bias toward deletions in this genus (Ramos-Onsins and Aguadé 1998; Petrov et al. 2000; Ometto et al. 2005).

Divergence between the Or22a and Or22b copies was not homogeneously distributed along the copies, as reflected by the second exon exhibiting the highest number of nonsynonymous substitutions not only between extant copies (results not shown) but most importantly also in the early history of the duplicates. The high differentiation detected at exon 2 is paralleled in extant copies by a relatively high proportion of sites affected by length differences in the second intron (ca. 60%) as compared with introns 1 (0%) and 3 (ca. 40%). The rapid accumulation of both point substitutions and fixed length changes might have contributed to counteract the homogenizing role of gene conversion between the young duplicates and also later in their history (Teshima and Innan 2004).

The copy-number polymorphism resulting from segregation of the chimeric Or22ab gene in some populations might be compatible with at least some functions being redundant between copies. The chimeric gene would be expressed not only at the same neurons, similarly to the Or22a and Or22b genes, but most importantly also at the same life stages as the Or22a gene, with which it shares its complete 5' flanking region. Further functional characterization of the Or22a and Or22b genes as well as of the chimeric form will undoubtedly contribute to disentangle putative redundant from novel functions between copies.

In summary, the evolutionary history of the Or22 region is rather complex, not only because it has suffered a recent and successful duplication event but also because unequal crossing-over between copies has led to a rather successful chimeric copy. Moreover, the comparative sequence analysis within and between copies and species has uncovered the footprint of positive selection acting at least in its recent history. Pinpointing the putative targets of recent events will require their further characterization both at the comparative sequence level and at the functional level.

Supplementary Material

Supplementary figures S1–S3 and table S1 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments


Literature Cited


