

Positive and Negative Selection in the β -Esterase Gene Cluster of the *Drosophila melanogaster* Subgroup

Evgeniy S. Balakirev,¹⁻³ Maria Anisimova,^{4,5} Francisco J. Ayala¹

¹ Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA

² Academy of Ecology, Marine Biology, and Biotechnology, Far Eastern National University, Vladivostok 690600, Russia

³ Institute of Marine Biology, Vladivostok 690041, Russia

⁴ LIRMM, Université Montpellier II, Montpellier Cedex 5, 34392 France

⁵ Department of Biology and Center for Mathematics and Physics in the Life Sciences and Experimental Biology, University College London, London WC1E 6BT, UK

Received: 7 June 2005 / Accepted: 20 December 2005 [Reviewing Editor: Dr. Martin Kreitman]

Abstract. We examine the pattern of molecular evolution of the β -esterase gene cluster, including the *Est-6* and ψ *Est-6* genes, in eight species of the *Drosophila melanogaster* subgroup. Using maximum likelihood estimates of nonsynonymous/synonymous rate ratios, we show that the majority of *Est-6* sites evolves under strong (48% of sites) or moderate (50% of sites) negative selection and a minority of sites (1.5%) is under significant positive selection. *Est-6* sites likely to be under positive selection are associated with increased intraspecific variability. One positively selected site is responsible for the EST-6 F/S allozyme polymorphism; the same site is responsible for the EST-6 functional divergence between species of the *melanogaster* subgroup. For ψ *Est-6* 83.7% sites evolve under negative selection, 16% sites evolve neutrally, and 0.3% sites are under positive selection. The positively selected sites of ψ *Est-6* are located at the beginning and at the end of the gene, where there is reduced divergence between *D. melanogaster* and *D. simulans*; these regions of ψ *Est-6* could be involved in regulation or some other function. Branch-site-specific analysis shows that the evolution of the *melanogaster* subgroup underwent

episodic positive selection. Collating the present data with previous results for the β -esterase genes, we propose that positive and negative selection are involved in a complex relationship that may be typical of the divergence of duplicate genes as one or both duplicates evolve a new function.

Key words: *Drosophila melanogaster* subgroup — *Est-6* — ψ *Est-6* — Positive selection — Negative selection — d_N/d_S rate ratio — Intergene

Introduction

The conventional model of gene evolution by duplication is based on the assumption that one gene copy is sufficient to perform the respective function, so that a gene duplication is redundant and has no effect on fitness (Ohno 1970; Kimura and King 1979). After gene duplication, either one paralogue will acquire a new function or, more often, one paralogue will be lost through the fixation of null alleles (Ohno 1970; Nei and Roychoudhury 1973; Bailey et al. 1978; Ohta 1988; Walsh 1995). The process of nonfunctionalization (pseudogenization) is relatively fast and usually occurs in the first few millions years after duplication if the duplicated gene is not under any

Correspondence to: Evgeniy S. Balakirev; c/o Francisco J. Ayala, Department of Ecology and Evolutionary Biology, 321 Steinhaus Hall, University of California, Irvine, CA 92697-2525, USA; email: esbalak@bio.dvgu.ru

selection (Lynch and Conery 2000). However, analysis of sequence data tends to confirm the conclusion that a substantial proportion of duplicates is retained in evolution (e.g., Nadeau and Sankoff 1997; Hileman and Baum 2003; Zhang 2003). The subfunctionalization model (or duplication-degeneration-complementation [DDC] model) seeks to resolve the observed discrepancy between theory and experimental data (Force et al. 1999; Lynch and Force 2000). The model assumes that degeneration of both paralogues occurs during a period of relaxed selection following the duplication event. If a population is small enough, then both paralogous copies might accumulate mutations that impair different functions of the ancestral gene, so that each daughter gene adopts part of the functions of the parental gene and none of the paralogues is capable of substituting for the ancestor; both copies are partially degraded by mutations to the extent that their joint expression is necessary to fulfill the essential functions of the ancestral locus (Force et al. 1999; Stoltzfus 1999; Lynch and Force 2000). The DDC model is based on the same initial assumption of redundancy as the classical model (Ohno 1970; Kimura and King 1979) and it is applicable mostly to small populations because the probability of subfunctionalization approaches zero in large populations (Lynch et al. 2001). In contrast to the classical theory of duplication and to the subfunctionalization model, Kondrashov et al. (2002) have argued that the retention of duplication is an adaptive process. They show that paralogues evolve under purifying selection at a significantly faster rate than unduplicated genes with a similar level of divergence. A majority of fixed duplicated genes increase fitness when present at two or more copies in a genome and thus may have been subject to purifying selection from the moment of duplication (Kondrashov et al. 2002).

The persistence of functional duplications becomes an even more challenging issue when considering the retention of duplications turned into pseudogenes. Some duplicated genes are maintained in the genome for a long time for a specific function but may later become pseudogenes because of relaxation of functional constraints (e.g., Rouquier et al. 2000). Moreover, pseudogenes may be functional (for a review see, Balakirev and Ayala 2003a) and thus be long maintained in the genome. True pseudogenes are thought to be free of purifying selection if the ancestral gene remains functional; thus accumulation of recurrent mutations (detrimental or not) would lead to degenerative disintegration and melting into the genomic background. However, even in the case of fixation of null alleles and complete pseudogenization of the duplicate, the process of genetic degeneration is not so obvious. In fact, cases of strong genetic degradation have been described only

for obligate intracellular pathogens as a result of long-term association with eukaryotic hosts (Andersson and Andersson 2001; Moran 2002). Pseudogenes are widespread (Harrison et al. 2001, 2002, 2003; Harrison and Gerstein 2002; Torrents et al. 2003), and contrary to theoretical predictions, they often maintain their structure as do functional entities of genome. In many cases strong conservation and even expression of pseudogene sequences have been observed (for a recent review see Balakirev and Ayala 2003a).

Est-6 and $\psi Est-6$ have very similar exon-intron structure (Oakeshot et al. 1987; Collet et al. 1990) and are closely linked, located on the left arm of chromosome 3 of *Drosophila melanogaster* (at 68F7-69A1 in the cytogenetic map). We have previously investigated the evolution of the β -*esterase* gene cluster, including *Est-6* and $\psi Est-6$, in four natural populations of *D. melanogaster* (Balakirev and Ayala 1996, 2003b,c, 2004; Balakirev et al. 1999, 2002, 2003; Ayala et al. 2002). We have now investigated the β -*esterase* gene cluster in eight species of the *Drosophila melanogaster* species subgroup. We estimate the ratio of nonsynonymous-to-synonymous substitutions, $\omega = d_N/d_S$, using maximum likelihood models developed by Yang (1998), Nielsen and Yang (1998), Yang et al. (2000), and Yang and Nielsen (2002). The application of these models has led to detecting positive selection in genes for which the selective hypothesis was rejected using pairwise comparisons between sequences (Zanotto et al. 1999; Yang et al. 2000; Bielawski and Yang 2001, 2003). The accuracy and power of the likelihood ratio tests for detecting positive selection have been extensively tested (Anisimova et al. 2001, 2002, 2003). These models make it possible also to identify critical amino acids under diversifying selection, without knowledge of the functionally important domains. Using this approach, we show that positive Darwinian selection promotes the polymorphism and divergence of both *Est-6* and $\psi Est-6$ and that there is a complex relationship between positive and negative selection in a process that involves functional divergence of duplicate genes and evolution of a new function.

Materials and Methods

Drosophila Strains and Species

The 78 strains of *D. melanogaster* are from random samples of wild flies collected in Africa (Zimbabwe), Europe (Barcelona, Spain), North America (El Rio, Acampo, California), and South America (Venezuela). The non-African strains were made fully homozygous for the third chromosome by crosses with balancer stocks (Seager and Ayala 1982; see Balakirev et al. 2002; Balakirev and Ayala 2003b). Chung-I Wu kindly provided the *D. melanogaster* strains from East Africa (Sengwa and Harare, Zimbabwe). *D. sechellia*,

D. mauritiana, *D. erecta*, *D. teissieri*, and *D. orena* strains were obtained from the *Drosophila* Species Stock Center (Bowling Green, OH). *D. simulans* strain is from Ayala's laboratory.

DNA Sequence Analysis

The procedures for DNA extraction, amplification, and sequencing have been described previously (Balakirev et al. 1999, 2002, 2003). For each line, the sequences of both strands were determined, using 24 overlapping internal primers spaced, on average, 350 nucleotides. At least two independent PCR amplifications were sequenced in both directions to prevent PCR or sequencing errors. The population data for *D. melanogaster* are from Balakirev and Ayala (2003b, c, 2004); see GenBank accessions AF147095–AF147102, AF150809–AF150815, AF217624–AF217645, AF526538–AF526559, AY247664–AY247713, AY247987–AY248036, AY368077–AY368109, and AY369088–AY369115. The data for other *D. melanogaster* subgroup species are from Balakirev et al. (2005); see GenBank accessions AY695919–AY695924. The *D. yakuba Est-6* sequence (Oakeshott et al. 2001) is from GenBank (AJ279007). The esterase sequences were assembled using the program SeqMan (Lasergene; DNASTAR, Inc., 1994–1997). Multiple alignment was carried out manually and using the program CLUSTAL W (Thompson et al. 1994). The program DIVERGE, version 1.04 (Gu 1999), was used to analyze the gene functional divergence after duplication or speciation.

