doi:10.1111/j.1420-9101.2008.01672.x

# Selection efficiency and effective population size in *Drosophila* species

N. PETIT & A. BARBADILLA

*Group of Genomics, Bioinformatics and Evolution, Departament de Genètica i Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona, Bellaterra, Spain*

*Keywords:*

evolution;
genetic drift;
neutral theory;
population size;
synonymous polymorphism.

## Abstract

A corollary of the nearly neutral theory of molecular evolution is that the efficiency of natural selection depends on effective population size. In this study, we evaluated the differences in levels of synonymous polymorphism among *Drosophila* species and showed that these differences can be explained by differences in effective population size. The differences can have implications for the molecular evolution of the *Drosophila* species, as is suggested by our results showing that the levels of codon bias and the proportion of adaptive substitutions are both higher in species with higher levels of synonymous polymorphism. Moreover, species with lower synonymous polymorphism have higher levels of nonsynonymous polymorphism and larger content of repetitive sequences in their genomes, suggesting a diminished efficiency of selection in species with smaller effective population size.

## Introduction

The neutral theory of molecular evolution claims that most evolutionary changes at the molecular level are not caused by positive selection, but by a random fixation of selectively neutral variants through the cumulative effect of sampling drift on the continued input of new mutations. According to the neutral theory, the rate of mutant substitutions in evolution is equal to the neutral mutation rate, and this rate is independent of the population size and environmental conditions (Kimura, 1983). The neutral theory assumes that mutations can be classified as neutral or deleterious. However, Ohta (1976) proposed that most 'neutral mutations' are indeed slightly deleterious, rather than strictly neutral. In this nearly neutral theory, the fate of a mutation depends on the relative forces of selection and drift (Ohta, 2002). In recent years, slightly advantageous as well as slightly deleterious amino acid substitutions have been shown to occur in protein evolution (Fay *et al.*, 2001; Eyre-Walker *et al.*, 2002; Ohta, 2002). However, the proportion of substitutions that have been driven by positive selection seems to be substantial both in mammals and *Drosophila* (around 30–70%), showing that the fraction of positively selected mutations is not as small as would be expected by neutral theory (Fay *et al.*, 2002; Smith & Eyre-Walker, 2002; Begun *et al.*, 2007; Macpherson *et al.*, 2007; Shapiro *et al.*, 2007; Studer *et al.*, 2008). Overall, both genetic drift and positive selection are the driving forces of molecular evolution, but the general relative impact of both forces in shaping the observed patterns of nucleotide variability is still under discussion.

The effective population size ($N_e$) determines the degree to which gene frequencies are faithfully transmitted across generations (Wright, 1931) and it is a key factor in the nearly neutral theory of molecular evolution, because the fate of a mutation is determined by the product $N_e s$. When population size is small, genetic drift may outweigh the force of selection, leading to the loss of adaptive genetic variation and the fixation of deleterious alleles (Kimura *et al.*, 1963). Evidence showing that selection efficiency and effective population size are positively correlated is increasing in the last years. Lynch & Conery (2003) have proposed that the changes in genome complexity from prokaryotes to multicellular eukaryotes, including gene number, intron abundance and mobile genetic elements, emerged passively in response to long-term population-size reductions.

*Correspondence:* Natalia Petit, Departament de Genètica i Microbiologia, Facultat de Biociències, Universitat Autònoma de Barcelona, 08193 Bellaterra, Barcelona, Spain.
Tel.: +34 935 812730; fax: +34 935 812387; e-mail: natalia.petit@uab.cat

According to this hypothesis, much of the restructuring of eukaryotic genomes was initiated by nonadaptive processes, which in turn provided novel substrates for the secondary evolution of phenotypic complexity by natural selection (Lynch & Conery, 2003). Vicario *et al.* (2007) found differences in the levels of genomic codon bias, the preferred use of one or more degenerate codons, in *Drosophila* species, and suggested that the differences in selection for codon bias could be explained for differences in effective population size. Bakewell *et al.* (2007) showed that both positive and purifying selection seems to be higher in chimpanzees than in humans. These results are explained by the reduced efficacy of natural selection in humans because of their smaller long-term effective population size. Moreover, a more recent study testing pattern of positive selection among six genomes of mammals found an increased estimate of the rate of evolution in proteins of hominids due to weakened purifying selection, owing to reduced effective population size (Kosiol *et al.*, 2008).