Maximum Likelihood Analysis

We measure selective pressure by the ratio $\omega = d_N/d_S$, where d_N is the rate of nonsynonymous substitutions per nonsynonymous site and d_S is the rate of synonymous substitutions per synonymous site. Deleterious mutations are eliminated by negative (purifying) selection, causing the synonymous rate to be higher than the nonsynonymous rate ($\omega < 1$); $\omega > 1$ is an indication of positive selection. In neutrally evolving genes the nonsynonymous and synonymous rates are expected to be similar, $\omega \approx 1$. We test for neutrality using a likelihood ratio test (LRT) by comparing the log-likelihood values l_0 and l_1 obtained under the null hypothesis “ ω fixed at 1” and the alternative “ ω is estimated.” LRTs can be performed only for a pair of nested hypotheses.

Averaging ω across the sites in the sequence does not affect the results of the LRT for neutrality. When the null “ $\omega = 1$ ” cannot be rejected, we cannot conclude that a gene is evolving neutrally since an estimate of ω close to 1 could be an artifact of averaging across sites. Since assuming a uniform ω ratio on a functional gene makes it impossible to infer levels of selective pressures at particular sites and reduces the power of detecting positive selection, models of heterogeneous selective pressures were also used (Yang 1998; Yang et al. 2000; Bielawski and Yang 2001, 2004; Yang and Nielsen 2002). An empirical Bayesian approach was used to infer sites under different functional constraints and hence to investigate patterns of gene evolution. The direction of the amino acid change was inferred using ancestral reconstruction by maximum likelihood.

Models for Detecting Heterogeneous Selective Pressures on a Protein; Site-Specific Models

We investigate the variability of selective constraints on *Est-6* and $\psi Est-6$ with the following models of variable selective pressures among sites: M0 (one ratio), M1 (neutral), M2 (selection), M3 (discrete), M7 (β), and M8 (β & ω) (Yang et al. 2000). The simplest model, M0, assumes one ω ratio for all sites; model M1 allows two classes, conserved sites with $\omega_0 = 0$ and neutrally evolving sites

with $\omega_1 = 1$; model M7 allows several site classes with ω ratios drawn from the β distribution $B(p, q)$ and, hence, limited between 0 and 1. The three models, M0, M1, and M7, are taken as null hypotheses in the LRTs against their alternative models, M3, M2, and M8, respectively. Model M3 allows K discrete site classes with ω ratios $\omega_0, \omega_1, \dots, \omega_{K-1}$ taken in proportions p_0, p_1, \dots, p_{K-1} . Here we use $K = 3$ as suggested by Yang et al. (2000); note that $K = 1$ in M0. Model M2 adds an extra class to M1 with an ω_2 estimated from the data. Similarly, model M8 adds one discrete class to M7 with ω estimated from the data. Thereby, we consider three LRTs: (i) M0 (one ratio) vs. M3 (discrete), (ii) M1 (neutral) vs. M2 (selection), and (iii) M7 (β) vs. M8 (β & ω). The LRT comparing M0 and M3 tests for variability of selective pressure among sites, whereas LRTs of M1 vs. M2 and M7 vs. M8 are tests for diversifying selection.

We analyze data using null and alternative models and calculate the LRT statistic $2\Delta l$ (twice the log-likelihood difference). Maximum likelihood computation is performed with the codeml program from PAML (Yang 1997). The distribution of $2\Delta l$ among the replicates is then compared with a χ^2 distribution, where the degrees of freedom ν are the difference in the number of free parameters between the two nested models; we use $\nu = 4$ for the comparison M0–M3 and $\nu = 2$ for the comparisons M1–M2 and M7–M8. Note, however, that neither of these test statistics follows the χ^2 distribution, causing the LRTs to be conservative (Anisimova et al. 2001). The significance of positive selection was also assessed with a modified M7–M8 test (Swanson et al. 2003), which yielded the same results.

Branch-Specific Models

Codon models that allow independent ω ratios in different branches of a topology enable the analysis of changes in selective pressure with time (Yang 1998; Bielawski and Yang 2001). The simplest model that does not allow variation along lineages is the one-ratio model R1 (equivalent to M0 above). The most flexible lineage-specific model is the free-ratio model, which assumes different ω ratios for all branches of the topology. Intermediate models can be constructed by constraining some branches to the same ω ratio. In a simple case where there is only one duplication event, model R2 (two-ratio model) assumes two independent ω ratios: for example one for branches preceding the duplication event and another for all branches following the duplication event. Models R1 and R2 are nested, and the LRT comparing them examines whether there is a difference between average selective constraints before and after the duplication. R3 assumes three independent ω ratios: for example, one for the branches preceding the duplication event and another two for the branches following the duplication event in two clades. This model allows selective pressure to vary between paralogous genes. The LRT of models R2 and R3 tests whether, after the duplication event, the two paralogues evolve under different selective constraints.

Branch-Site Models: Lineage Specific Changes of Selective Pressure at Specific Codon Sites

Yang and Nielsen (2002) combine the site-specific and branch-specific models in order to detect episodic changes in selective constraints across specific sites. These models are useful for detecting adaptive evolution in the branches immediately following a duplication event. The branches of interest are referred to as “foreground” branches and the other branches are called “background” branches. Selective constraints are assumed to vary across sites, but at a subset of sites selective constraints also vary along “foreground” branches. The simplest model has four site classes: two site classes are uniform along the topology (with ratios ω_0, ω_1);

the other two classes allow the ω ratio at a fraction of sites to vary (to ω_2) along the foreground branches. Two versions of branch-specific models are implemented by Yang and Nielsen (2002). In model A the two uniform classes have fixed ratios: $\omega_0 = 0$ and $\omega_1 = 1$. Hence, ω can differ from 0 and 1 only in the foreground branches. The LRT comparing MA and M1 (neutral) is a good test for episodic positive selection. In model B, ratios ω_0 and ω_1 are free parameters so that some sites can be under positive selection throughout the phylogeny. Model B can be compared with M3 (discrete) with $K = 2$ site classes.

Another branch-site model (model D) has recently been proposed (Bielawski and Yang 2004). This model is powerful for detecting differences in selective constraints in the evolution of two paralogues. It allows K site classes (implemented for $K = 2$ and 3). $K - 1$ classes have a ω ratio uniform over time. The remaining class allows sites to evolve under different selective constraints in different parts of the topology. Case 1 requires a rooted topology and distinguishes the two postduplication event clades by allowing some sites to evolve with different ω ratios. Case 2 does not require a root and allows some sites to evolve with different ω ratios in background and foreground branches.

Arrangement of Data Analyses

The *Est-6* and ψ *Est-6* genes from seven species of the *D. melanogaster* subgroup and 78 lines of *D. melanogaster* were aligned; the gaps, termination codons, and introns were removed. The total sequence length was 1608 bp. Based on these sequences we formed three datasets. The first dataset, (a) *Est-6* (7) and ψ *Est-6* (7), consists of the sequences of the *Est-6* and ψ *Est-6* genes from seven species of *D. melanogaster* subgroup. The second dataset, (b) *Est-6* (85) and ψ *Est-6* (85), is composed of dataset (a) plus 78 sequences of the *Est-6* and ψ *Est-6* genes from four natural populations of *D. melanogaster*. The third dataset, (c) *Est-6* (8), includes only *Est-6* sequences from eight species of *D. melanogaster* subgroup. The seven *Est-6* sequences of this dataset are from dataset (a), while the *Est-6* sequence of *D. yakuba* is obtained from the literature (Oakshott et al. 2001; GenBank accessions AJ279007). The three types of maximum likelihood analysis described above were performed for each dataset.

Results

(a) *Est-6* and ψ *Est-6* Dataset (Seven Species)

Site-Specific Models. The likelihood ratio tests (LRT) of neutrality are highly significant for both *Est-6* ($p < 10^{-8}$) and ψ *Est-6* ($p < 10^{-4}$). This means that none of the genes evolves neutrally (with $\omega = 1$, on average) and strong selective forces are involved: $\omega = 0.28$ for *Est-6* and $\omega = 0.22$ for ψ *Est-6*. Site-specific models (Table 1) are used to explore differences in evolutionary patterns between *Est-6* and ψ *Est-6*. LRTs comparing M0 and M3 are highly significant, indicating heterogeneous selective pressure along the sequences for both genes. For the *Est-6* gene the estimates suggest that 45.0% of sites are highly conserved with $\omega_0 = 0.00$, 49.0% of sites evolve under purifying selection with $\omega_1 = 0.40$, and the remaining 6.0% of sites evolve under positive selection with $\omega_2 = 1.62$. The distribution of selective constrains across ψ *Est-6* is different: 85% of sites

Table 1. Maximum likelihood (ML) estimates for site-specific models of the β -*esterase* genes in seven species of the *D. melanogaster* subgroup