Multispecies data on nucleotide diversity is a very useful source of information that could be used to test the expected relationship between population size and selection at the DNA level. *Drosophila* species have been the focus of genetic and evolutionary studies for decades, and now the availability of the sequences of 12 genomes of these species makes them the focus of evolutionary genomics (Drosophila 12 Genomes Consortium, *et al.*, 2007). The availability of nucleotide polymorphism data in these species ['*Drosophila* Polymorphism Data Base' (DPDB); Casillas *et al.*, 2005, 2007] allow us to test the hypothesis that effective population size and selection efficiency are positively correlated. As synonymous polymorphism is thought to be mainly neutral ($\pi_s \approx \theta \approx 4N_e\mu$; where $\theta$ is the average heterozygosity per site for neutral mutations and $\mu$ is the neutral mutation rate per nucleotide site per generation, Kimura, 1991), the differences detected in the levels of synonymous polymorphism among species could be attributed to differences in effective population size. In this study, we assessed the differences in the levels of synonymous polymorphism among *Drosophila* species, evaluating different potential causes. Then, we tested the hypothesis that the selection efficiency is positively correlated with effective population size by relating the levels of synonymous polymorphism with the levels of codon bias, purifying and adaptive selection. The results show that both, levels of codon bias and proportion of adaptive substitutions, are higher in species with higher levels of synonymous polymorphism. Moreover, species with low synonymous polymorphism have proportionally higher levels of nonsynonymous polymorphism and content of repetitive sequences in their genomes, suggesting a diminished efficiency of the purifying selection in species with smaller population size. Our results provide evidence in favour of the importance of effective population size as a factor shaping the patterns of molecular evolution even within a group of closely related species such as those belonging to the *Drosophila* genus.

## Materials and methods

### Polymorphism data

Polymorphism estimates ($\pi_s$ and $\pi_n$; Nei & Gojobori, 1986) and Tajima's *D*-values (Tajima, 1989b) by gene, were obtained from the DPDB http://dpdb.uab.es; Casillas *et al.*, 2005). DPDB provides polymorphism data by gene and species in the *Drosophila* genus, giving several evaluations of the quality of the estimates and additional information on the genes, such as the coordinates to the *Drosophila melano*gaster genome and the Gene Ontology classification. A link to the source information of the sequences in GenBank and EMBL, and the possibility of reanalysing any gene–specie dataset is also provided in the web page. To get reliable estimates of $\pi_s$, the data was filtered according to the following criteria:

1 Only species with more than five analysed genes and $\pi_s$ estimates based on four or more sequences were considered.
2 Alignments with total gap length or length differences larger than 30% were discarded.
3 Genes from transposable elements (gag, pol, etc.) and pseudogenes were discarded ($N = 9$).

### Datasets

The filtered dataset consisted of 751 polymorphism estimates, belonging to 482 different genes of 15 *Drosophila* species: *D. americana* (dame), *D. arizonae* (dari), *D. kikkawai* (dkik), *D. mauritiana* (dmau), *D. melanogaster* (dmel), *D. miranda* (dmir), *D. mojavensis* (dmoj), *D. persimilis* (dper), *D. pseudoobscura* (dpse), *D. santomea* (dsan), *D. sechellia* (dsec), *D. simulans* (dsim), *D. subobscura* (dub), *D. virilis* (dvir) and *D. yakuba* (dyak); and of four different groups of species (*melonagaster*, *obscura*, *virilis* and *repleta*). The estimates of $\pi_s$ in the dataset were based on alignments of haplotype sequences with an average of 16.22 sequences and 1,181.62 analysed sites for each estimate, with an average and SD of $\pi_s$ values of 0.0165 and 0.0063, respectively. The DPDB accession number and the $\pi_s$ and Tajima's *D*-values by species and genes are listed in supplementary Table S1. African and derived populations of *D. melanogaster* and *D simulans* seem to have different levels of synonymous polymorphism (Andolfatto, 2001; Eyre-Walker, 2002; Nolte & Schlotterer, 2008), therefore, we grouped gene sequences of these species according their origin and recalculated the $\pi_s$ values by gene. Because the data of *D. melanogaster* and *D. simulans* account for nearly 60% of the dataset, the same analyses were also performed on a second dataset (Dataset 2, $N = 175$, see Table S1) which contains a subset of 30 random genes for *D. melanogaster* and 30 random genes for *D. simulans* and the total data from the other species.