LRT	df	ML estimates under the null	l_0	ML estimates under the alternative	l_1	p value
<i>Est-6</i> M0 vs. M3	4	M0 (one ratio): $\omega = 0.28$	-4285.44	M3 (discrete with $K = 3$): $\omega_0 = 0.00, \omega_1 = 0.40, \omega_2 = 1.62$ $p_0 = 0.45, p_1 = 0.49, (p_2 = 0.06)$	-4263.81	$< 10^{-8}$
	2	M1 (neutral): $p_0 = 0.62, p_1 = 0.38$	-4269.26	M2 (selection): $\omega_2 = 0.21, (p_2 = 0.48)$	-4264.06	$< 10^{-2}$
	2	M7 (B): $p = 0.22, q = 0.54$	-4264.37	M8 (B& ω): $p_0 = 0.95, p = 0.40, q = 1.32$ $(p_1 = 0.05), \omega = 1.62$	-4263.92	0.64
ψ <i>Est-6</i> M0 vs. M3	4	M0 (one ratio): $\omega = 0.22$	-4415.47	M3 (discrete with $K = 3$): $\omega_0 = 0.11, \omega_1 = 0.11, \omega_2 = 0.91$ $p_0 = 0.21, p_1 = 0.64, (p_2 = 0.15)$	-4401.83	$< 10^{-4}$
	2	M1 (neutral): $p_0 = 0.64, p_1 = 0.36$	-4416.10	M2 (selection): $\omega_2 = 0.12,$ $(p_2 = 0.86) p_0 = 0.00, p_1 = 0.14$	-4401.87	$< 10^{-6}$
	2	M7 (B): $p = 0.37, q = 1.22$	-4402.42	M8 (B& ω): $p_0 = 0.85, p = 12.83, q = 99$ $(p_1 = 0.15), \omega = 0.92$	-4401.85	0.56

Note. The proportion of sites in the class with ω_j is denoted p_j . Statistically significant parameter estimates of positive selection are in boldface. Parameters for model M2 were obtained at the suboptimal peak (see Anisimova and Yang 2004).

evolve under strong purifying constraint with $\omega_0 = \omega_1 = 0.11$ and 15% of sites evolve with $\omega_1 = 0.91$, which could be indicating neutral evolution. None of the LRTs suggests the existence of sites evolving by positive selection for $\psi Est-6$. Comparison of models M1 (neutral) and M2 (selection) is also significant for both genes; the M2 model fits the data better allowing an extra class of sites evolving under weak purifying selection. The LRTs comparing models M7 (β) and M8 ($\beta&\omega$) are not significant (note, however, that for the *Est-6* gene the model M8 suggests that 5% of sites are under positive selection with $\omega = 1.62$, yet it does not fit the data significantly better than the model M7).

Branch-Specific Models. In order to further investigate the evolutionary patterns in the *Est-6* and $\psi Est-6$ clades we use branch-specific and branch-site models. Table 2 summarizes the results of the LRTs performed on the combined *Est-6* and $\psi Est-6$ dataset using models that allow variation of selective constraints only along evolutionary history. Figure 1 shows an unrooted NJ tree used in the analysis with background/foreground partitions. The branch connecting the *Est-6* and $\psi Est-6$ clades is denoted as a background branch, and the ω ratio for this branch as ω_{backgr} . The others are foreground branches with ω ratios as $\omega_{foregr-6}$ in the *Est-6* clade and $\omega_{foregr-P}$ in the $\psi Est-6$ clade. The LRT comparing models M0 (one ratio) and MR2 (two ratios) is significant, indicating that selective pressure is significantly different on the background and foreground branches; i.e., immediately after the duplication event most of the sites in both clades evolve under strong purifying selection ($\omega_{backgr} = 0.09$), but later negative selection is weakened ($\omega_{foregr} = 0.24$) (Fig. 1). This observation is in good agreement with the results of Kondrashov et al. (2002) showing that both paralogues produced by a duplication are subject to purifying selection.

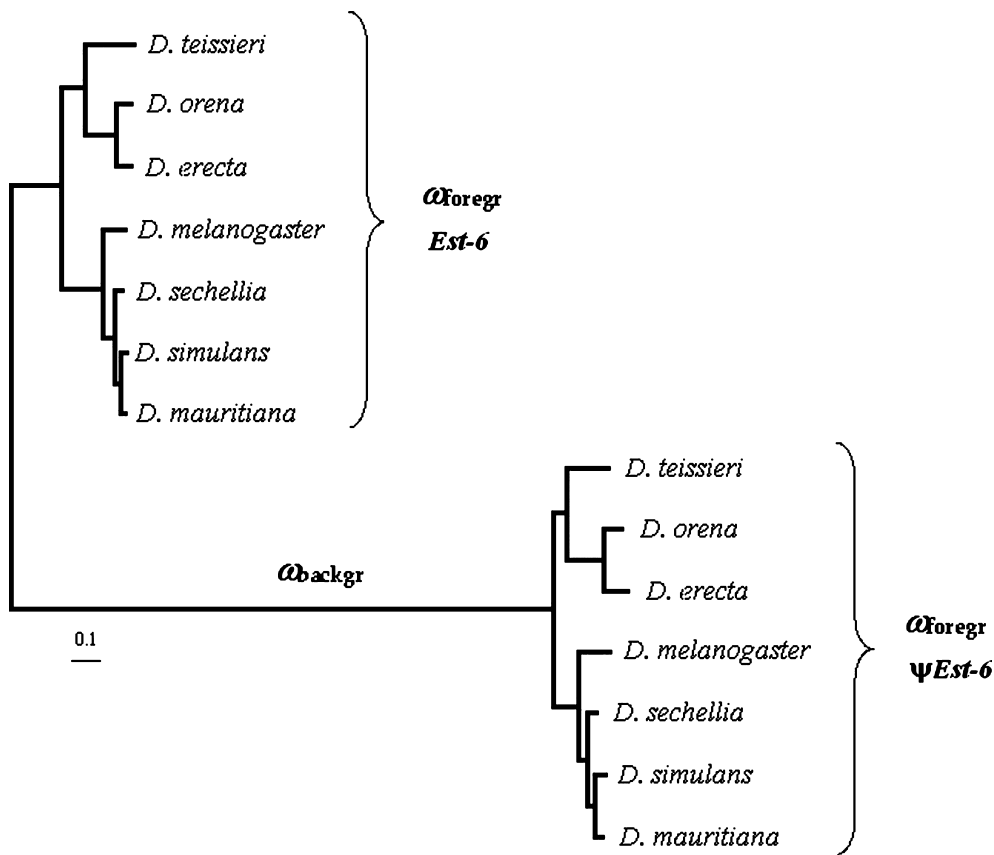
A good test for analyzing the differences in functional constraint between the two paralogues is the LRT comparing models R2 (two ratios) and R3 (three ratios). This test is not significant: selective pressures averaged over sites appear to be similar in both clades. Models that allow variation of selective constraints over time as well as across sites confirm positive selection on *Est-6* and $\psi Est-6$, but they are unable to pick up significant differences in selective pressure between the two genes (results not shown).

Branch-Site Models. We use branch-site models to explore the evolution of the $\psi Est-6$ and *Est-6* clades (Table 3). Two ways of defining background/foreground branches are used. First, we define the branch connecting the two clades as background and all other branches as foreground (as in Fig. 1). Such interpretation does not require a rooted tree. Using this

Table 2. Maximum likelihood (ML) estimates for branch-specific models for the β -esterase genes in seven species of the *D. melanogaster* subgroup

LRT	df	ML estimates under the null	l_0	ML estimates under the alternative	l_1	p value
M0 vs. MR2	1	M0 (one ratio): $\omega = 0.1981$ ($\omega = \omega_{backgr} = \omega_{foregr}$)	-7950.34	MR2 (two ratios): $\omega_{backgr} = 0.09, \omega_{foregr} = 0.24$	-7933.80	$< 10^{-8}$
MR2 vs. MR3	1	MR2 (two ratios): $\omega_{backgr} = 0.09, \omega_{foregr} = 0.24$ ($\omega_{foregr} = \omega_{foregr-P} = \omega_{foregr-6}$)	-7933.80	MR3 (three ratios): $\omega_{backgr} = 0.09, \omega_{foregr-P} = 0.22,$ $\omega_{foregr-6} = 0.26$	-7933.16	0.26

Note. See Table 1 and Fig. 1 for notations.



Model R1: $\omega_{\text{backgr}} = \omega_{\text{foregr}} - \psi_{\text{Est-6}} = \omega_{\text{foregr}} - \text{Est-6}$

Model R2: $\omega_{\text{backgr}} \neq \omega_{\text{foregr}} - \psi_{\text{Est-6}} = \omega_{\text{foregr}} - \text{Est-6}$

Model R3: $\omega_{\text{backgr}} \neq \omega_{\text{foregr}} - \psi_{\text{Est-6}} \neq \omega_{\text{foregr}} - \text{Est-6}$

Fig. 1. Unrooted neighbor-joining tree of the β -esterase genes based on the coding sequence (exon I + exon II). Branch lengths are estimated by ML. Background/foreground and *Est-6*/ ψ *Est-6* partitions are as used in branch model definitions.

assumption two LRTs were performed: MA vs. M1 and MB vs. M3 (with $K = 2$). The LRT comparing models MA and M1 is significant. This is consistent with the results from site-specific models. Approximately 60% of sites have undergone a change in selective pressure in the background branches. Although under MB a small fraction of sites has a high ω ratio ($\omega_2 = 26.04$) in the foreground branches, the LRT comparing MB and M3 is not significant.

Keeping the same definition of background and foreground, we have performed the LRT of MD (case 1) and M3 assuming a rooted tree comparison (Fig. 2A). Model MD fits the data significantly better. In the background (connecting) branch the two classes of sites have ω ratios 0.03 (strong purifying selection) and 0.28 (weak purifying selection); later the functional constraints are relaxed: the two classes of sites evolve with $\omega_1 = 0.28$, then $\omega_2 = 0.50$.

More interesting yet is to compare the evolution patterns of the two clades since the duplication event.

It is convenient to consider the duplication event to be at the root of a tree on the branch connecting the two clades. We now define all branches of the ψ *Est-6* clade as background branches and all branches of the *Est-6* clade as foreground branches (Fig. 2B). The LRT comparing MD (case 2) and M3 for $K = 2$ is not significant, but adding an extra class improves the fit. The LRT comparing models MD (case 2) and M3 for $K = 3$ is significant (at 5%), and the ML estimates indicate that the pattern of evolution is significantly different in the two clades. Two classes of sites evolve with constant ω ratios of $\omega_0 = 0.05$, $\omega_1 = 0.58$, in both the ψ *Est-6* and the *Est-6* clades. The remaining 16% of sites (belonging to the third class) evolve with different ω ratios in each clade: $\omega = \omega_2 = 0.14$ in ψ *Est-6* and $\omega = \omega_3 = 0.70$ in *Est-6*. Accordingly, 16% of the sites are conserved in ψ *Est-6* but evolve under relaxed functional constraints in *Est-6*. This is also consistent with ω ratio estimates under a free ratio model.

Table 3. Maximum likelihood (ML) estimates for branch-site models for the β -esterase genes in seven species of the *D. melanogaster* subgroup

Models	df	ML null estimates	l_0	ML alternative estimates	l_1	p value
M1 vs. MA Unrooted tree	2	M1 (neutral): $p_0 = 0.46, p_1 = 0.54$	-7966.97	MA: $p_0 = 0.25, p_1 = 0.14$ $\omega_2 = 0.22 (p_2 + p_3 = 0.60)$	-7862.21	$< 10^{-45}$
M3 vs. MB Unrooted tree	2	M3 (with $K = 2$): $p_0 = 0.64, p_1 = 0.36$ $\omega_0 = 0.06, \omega_1 = 0.52$	-7857.31	MB: $\omega_0 = 0.06, \omega_1 = 0.53, \omega_2 = 26.04$ $p_0 = 0.64, p_1 = 0.35 (p_2 + p_3 = 0.01)$	-7857.26	0.95
M3 vs. MD (case 1) $K = 2$ Rooted tree	1	M3 (with $K = 2$): $p_0 = 0.64, p_1 = 0.36$ $\omega_0 = 0.06, \omega_1 = 0.52$	-7857.31	MD (case 1) ($K = 2$): $p_1 = 0.55, p_2 = 0.45$ $\omega_0 = 0.03, \omega_1 = 0.28, \omega_2 = 0.50$	-7854.24	0.01
M3 vs. MD (case 2) $K = 3$ Rooted tree	1	M3 (with $K = 3$): $p_0 = 0.33, p_1 = 0.53 (p_2 = 0.14)$ $\omega_0 = 0.00, \omega_1 = 0.22, \omega_2 = 0.84$	-7851.79	MD (case 2) ($K = 3$): $p_0 = 0.59, p_1 = 0.25 (p_2 = 0.16)$ $\omega_0 = 0.05, \omega_1 = 0.58,$ $\omega_2 = 0.14, \omega_3 = 0.70$	-7849.85	< 0.05

Note. See Table 1 and Fig. 2 for notations.

(b) *Est-6* and ψ *Est-6* Dataset (85 Sequences)

Site-Specific Models. The LRTs of neutrality are highly significant for both *Est-6* ($p < 10^{-25}$) and ψ *Est-6* ($p < 10^{-35}$). This means that neither gene evolves neutrally (with $\omega = 1$, on average) and strong selective forces are involved: $\omega = 0.258$ for *Est-6* and $\omega = 0.262$ for ψ *Est-6*. All LRTs are highly significant (Table 4) and suggest the existence of a small fraction of sites (1–3%) under positive selection. For the two genes' combined data, the ML parameter estimates under M3 (discrete) indicate that approximately 49% of sites are highly conserved with $\omega_0 = 0.02$, 48% of sites evolve under purifying selection with $\omega_1 = 0.38$, and the remaining 3% of sites evolve under positive selection with $\omega_2 = 1.86$. For the *Est-6* gene the estimates suggest that 48.1% of sites are highly conserved with $\omega_0 = 0.003$, 50.4% of sites evolve under purifying selection with $\omega_1 = 0.41$, and the remaining 1.5% of sites evolve under positive selection with $\omega_2 = 4.54$. The majority of ψ *Est-6* sites (83.7%) evolve under strong purifying constraint with $\omega_0 = 0.11$, 16% of sites are neutral with $\omega_1 = 1.1$, and 0.3% of sites are under positive selection with $\omega_2 = 8.63$. For both *Est-6* and ψ *Est-6* the ratio values of d_N/d_S are significantly greater than one, even with the Bonferroni correction. LRTs comparing models M1 (neutral) vs. M2 (selection) and M7 (β) vs. M8 ($\beta\&\omega$) are also significant for both separate and combined analyses. All tests support the presence of sites under positive selection.

The sites predicted to be under positive selection using the Bayesian method are shown in Table 5. The few sites inferred to evolve under positive selection are different in the two clades, which might be an indication of functional differences between *Est-6* and ψ *Est-6*. The *Est-6* sites with very high posterior probabilities are (coordinates of Oakeshott et al. 2001) Thr37 → Ile, Asn237 → Asp, Ala247 → Thr, Leu388 → Val, and Ser487 → Ala. The ψ *Est-6* sites with very high posterior probabilities are Thr12 → Ser and Thr460 → Val.

The amino acid replacement Asn237 → Asp is responsible for the F/S allozyme polymorphism of *EST-6* and it is located in the region with an elevated level of silent polymorphism (Balakirev et al. 1999, 2002; Balakirev and Ayala 2003b). We have suggested previously that this site is under balancing selection. The present analysis confirms our previous suggestion. The other *Est-6* sites evolving under positive selection are located in regions of elevated replacement polymorphism (data not shown). The strong signal of positive selection in ψ *Est-6* is surprising but not totally unexpected. The Kelly (1997) and Wall (1999) tests reject the neutral model for ψ *Est-6* (Balakirev and Ayala 2003c, 2004); moreover, there is

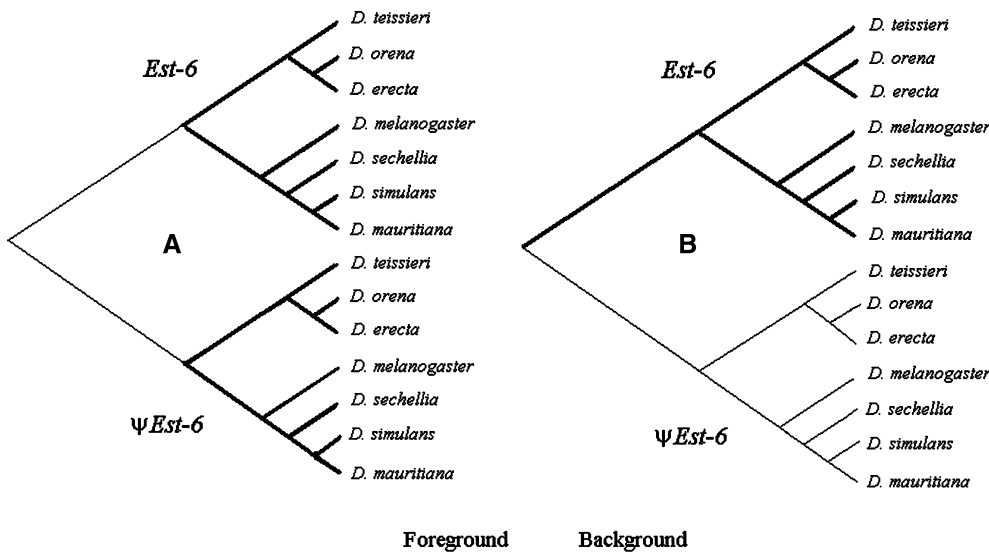


Fig. 2. Rooted maximum-likelihood tree of the β -esterase genes (not to scale). Thin and thick lines are for background and foreground branches, respectively. **A** Model D, case 1. **B** Model D, case 2.

strong haplotype structure and significant linkage disequilibrium within the gene (Balakirev and Ayala 2003c, 2004). The two positively selected sites within $\psi Est-6$ are located at the beginning and at the end of the gene (amino acid positions 12 and 460; coordinates of Oakeshott et al. 2001), where the level of interspecific divergence between *D. melanogaster* and *D. simulans* is significantly decreased (Balakirev and Ayala 2003c, 2004). Healy et al. (1996) have shown that sequences at the beginning of the $\psi Est-6$ transcription unit contain elements that modulate the expression of *Est-6*. Moreover, Brady and Richmond (1992) detected some sequence similarity in this region between $\psi Est-6$ (*D. melanogaster*) and its orthologue *Est-5A* (*D. pseudoobscura*). The obvious decrease in interspecific divergence and presence of sites under positive selection may indicate functionally important regions within $\psi Est-6$, such as sites participating in the regulation of the *Est-6* gene (Healy et al. 1996).