To perform paired comparisons among species we used an orthologous sets of genes (Table S2).

### Functional classification of the genes

The genes in the dataset were classified according to their Gene Ontology categories (biological process at level 3), using the web site FatiGO (http://fatigo.bioinfo.cipf.es, Al-Shahrour *et al.*, 2005). The 10 most well-represented categories in the dataset were used to group the genes by their functions: anatomical structural development (15.5%); cell communication (16.5%); cellular metabolic process (53.1%); cellular organization and biogenesis (14.1%); cellular developmental process (15.9%); macromolecule metabolic process (46.8%); multicellular organismal development (26.3%); primary metabolic process (50.4%); regulation of biological process (27.3%) and sexual reproduction (10.9%).

### Orthologous genomic sequences

Genomic coding sequences of *D. yakuba*, *D. melanogaster*, *D. mojavensis* and *D. pseudoobscura* were retrieved from the precomputed alignments of genome sequences in the Vista Genome Browser (http://pipeline.lbl.gov/cgi-bin/gateway2; Couronne *et al.*, 2003).

### Genome estimates

The total percentage of repeat sequences in the genomes (%repeat1) was obtained from the genome database of the University of California in Santa Cruz, UCSC web page as calculated with RepeatMasker (http://genome.ucsc.edu/, Karolchik *et al.*, 2003). Genome size, and repeat coverage (%repeat2) were obtained from Drosophila 12 Genomes Consortium, 2007.

The total genomic codon bias as measured by effective number of codons (ENC); Wright, 1990) was obtained from Vicario *et al.* (2007).

### Synonymous polymorphism differences analysis

The variable $\pi_s$ is not normally distributed; therefore, the $\pi_s$ values were transformed to the natural logarithm, giving a normally distributed dataset which allowed us to perform parametric analyses (ANOVAS and regression analysis).

As different types of genes are included in the 15 analysed species, the datasets by species could contain a biased sample of genes with high codon bias or with high rate of selective sweeps which would decrease the level of synonymous polymorphism in a particular species. Therefore, a two-factor (gene and species) analysis was performed. A *post hoc* Tukey test was performed to determine what species pairs are significantly different in their levels of synonymous polymorphism.

### Differences in selective constraints among genes

To test further the different selective constraints on different genes sampled across the species, a two tailed paired *t*-tests (for orthologous genes in pairs of species, see Table S2), was carried out to test pairwise differences among species. Moreover, the effect of the differences in selective constraints among genes was tested using the functional classification of genes by their Gene Ontology. As each gene can belong to different Gene Ontology categories, we performed 10 unbalanced analyses of variance with the underlying model: $\pi_s$ = species + GO. Where GO is a component that classifies the genes as belonging or not to one of the 10 more represented GO categories in the dataset at the level 3 of biological process.

### Phylogenetic relationships among species

The potential effect of the phylogenetic relationship among the species (Felsenstein 1985) on the differences in the levels of synonymous polymorphism was tested by a fully nested ANOVA, grouping the species in subgenus and groups of species (subgenus: *Sophophora* and *Drosophila* and groups: *melanogaster*, *virilis*, *repleta* or *obscura*).

### Deviations from neutral equilibrium

Differences in the values of the Tajima's *D* statistic were used to test possible differences in deviations from neutral equilibrium. The differences in Tajima's *D* statistic were tested by an unbalanced ANOVA and Tukey tests. Coalescent simulations were performed with intermediate values of recombination using DnaSP (Rozas *et al.*, 2003), giving the mean values of segregating sites, samples and number of sites. Recombination values by gene were estimated using DnaSP (Rozas *et al.*, 2003). As the levels of recombination varies widely among genes (mean ± SD, $r = 71 \pm 116.45$), we performed the coalescent simulations for three values of recombination by gene (10, 50 and 100).

### Selection for preferred codons

The relationship between $\pi_s$ and the levels of selection for preferred codons was tested by correlations and regression analyses in the full dataset and in the Dataset 2. Bootstrapping of the correlations was performed using Resampling software from D. C. Howell (http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html) to test the robustness of the correlation.

### Purifying selection

Differences in levels of purifying selection among species was tested in two ways: (1) estimating the slope of the regression between synonymous and nonsynonymous polymorphism levels within each species and then

regressing the slope against the mean synonymous polymorphism value by species; and (2) performing correlations and regression analyses between the percentage of repeat sequences and $\pi_s$ in the full dataset and in the Dataset 2, and testing the robustness of the correlation by bootstrapping.