Overall, the majority of the *Est-6* and $\psi Est-6$ sites evolve under strong purifying selection, which implies an existence of strong functional constraint on both genes. The presence of *Est-6* sites evolving under positive selection is in accordance with our previous data on population variability of the gene in *D. melanogaster* (Balakirev et al. 1999, 2002; Balakirev and Ayala 2003b, c) and also with the interspecific data of Oakeshott et al. (2001), who have detected accelerated *Est-6* amino acid sequence change during the evolution of the gene in the *D. melanogaster* subgroup. The existence of positively selected sites in $\psi Est-6$ is an interesting feature of our results, indicating diversifying selective pressure at those sites.

For the population data only (78 strains of *D. melanogaster*) there is a stronger signal of positive

selection (Table 6) than for dataset (b) (*Est-6* [85] and $\psi Est-6$ [85]); also, there are more sites inferred to be under positive selection (data not shown). However, due to the low divergence of sequences at the population level, it is necessary to take the results of site difference with caution since the Bayesian prediction can be inaccurate in this case (Anisimova et al. 2002).

(c) *Est-6* Dataset (Eight Species)

Oakeshott et al. (2001) have detected accelerated *EST-6* sequence change within the *D. melanogaster* subgroup. They attribute this acceleration to either positive directional selection or a relaxation of selective constraint, or a mixture of both. The authors use the Tajima (1993) relative-rate test, which does not allow the resolution of this ambiguity. We seek to clarify this situation by analyzing the *Est-6* gene in the eight species of the *D. melanogaster* subgroup using the site- and branch-specific models.

Site-Specific Models. The LRT comparing models M0 and M3 indicates that 11% sites are under positive selection, $\omega_2 = 1.43$ (Table 7). However, the LRTs of models M1 vs. M2 and M7 vs. M8 do not provide significant evidence of positive selection. Nevertheless, positive selection cannot be ruled out; the absence of strong evidence for positive selection may be due to the small sample size and little divergence.

Branch-Specific Models. Three hypotheses are tested using branch-specific models. First, only the *D. yakuba* is defined to be in the foreground. The two-ratio model yields $\omega_{yak} = 0.23$ for *D. yakuba* and $\omega_{others} = 0.29$ for all the remaining branches,

Table 4. Maximum likelihood (ML) estimates for site-specific models of the β -esterase genes in 78 strains of *D. melanogaster* and seven species of the *D. melanogaster* subgroup

Models	df	ML null estimates	l_0	ML alternative estimates	l_1	P value
<i>Est-6</i> M0 vs. M3	4	M0 (one-ratio): $\omega = 0.2577$	-5044.56	M3 (discrete with $K = 3$): $\omega_0 = 0.003$, $\omega_1 = 0.41$, $\omega_2 = 4.54$ $p_0 = 0.481$, $p_1 = 0.504$, ($p_2 = 0.015$)	-4980.59	$< 10^{-25}$
M1 vs. M2	2	M1 (neutral): $p_0 = 0.65$, $p_1 = 0.35$	-5011.40	M2 (selection): $\omega_2 = 8.46$, ($p_2 = 0.007$) $p_0 = 0.64$, $p_1 = 0.35$	-4996.46	$< 10^{-6}$
M7 vs. M8	2	M7 (B): $p = 0.22$, $q = 0.65$	-4999.14	M8 (B&C): ($p_1 = 0.012$) $\omega = 5.12$ $p_0 = 0.988$, $p = 0.43$, $q = 1.52$	-4980.62	$< 10^{-8}$
ψ <i>Est-6</i> M0 vs. M3	4	M0 (one-ratio): $\omega = 0.2616$	-5700.66	M3 (discrete with $K = 3$): $\omega_0 = 0.10$, $\omega_1 = 1.10$, $\omega_2 = 8.63$ $p_0 = 0.84$, $p_1 = 0.16$, ($p_2 = 0.003$)	-5614.13	$< 10^{-35}$
M1 vs. M2	2	M1 (neutral): $p_0 = 0.64$, $p_1 = 0.36$	-5642.48	M2 (selection): $\omega_2 = 10.09$, ($p_2 = 0.003$) $p_0 = 0.64$, $p_1 = 0.36$	-5629.66	$< 10^{-5}$
M7 vs. M8	2	M7 (B): $p = 0.17$, $q = 0.52$	-5628.99	M8 (B&C): $p_0 = 0.997$, $p = 0.20$, $q = 0.61$ ($p_1 = 0.003$) $\omega = 7.64$	-5615.94	$< 10^{-5}$

Note. See Table 1 for notations.

with log-likelihood ($l = -4628.06$, which does not offer significant improvement over the one-ratio model M0 (Table 8). This means that in our dataset there is no evidence that ω along the whole tree differs significantly from ω in the *D. yakuba*. In a second test, *D. yakuba* and *D. teissieri*, together with the branch leading to their MRCA, are taken as foreground (Fig. 3). Again, the two-ratio model does not provide a significantly better fit than M0, with $l = -4628.23$ and $\omega_{\text{yak+tei}} = 0.284$, $\omega_{\text{others}} = 0.279$. Finally, we test the long branch (Fig. 3) separating *D. yakuba*, *D. teissieri*, *D. erecta*, and *D. orena* from the other species: one ω ratio is assumed for the specified branch and ω is averaged for all other branches. The two-ratio model gives a significantly better fit than M0 (after Bonferroni correction for multiple testing) with $l = -4616.52$; the estimates of selection pressure are $\omega_{\text{longest}} = 0.72$, $\omega_{\text{others}} = 0.21$ (Table 8). The constraints of this model are, then, relaxed by allowing the two clades separated by the “longest” branch to have different ω ratios, and different from ω_{longest} , so that the model becomes a three-ratio model. However, the fit is not significantly better (Table 8). The free-ratio model also does not provide a significantly better fit than the two-ratio model, with $l = -4610.69$ and $P = 0.3$ (estimates of ω ratios for each branch are shown in Fig. 3).

Since we find that the “long” branch separating the *D. yakuba*, *D. teissieri*, *D. erecta*, *D. orena* clade and the *D. melanogaster*, *D. sechellia*, *D. mauritiana*, *D. simulans* clade evolves under significantly higher selective pressure (averaged over sites) than other branches, we further investigate evolution on this branch (foreground) using the branch-site model MB. Model B fits the data significantly better (Table 8). Moreover, the ω ratios suggest that some sites along the foreground branch underwent episodic positive selection. More precisely, according to model MB there are generally two classes of sites. The first class of sites evolves with $\omega_0 = 0.06$ (strong purifying selection): 65% of all sites in the sequence belong to this class and evolve at a constant rate ω_0 ; but 3.6% of all sites change from rate ω_0 to $\omega_2 = 2.29$ (positive selection) in the foreground (Fig. 3). The second class of sites evolves with $\omega_1 = 1.18$ (which could be taken as neutral): 14% of all sites in the sequence belong to this class and evolve at a constant rate ω_1 , but 17.4% of all sites change from ω_1 to $\omega_2 = 2.29$ (positive selection) in the foreground.

Overall, the analysis shows episodic positive selection during the *Est-6* evolution in the *D. melanogaster* subgroup. The signal of positive selection is most obvious when we compare the lineage of *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia* with the lineage of *D. yakuba*, *D. teissieri*, *D. erecta*, and *D. orena* (Fig. 3). This division coincides with a change in the quaternary structure of EST-6, which is

Table 5. Sites under positive selection predicted by site-specific models (with posterior probabilities) for the β -*esterase* genes in 78 strains of *D. melanogaster* and 7 species of the *D. melanogaster* subgroup

Model	<i>Est-6</i>	ψ <i>Est-6</i>
M2	54(T) 1.00* , 253(N) 0.94, 402(V), 0.82, 500(S) 1.00*	29(T) 0.98, 473(T) 1.00*
M3	54(T) 1.00* , 253(N) 1.00* , 263(A) 0.95 , 402 (V) 0.99* , 500(S) 1.00*	29(T) 1.00* , 473(T) 1.00*
M8	54(T) 1.00* , 253(N) 0.99* , 263(A) 0.81, 402(V) 0.96 , 500(S) 1.00*	29(T) 0.99* , 473(T) 1.00*

Note. Statistically significant values are in boldface. The amino acids in parentheses are from *D. melanogaster* reference line ER-S-26F.

homodimeric in *D. erecta*, *D. teissieri*, *D. orena*, and *D. yakuba* but monomeric in *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia* (Oakeshott et al. 1990; Richmond et al. 1990). This change in quaternary structure of the enzyme is moreover associated with changes in the *Est-6* expression pattern. In *D. melanogaster*, *D. simulans*, *D. mauritiana*, and *D. sechellia* the gene is mostly expressed in the male sperm ejaculatory duct and its product (EST-6) is transferred by males to females in the seminal fluid during copulation (Richmond et al. 1980) and affects the female's consequent behavior and mating proclivity (Gromko et al. 1984). However, in *D. erecta*, *D. teissieri*, *D. orena*, and *D. yakuba* species, *Est-6* is mostly expressed in the hemolymph and has no reproduction-related function (Oakeshott et al. 1990; Korochkin 1995). Oakeshott et al. (2001) suggest that changes in three amino acid sites (Ile237 \rightarrow Asn, Gly241 \rightarrow Glu, and Leu242 \rightarrow Ser) cause the dimer-monomer transition of EST-6. This suggestion is based on the data on similar noncovalent subunit binding sites found in the ADH of *D. lebanonensis* (Benach et al. 1998). However Gu's (1999) test predicts with a high posterior probability (0.7894) that the single amino acid change Ile 237 \rightarrow Asn, Thr is responsible for the functional divergence of EST-6 between the lineage of *D. erecta*, *D. teissieri*, *D. orena*, *D. yakuba* and the lineage of *D. melanogaster*, *D. simulans*, *D. mauritiana*, *D. sechellia*. Interestingly, this site is also responsible for the F/S polymorphism in *D. melanogaster* (Asn237 \rightarrow Asp). A single amino acid change is also responsible for the dimerization of the resisting peptide hormone in multiple species (Banerjee and Lazar 2001).