### Adaptive selection

The average proportion ($\overline{\alpha}$) and average number of adaptive substitutions ($\bar{a}$) was estimated following eqns 1–3 of Smith & Eyre-Walker (2002), for all species pairs with evidences of significant differences in their values of synonymous polymorphism.

$$a = D_n - Ds\frac{P_n}{P_s} \quad (1)$$

$$\alpha = 1 - \frac{D_s P_n}{D_n P_s} \quad (2)$$

and

$$\overline{\alpha} = 1 - \frac{\overline{D_s}}{\overline{D_n}}\left(\frac{\overline{P_n}}{\overline{P_s + 1}}\right) \quad (3)$$

The numbers of synonymous polymorphic sites ($P_s$), nonsynonymous polymorphic sites ($P_n$), synonymous fixed changes ($D_s$) and nonsynonymous fixed changes ($D_n$) were calculated using DnaSP (Rozas *et al.*, 2003) and the website http://mkt.uab.es (Egea *et al.*, 2008). To enable comparisons of $\overline{\alpha}$ between pairs of species, the gene alignments of each species were aligned with the genomic coding sequences of an outgroup species to the tested lineage (see Tables S7 and S5), using the Muscle alignment program (Edgar, 2004). We excluded the dsim–dsan comparison, as no data for nonsynonymous polymorphism for dsan was available. A potential bias could arise in the estimation of the number of polymorphic sites ($P_n$ and $P_s$) as a result of the differences in the number of sequences being compared between pairs of species. To avoid this bias, the same number of sequences was used in each gene-to-gene comparison. All the alignments were revised by eye and are available from the authors on http://bioinformatica.uab.es/alignmentsPetit&Barbadilla2008JBE/align_alfa.rar. Differences in the number of genes compared in Tables 2 and 5 are either due to the bad quality of alignments or to the lack of orthologue sequences to calculate divergence. Confidence intervals of $\overline{\alpha}$ were obtained by bootstrapping the values of $\alpha$ by gene. The effect of each gene on the estimates of $\overline{\alpha}$ based on 10 or fewer genes were evaluated by resampling without replacement, calculating the $\overline{\alpha}$ in $N$ datasets (where $N$ is the number of genes), while removing a different gene each time. The effect of low frequency polymorphism on $\overline{\alpha}$ was tested for the dmel–dsim comparison using the mkt website http://mkt.uab.es (Egea *et al.*, 2008).

All statistical analyses were made using the SAS 9.1 statistical package (SAS Institute Inc., Cary, NC, USA).

## Results

### Differences in the levels of synonymous polymorphism among species

The two-factor ANOVA testing the differences in the levels of synonymous polymorphism among *Drosophila* species showed significant differences only for the species factor. Neither the gene factor nor the interaction between species and genes were significant (Table 1). As the data of *D. melanogaster* and *D. simulans* account for 60% of the dataset, the same analysis was performed with a second dataset (Dataset 2; $N$ = 175; see methods). The analysis showed similar result (Table 1). Figure 1 graphs the different mean $\pi_s$ values per species.

To evaluate which species differed in their $\pi_s$ values, we performed a *post hoc* pairwise Tukey test for pairs of species. Because single gene effects can remain undetected among the 482 different genes analysed by the ANOVA, we also performed a paired *t*-test among the orthologous genes of pairs of species (Table 2). The results of the Tukey test showed that the pairs of species dmel–dsim, dsim–dmir, dpse–dmir, dyak–dsim, dmir–dmel, dsim–dsan, dmir–dari, dmir–dame, dmir–dkik and dkik–dsan are significantly different in their $\pi_s$ levels. Paired *t*-test were coincident for several pairs of species with orthologous genes data (Table 2). For African and derived populations of *D. melanogaster* and *D. simulans*, $\pi_s$ was, on average, significantly larger for *D. simulans* than *D. melanogaster* in non-African populations of both species, but the differences were borderline significant in African populations (Table 2).