Discussion

We examine the patterns of molecular evolution of the β -*esterase* gene cluster, including *Est-6* and ψ *Est-6*, in seven (plus *D. yakuba* in some analyses) species of the *Drosophila melanogaster* species subgroup. The results provide strong support for variable selection intensity among amino acid sites. There are also significant rate differences between different *D. melanogaster* lineages in the *Est-6* gene phylogeny, indicating accelerated rates of nucleotide substitution, associated with a change of the enzyme

quaternary structure and function. There is statistically significant evidence supporting the hypothesis of positive selection driving the evolution of the β -*esterase* gene cluster. The maximum likelihood analysis identifies a number of sites under diversifying Darwinian selection, in both *Est-6* and ψ *Est-6*, and demonstrates the importance of examining numerous sequences in order to detect positive selection at individual codon (amino acid) sites.

We have found significant heterogeneity in the substitution patterns between the two genes, suggesting that each gene has evolved under different selective constraints. Using maximum likelihood estimates of nonsynonymous/synonymous rate ratios ($\omega = d_N/d_S$), we show that most amino acid sites (98.5%) in *Est-6* are under strong or moderate selective constraint, while a few sites (1.5%) are under significant diversifying selection. For ψ *Est-6*, the distribution of selective constraints is different. Although most amino acid sites (83.7%) evolve under purifying constraint, there is a significant proportion of sites (16%) that evolve neutrally. Interestingly, there are two sites within ψ *Est-6* that evolve under strong positive selection. These sites are located at the beginning and the end of the ψ *Est-6* coding region (amino acid positions 12 and 460 in Oakeshott et al. [2001] coordinates) and coincide with the regions of elevated level of intraspecific variability and decreased interspecific divergence. The presence of positively selected sites within both genes is consistent with our previous results for *D. melanogaster* and provides additional support for our hypothesis concerning the functions of *Est-6* and ψ *Est-6*, namely, that after the duplication event *Est-6* has retained the esterase coding function, while ψ *Est-6* has lost that function but has evolved new ones, and that it may now operate in conjunction with *Est-6* as an intergene (Balakirev and Ayala 2003a, c, d, 2004).

We have characterized the patterns of evolution in the β -*esterase* gene cluster using different approaches, which include molecular population analysis (Balakirev and Ayala 1996, 2003b, c, 2004; Balakirev et al. 1999, 2002, 2003; Ayala et al. 2002), entropy and GC-content analyses (Balakirev et al. 2003, 2005), and codon substitution models that allow for variable nonsynonymous/synonymous rate ratios ($\omega = d_N/d_S$) among sites and between branches (present work). All approaches reveal significantly

Table 6. Maximum likelihood (ML) estimates for site-specific models of the β -esterase genes in 78 strains of *D. melanogaster*

LRT	df	ML estimates under the null	l_0	ML estimates under the alternative	l_1	P value
<i>Est-6</i> M0 vs. M3	4	M0 (one ratio): $\omega = 0.25$	-3007.27	M3 (discrete with $K = 3$): $\omega_0 = 0.04$, $\omega_1 = 3.67$, $\omega_2 = 15.29$ $p_0 = 0.97$, $p_1 = 0.015$, ($p_2 = 0.01$)	-2940.36	$< 10^{-27}$
M1 vs. M2	2	M1 (neutral): $p_0 = 0.90$, $p_1 = 0.10$	-2966.66	M2 (selection): $\omega_2 = 13.21$, ($p_2 = 0.01$) $p_0 = 0.91$, $p_1 = 0.08$	-2940.66	$< 10^{-11}$
M7 vs. M8	2	M7 (B): $p = 0.005$, $q = 0.05$	-2966.66	M8 (B& ω): $p_0 = 0.985$, $p = 0.02$, $q = 0.31$ ($p_1 = 0.015$) $\omega = 12.58$	-2940.68	$< 10^{-11}$
ψ^{Est-6} M0 vs. M3	4	M0 (one ratio): $\omega = 0.32$	-3607.66	M3 (discrete with $K = 3$): $\omega_0 = 0.16$, $\omega_1 = 2.85$, $\omega_2 = 17.13$ $p_0 = 0.89$, $p_1 = 0.10$, ($p_2 = 0.004$)	-3512.45	$< 10^{-39}$
M1 vs. M2	2	M1 (neutral): $p_0 = 0.83$, $p_1 = 0.17$	-3537.67	M2 (selection): $\omega_2 = 5.46$, ($p_2 = 0.04$) $p_0 = 0.84$, $p_1 = 0.12$	-3514.69	$< 10^{-9}$
M7 vs. M8	2	M7 (B): $p = 0.01$, $q = 0.06$	-3542.69	M8 (B& ω): $p_0 = 0.95$, $p = 0.01$, $q = 0.07$ ($p_1 = 0.05$) $\omega = 4.90$	-3514.89	$< 10^{-12}$

Note. See Table 1 for notations.

Table 7. Maximum likelihood estimates for site-specific models of the *Est-6* gene in eight species of the *D. melanogaster* subgroup

Models	df	ML null estimates	l_0	ML alternative estimates	l_1	P value
M0 vs. M3	4	M0 (one ratio): $\omega = 0.2804$	-4628.23	M3 (discrete with $K = 3$): $\omega_0 = 0.00$, $\omega_1 = 0.35$, $\omega_2 = 1.43$ $p_0 = 0.48$, $p_1 = 0.41$, ($p_2 = 0.11$)	-4589.05	$< 10^{-15}$
M1 vs. M2	2	M1 (neutral): $p_0 = 0.64$, $p_1 = 0.36$	-4593.92	M2 (selection): $\omega_2 = 0.07$, ($p_2 = 0.76$) $p_0 = 0.00$, $p_1 = 0.24$	-4589.30	$< 10^{-2}$
M7 vs. M8	2	M7 (B): $p = 0.14$, $q = 0.34$	-4589.95	M8 (B& ω): $p_0 = 0.91$, $p = 0.30$, $q = 1.26$ ($p_1 = 0.09$) $\omega = 1.50$	-4589.07	0.41

Note. See Table 1 and Fig. 1 for notations.

Table 8. Maximum likelihood estimates for branch-specific models of the *Est-6* gene in eight species of the *D. melanogaster* subgroup

Models	df	ML null estimates	l_0	ML alternative estimates	l_1	P value
M0 vs. MR2	1	M0 (one ratio): $\omega = 0.2804$ ($\omega = \omega_{\text{others}} = \omega_{\text{longest}}$)	-4628.23	MR2 (two ratios): $\omega_{\text{others}} = 0.22$, $\omega_{\text{longest}} = 0.72$	-4616.52	$< 10^{-5}$
MR2 vs. MR3	1	MR2 (two ratios): $\omega_{\text{others}} = 0.22$, $\omega_{\text{longest}} = 0.72$ ($\omega_{\text{others}} =$ $\omega_{\text{yak+tei+ore+ere}} = \omega_{\text{sim+mau+sec+mcl}}$)	-4616.52	MR3 (three ratios): $\omega_{\text{longest}} = 0.72$, $\omega_{\text{yak+tei+ore+ere}} = 0.23$, $\omega_{\text{sim+mau+sec+mcl}} = 0.18$	-4616.01	0.31
M3 vs. MB	2	M3 (with $K = 2$): $p_0 = 0.78$, $p_1 = 0.22$ $\omega_0 = 0.08$, $\omega_1 = 1.1$	-4589.21	MB: $\omega_0 = 0.06$, $\omega_1 = 1.18$, $\omega_2 = 2.29$ $p_0 = 0.65$, $p_1 = 0.14$ ($p_2 + p_3 = 0.21$)	-4575.02	$< 10^{-6}$

Note. See Table 1 and Fig. 1 for notations.

different patterns in *Est-6* and $\psi Est-6$. The *Est-6* gene shows a typical pattern of molecular evolution, as revealed in many other functional genes, which involves signals of positive selection; the $\psi Est-6$ gene combines characteristics of evolution as are expected for both functional and nonfunctional genes (Balakirev and Ayala 1996, 2003a, c, d, 2004; Balakirev et al. 2003, 2005). Some of the observations, but not others, are consistent with the hypothesis that $\psi Est-6$ is a pseudogene, a state of affairs that obtains in other putative pseudogenes in *Drosophila* as well as in several other organisms (Balakirev and Ayala 2003a, d).

We have advanced the “intergene” concept to explain the observed evolutionary features of *Est-6* and $\psi Est-6$, as well as numerous other examples of conserved, expressed, and functional pseudogenes (Balakirev and Ayala 2003a, c, d, 2004). This intergene concept proposes that pseudogenes arisen by duplication are likely to interact with their ancestral genes, forming a functional complex (“intergene”), in which the new duplicate fulfills new functions derived from the function of the ancestral gene. We define the intergene as a functionally interacting entity, in which each separate component (gene and pseudogene) cannot successfully accomplish a newly evolved functional role. Pseudogene duplications persist in populations because they participate in the ancestral gene’s regulation and/or enhance its allelic variation. The *Est-6*/ $\psi Est-6$ complex in *D. melanogaster* may exemplify such an intergene. The *Est-6* gene plays the structural role (coding for the *EST-6* enzyme) in the complex, while $\psi Est-6$ enhances genetic variation in the *Est-6* gene and contributes to regulate its expression (Healy et al. 1996). There is strong linkage disequilibrium between *Est-6* and $\psi Est-6$. We suggest that intergenic epistatic selection plays a significant role in the evolution of the β -*esterase* gene cluster, preserving $\psi Est-6$ from degenerative destruction and reflecting functional interactions between *Est-6* and $\psi Est-6$. Other examples of similar intergenes are *Ste* and [*Su(Ste)*] (Hardy et al. 1981; Livak 1984, 1990) in *Drosophila melanogaster*, *nNOS* and *pseudo-NOS* in the mollusk *Lymnaea stagnalis* (Korneev et al. 1999), the *cytokeratin 17* gene and *cytokeratin 17* pseudogene (Trojanovsky and Leube 1994) in humans, and the *Makorin1-Makorin1-p1* pseudogene in mice (Hirotsume et al. 2003), plus immune response and antigenic coding sequences in diverse organisms (Balakirev and Ayala 2003a).

The population dynamics of intergenes is unexplored. The present data suggest that, after the duplication event, the ancestral and derived genes (comprising an “intergene”) evolve mostly under purifying selection. This agrees with the results of Kondrashov et al. (2002) showing that two paralogues produced by a duplication are subject to

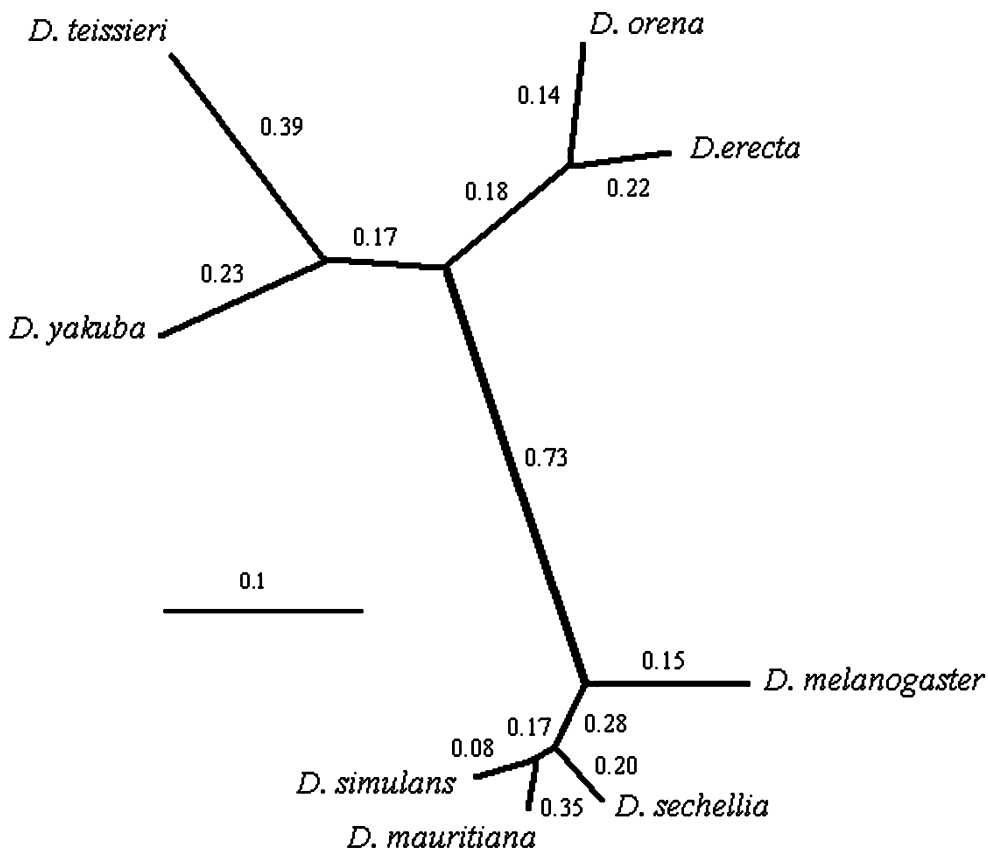


Fig. 3. Maximum-likelihood tree of the *Est-6* sequences in eight species of the *D. melanogaster* subgroup. The numbers at nodes are ω values.

purifying selection and do not experience a phase of neutral evolution. They hypothesize that gene duplications that persist in an evolving lineage are beneficial from the time of their origin due primarily to a protein dosage effect. We show that the persistence of the duplicate may be accounted for also by other explanations, which include regulatory interaction and generation of genetic variation. Kondrashov et al. (2002) suggest that duplications are likely to give rise to new functions at a later phase of their evolution, once a substantial degree of divergence is reached. A large divergence may not be necessary if the derived gene includes regulatory elements that readily interact positively with its homolog ancestor. In this case, an intergene or gene duplication may be immediately advantageous and maintained by selection; any new mutant that provides some additional functional role for the derived gene may further favor the duplicate's evolutionary persistence.

There are 16% of sites within $\psi Est-6$ that evolve neutrally but the neutral fraction is much less in *Est-6*. This difference may account for the higher entropy level and lower GC content of $\psi Est-6$ than of *Est-6* (Balakirev et al. 2003, 2005). The overwhelming fraction of sites evolves under strong negative selection in both *Est-6* and $\psi Est-6$. A small fraction of sites is subject to significant positive selection in both

genes. The strong negative selection prevents the degeneration process of the $\psi Est-6$ sequence, whereas positively selected sites reflect the functional importance of some regions or sites within the gene. Thus, there is a balance between negative and positive selection in intergene evolution that is a consequence of strict functional demands combined with the rise of new variability that enhances the functional possibilities of the β -esterase gene cluster.

Acknowledgments. We thank Elena Balakireva and Iria Blanco Barca for encouragement and help. We are grateful to W.M. Fitch, B. Gaut, R.R. Hudson, and A. Long for detailed and valuable comments. This work is supported by NIH Grant GM42397 to F.J. Ayala and a Medical Research Council (UK) studentship and postdoctoral fellowship from the French Research Ministry to M. Anisimova.

References

- Andersson JO, Andersson SGE (2001) Pseudogenes, junk DNA, and the dynamics of rickettsia genomes. *Mol Biol Evol* 18:829–839
- Anisimova M, Yang Z (2004) Molecular evolution of hepatitis delta virus antigen gene: Recombination or positive selection? *J Mol Evol* 59:815–826
- Anisimova M, Bielawski JP, Yang Z (2001) The accuracy and power of likelihood ratio tests to detect positive selection at amino acid sites. *Mol Biol Evol* 18:1585–1592