Different selective constraints in diverse functional categories of genes of *Drosophila* species could affect the $\pi_s$ levels, by mean of selection acting on linked nonsynonymous sites, (Begun & Aquadro, 1992; Haddrill *et al.*, 2007; McPherson *et al.*, 2007). We further investigated the effect of the differences in the genes sampled across the analysed species on the levels of $\pi_s$, by grouping the genes by their functional categories. We found a small effect of gene function on the synonymous polymorphism for genes belonging to categories related with

**Table 1** Results of the unbalanced analysis of variance of two factors testing the interaction between genes and species factors in two datasets. The dependent variable is the natural logarithm of $\pi_s$.

| | All data | | | Dataset 2 | | |
|---|---|---|---|---|---|---|
| Source | d.f. | *F* | *Pr > F* | d.f. | *F* | *Pr > F* |
| Species | 14 | 6.13 | 0.0049 | 14 | 29.47 | 0.0333 |
| Gene | 452 | 1.18 | 0.4265 | 190 | 8.70 | 0.1086 |
| Species x gene | 205 | 0.69 | 0.8343 | 91 | 5.04 | 0.1797 |

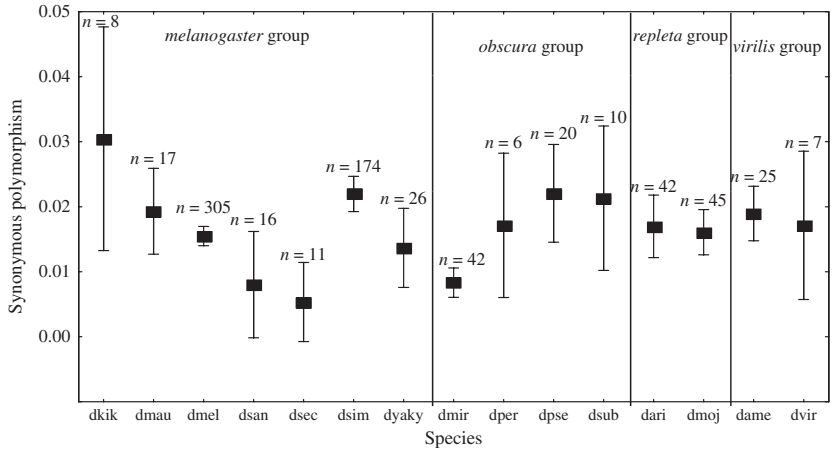


**Fig. 1** Mean $\pi_s$ values and confidence intervals at 95% for the 15 studied species of *Drosophila*.

**Table 2** Differences of the values of $\pi_s$ between pairs of species evaluated by paired *t*-test for orthologous genes. Significant differences are bold faced.

| sp1 | sp2 | *N* genes | Mean $\pi_s$ sp1 | Mean $\pi_s$ sp2 | *P* (*t*-test) |
|---|---|---|---|---|---|
| dame | dsim | 6 | 0.018 | 0.025 | 0.474 |
| dame | dvir | 6 | 0.015 | 0.009 | 0.123 |
| dari | dmoj | 42 | 0.017 | 0.016 | 0.796 |
| NA_dmel | NA_dsim | 91† | **0.013** | **0.020** | **0.00003** |
| A_dmel | A_dsim | 23‡ | 0.015 | 0.032 | 0.063 |
| dmel | dsec | 11 | **0.017** | **0.006** | **0.002** |
| dmel | dpse | 9 | **0.012** | **0.020** | **0.021** |
| dmel | dmir | 10 | 0.011 | 0.006 | 0.083* |
| dmel | dkik | 10 | 0.015 | 0.025 | 0.105 |
| dmel | dame | 6 | 0.014 | 0.019 | 0.061 |
| dmel | dmau | 16 | 0.019 | 0.019 | 0.790 |
| dmel | dyak | 10 | 0.012 | 0.010 | 0.399 |
| dmir | dper | 5 | 0.008 | 0.017 | 0.128 |
| dmir | dpse | 9 | 0.009 | 0.015 | 0.066* |
| dper | dpse | 5 | 0.013 | 0.018 | 0.190 |
| dpse | dsim | 10 | 0.021 | 0.027 | 0.522 |
| dsec | dmau | 11 | **0.006** | **0.021** | **0.032** |
| dsim | dsec | 11 | **0.026** | **0.006** | **0.007** |
| dsim | dmir | 10 | **0.023** | **0.006** | **0.006** |
| dsim | dsan | 5 | 0.010 | 0.025 | 0.109* |
| dsim | dyak | 9 | **0.021** | **0.008** | **0.045** |

*Data from non-African populations of these two species.
†Data from African populations of these two species.
‡Species with means $\pi_s$ levels significantly different by Tukey test.

metabolic process (more constrained than the rest of the genes) and genes related with sexual reproduction (less constrained than the rest of the genes). The results of these analyses are presented in Tables S3 and S4. The mean percentage of the variance explained by the differences between species was 14.38%, and that explained by functional differences was 2.6%. These results are consistent with the previous ANOVA and paired *t*-test analysis for orthologous genes, indicating that the $\pi_s$ differences among species are not due to differently sampled genes within each species.