- Anisimova M, Bielawski JP, Yang Z (2002) Accuracy and power of Bayes prediction of amino acid sites under positive selection. *Mol Biol Evol* 19:950–958
- Anisimova M, Nielsen R, Yang Z (2003) Effect of recombination on the accuracy of the likelihood method for detecting positive selection at amino acid sites. *Genetics* 164:1229–1236
- Ayala FJ, Balakirev ES, Sáez AG (2002) Genetic polymorphism at two linked loci, *Sod* and *Est-6*, in *Drosophila melanogaster*. *Gene* 300:19–29
- Bailey GS, Poulter RTM, Stockwell PA (1978) Gene duplication in tetraploid fish: model for gene silencing at unlinked duplicated loci. *Proc Natl Acad Sci USA* 75:5575–5579
- Balakirev ES, Ayala FJ (1996) Is esterase-P encoded by a cryptic pseudogene in *Drosophila melanogaster*? *Genetics* 144:1511–1518
- Balakirev ES, Ayala FJ (2003a) Pseudogenes: Are they “junk” or functional DNA? *Annu Rev Genet* 37:123–151
- Balakirev ES, Ayala FJ (2003b) Nucleotide variation of the *Est-6* gene region in natural populations of *Drosophila melanogaster*. *Genetics* 165:1901–1914
- Balakirev ES, Ayala FJ (2003c) Molecular population genetics of the β -*esterase* gene cluster of *Drosophila melanogaster*. *J Genet* 82:115–131
- Balakirev ES, Ayala FJ (2003d) Pseudogenes are not junk DNA. In: Wasser SP (ed) *Evolutionary theory and processes: modern horizons*. Kluwer Academic, Dordrecht, the Netherlands, pp 177–193
- Balakirev ES, Ayala FJ (2004) The β -*esterase* gene cluster of *Drosophila melanogaster*: Is ψ *Est-6* a pseudogene, a functional gene, or both?. *Genetica* 121:165–179
- Balakirev ES, Balakirev EI, Rodríguez-Trelles F, Ayala FJ (1999) Molecular evolution of two linked genes, *Est-6* and *Sod*, in *Drosophila melanogaster*. *Genetics* 153:1357–1369
- Balakirev ES, Balakirev EI, Ayala FJ (2002) Molecular evolution of the *Est-6* gene in *Drosophila melanogaster*: Contrasting patterns of DNA variability in adjacent functional regions. *Gene* 288:167–177
- Balakirev ES, Chechetkin VR, Lobzin VV, Ayala FJ (2003) DNA polymorphism in the β -*esterase* gene cluster of *Drosophila melanogaster*. *Genetics* 164:533–544
- Balakirev ES, Chechetkin VR, Lobzin VV, Ayala FJ (2005) Entropy and GC content in the β -*esterase* gene cluster of *D. melanogaster* subgroup. *Mol Biol Evol* 22:2063–2072
- Banerjee RR, Lazar MA (2001) Dimerization of resistin and resistin-like molecules is determined by a single cysteine. *J Biol Chem* 276:25970–25973
- Benach J, Atrian S, Gonzalez-Duarte R, Ladenstein R (1998) The refined crystal structure of *Drosophila lebanonensis* alcohol dehydrogenase at 1.9 Å resolution. *J Mol Biol* 282:383–399
- Bielawski JP, Yang Z (2001) Positive and negative selection in the DAZ gene family. *Mol Biol Evol* 18:523–529
- Bielawski JP, Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *J Struct Funct Genomics* 3:201–212
- Bielawski JP, Yang Z (2004) A maximum likelihood method for detecting functional divergence at individual codon sites, with application to gene family evolution. *J Mol Evol* 59:121–130
- Brady JP, Richmond RC (1992) An evolutionary model for the duplication and divergence of *esterase* genes in *Drosophila*. *J Mol Evol* 34:506–521
- Collet C, Nielsen KM, Russell RJ, Karl M, Oakeshott JG, Richmond RC (1990) Molecular analysis of duplicated esterase genes in *Drosophila melanogaster*. *Mol Biol Evol* 7:9–28
- Force A, Lynch M, Pickett FB, Amores A, Yan YL, Postlethwait J (1999) Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545
- Gromko MH, Gilbert DF, Richmond RC (1984) Sperm transfer and use in the multiple mating system of *Drosophila*. In: Smith RL (ed) *Sperm competition and the evolution of animal mating systems*. Academic Press, New York, pp 371–426
- Gu X (1999) Statistical methods for testing functional divergence after gene duplication. *Mol Biol Evol* 16:1664–1674
- Hardy RW, Tokuyasu KT, Lindsley DL (1981) Analysis of spermatogenesis in *Drosophila melanogaster* bearing deletions for *Y* chromosome fertility genes. *Chromosoma* 83:593–617
- Harrison PM, Gerstein M (2002) Studying genomes through the aeons: protein families, pseudogenes and proteome evolution. *J Mol Biol* 318:1155–1174
- Harrison PM, Echols N, Gerstein MB (2001) Digging for dead genes: an analysis of the characteristics of the pseudogene population in the *Caenorhabditis elegans* genome. *Nucleic Acids Res* 29:818–830
- Harrison PM, Kumar A, Lan N, Echols N, Snyder M, Gerstein M (2002) A small reservoir of disabled ORFs in the yeast genome and its implications for the dynamics of proteome evolution. *J Mol Biol* 316:409–419
- Harrison PM, Milburn D, Zhang Z, Bertone P, Gerstein M (2003) Identification of pseudogenes in the *Drosophila melanogaster* genome. *Nucleic Acids Res* 31:1033–1037
- Healy MJ, Dumancic MM, Cao A, Oakeshott JG (1996) Localization of sequences regulating ancestral and acquired sites of esterase 6 activity in *Drosophila melanogaster*. *Mol Biol Evol* 13:784–797
- Hileman LC, Baum DA (2003) Why do paralogues persist? Molecular evolution of *CYCLOIDEA* and related floral symmetry genes in Antirrhineae (Veronicaceae). *Mol Biol Evol* 20:591–600
- Hirotsune S, Yoshida N, Chen A, Garrett L, Sugiyama F, Takahashi S, Yagami, Ken-ichi, Wynshaw-Boris A, Yoshiki A (2003) An expressed pseudogene regulates the messenger-RNA stability of its homologous coding gene. *Nature* 423:91–96
- Kelly JK (1997) A test of neutrality based on interlocus associations. *Genetics* 146:1197–1206
- Kimura M, King JL (1979) Fixation of a deleterious allele at one of two “duplicate” loci by mutation pressure and random drift. *Proc Natl Acad Sci USA* 76:2858–2861
- Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplication. *Genome Biol* 3(2):research 8.1–8.9
- Korneev SA, Park J-H, O’Shea M (1999) Neuronal expression of neural nitric oxide synthase (nNOS) protein is suppressed by an antisense RNA transcribed from an NOS pseudogene. *J Neurosci* 19:7711–7720
- Korochkin LI (1995) Cloning, expression, and regulation of tissue-specific genes in *Drosophila*. *Genetika* 31:1029–1042
- Livak KJ (1984) Organization and mapping of a sequence on the *Drosophila melanogaster* *X* and *Y* chromosomes that is transcribed during spermatogenesis. *Genetics* 107:611–634
- Livak KJ (1990) Detailed structure of the *Drosophila melanogaster stellate* genes and their transcripts. *Genetics* 124:303–316
- Lynch M, Conery JS (2000) The evolutionary fate and consequences of duplicate genes. *Science* 290:1151–1155
- Lynch M, Force A (2000) The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473
- Lynch M, O’Hely M, Walsh B, Force A (2001) The probability of preservation of a newly arisen gene duplicate. *Genetics* 159:1789–1804
- Moran NA (2002) Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 108:583–586
- Nadeau JH, Sankoff D (1997) Comparable rates of gene loss and functional divergence after genome duplications early in vertebrate evolution. *Genetics* 147:1259–1266
- Nei M, Roychoudhury AK (1973) Probability of fixation of nonfunctional genes at duplicate loci. *Am Nat* 107:362–372

- Nielsen R, Yang Z (1998) Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936
- Oakeshott JG, Collet C, Phillis R, Nielsen KM, Russell RJ, Chambers GK, Ross V, Richmond RC (1987) Molecular cloning and characterization of esterase 6, a serine hydrolase from *Drosophila*. *Proc Natl Acad Sci USA* 84:3359–3363
- Oakeshott JG, Healy MJ, Game AY (1990) Regulatory evolution of the β -carboxyl esterases in *Drosophila*. In: Barker JSF, Starmer WT, MacIntyre RJ (eds) *Ecological and evolutionary genetics of Drosophila*. Plenum Press, New York, pp 359–387
- Oakeshott JG, van Papenrecht EA, Claudianos C, Morrish BC, Coppin C, Odgers WA (2001) An episode of accelerated amino acid change in *Drosophila* esterase-6 associated with a change in physiological function. *Genetica* 110:231–244
- Ohno S (1970) *Evolution by gene duplication*. Springer-Verlag, Heidelberg, Germany
- Ohta T (1988) Further simulation studies on evolution by gene duplication. *Evolution* 42:375–386
- Richmond RC, Gilbert DG, Sheehan KB, Gromko MH, Butterworth FM (1980) Esterase 6 and reproduction in *Drosophila melanogaster*. *Science* 207:1483–1485
- Richmond RC, Nielsen KM, Brady JP, Snella EM (1990) Physiology, biochemistry and molecular biology of the *Est-6* locus in *Drosophila melanogaster*. In: Barker JSF, Starmer WT, MacIntyre RJ (eds) *Ecological and evolutionary genetics of Drosophila*. Plenum Press, New York, pp 273–292
- Rouquier S, Blancher A, Giorgi D (2000) The olfactory receptor gene repertoire in primates and mouse: evidence for reduction of the functional fraction in primates. *Proc Natl Acad Sci USA* 97:2870–2874
- Seager RD, Ayala FJ (1982) Chromosome interactions in *Drosophila melanogaster*. I. Viability studies. *Genetics* 102:467–483
- Stoltzfus A (1999) On the possibility of constructive neutral evolution. *J Mol Evol* 49:169–181
- Swanson WJ, Nielsen R, Yang Q (2003) Pervasive adaptive evolution in mammalian fertilization proteins. *Mol Biol Evol* 20:18–20
- Tajima F (1993) Simple method for testing the molecular evolutionary clock hypothesis. *Genetics* 135:599–607
- Thompson JD, Higgins DG, Gibson TJ (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
- Torrents D, Suyama M, Zdobnov E, Bork P (2003) A genome-wide survey of human pseudogenes. *Genome Res* 13:2559–2567
- Troyanovsky SM, Leube RE (1994) Activation of the silent human cytokeratin 17 pseudogene-promoter region by cryptic enhancer elements of the cytokeratin 17 gene. *Eur J Biochem* 225:61–69
- Wall JD (1999) Recombination and the power of statistical tests of neutrality. *Genet Res* 74:65–79
- Walsh J (1995) How often do duplicated genes evolve new functions? *Genetics* 139:421–428
- 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, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol* 19:908–917
- Yang Z, Nielsen R, Goldman N, Pedersen AMK (2000) Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449
- Zanotto PM, Kallas EG, de Souza RF, Holmes EC (1999) Genealogical evidence for positive selection in the nef gene of HIV-1. *Genetics* 153:1077–1089
- Zhang J (2003) Evolution by gene duplication: an update. *TREE* 18:292–298