The two following subsections address other two possible reasons that could explain the differences in $\pi_s$ among species in the above analysis.

*Phylogenetic nonindependence among species*

A significant correlation between variables across species can arise just because of their shared phylogenetic history, even if the variables evolve independently (Felsenstein, 1985). Species belonging to different groups of species could differ in aspects of their life-history, such as generation time. Moreover, recombination rate seems to be different among groups of *Drosophila* species (Caceres *et al.*, 1999; Casillas *et al.*, 2007), which could affect the levels of synonymous polymorphism (Begun & Aquadro, 1992). We performed a fully nested ANOVA to test whether the differences in the levels of synonymous polymorphism could be explained by differences between subgenus or among groups of species. Differences between subgenus explained nothing of levels of synonymous polymorphism (Table 3), whereas differences among species groups explained a very low percentage of the total variance (1.42%, Table 3), and differences among species within groups explained a significant percentage (14.2%; Table 3). This indicates that the differences in the levels of synonymous polymorphism can be attributed mainly to within group differences among species.

**Table 3** Result of the fully nested analysis of variance grouping the species by their phylogenetic relationship (subgenus and groups). The dependent variable is the natural logarithm of $\pi_s$.

| Source | d.f. | *F* | *P* | Var Comp | % of Total |
|---|---|---|---|---|---|
| Subgenus | 1 | 0.64 | 0.507042 | −0.040* | 0 |
| Group | 2 | 1.05 | 0.383304 | 0.02 | 1.48 |
| Species | 11 | 6.30 | **6.70E$^{-10}$** | 0.14 | 14.2 |
| Error | 604 | | | 0.83 | 84.31 |
| Total | 618 | | | 0.98 | |

*Value of the component of variance is negative, and the% of the total is estimated to be zero.

*Deviations from the neutral equilibrium*

Deviation from the neutral equilibrium could be produced by selective events, such as selective sweeps, or demographic events, such as population bottlenecks (Tajima, 1989a, b). Demographic events would generate an overall deviation across the whole genome (Fay *et al.*, 2002), but these differences in $\pi_s$ are likely due to recent demographic events, rather than to long-term effective population size. The *D* statistics of the Tajima test (Tajima, 1989b) measures deviations from the neutral model. Estimated mean values of *D* are shown in Table S5. These *D*-values by species are within the expected values of coalescent simulations with intermediate recombination rate by genes under a neutral infinite-sites model assuming constant population size (Table S5). Combined results of the ANOVA and Tukey test (see Materials and methods) showed that mean Tajima's *D* is significantly higher in *D. simulans* than in *D. yakuba*. The detailed analyses are presented in supplementary Results. Overall, the analyses suggest that the bulk of the differences in $\pi_s$ are not explained by short time reductions in effective population size of the species.

## Testing the hypothesis that selection efficiency is positively correlated with effective population size

The analysis of the previous section indicates that the differences in $\pi_s$ observed in the dataset are mainly explained by differences among species, which could be a reflection of differences in effective population size. In the next sections we will test the hypothesis that selection efficiency and effective population size are positively correlated, assuming that differences in synonymous polymorphism are representative of differences in effective population size ($N_e$). Differences in selection efficiency were tested by differences in: (1) selection for preferred codons, (2) purifying selection and (3) adaptive selection.

*Selection for preferred codons in Drosophila species*

To test the hypothesis that the differences in the levels of codon bias are due to differences in the effective population sizes, we performed analyses of correlation and regression between $\pi_s$ values by gene and species and average genomic ENC estimates. ENC (Wright, 1990) measures the deviation from the equal use of codons, taking higher values when levels of codon bias are lower. We found that $\pi_s$ is negatively correlated with ENC ($r_{Spearman} = -0.149$, $P = 0.0003$, $N = 594$; Fig. 2a). The regression, as measured by the *F* statistic of the regression (Sokal & Rohlf, 1995) was significant ($r^2 = 0.037$, $F_{1,538} = 20.7$; $P < 0.0001$; $n = 540$). This result could be biased by the unequal number of $\pi_s$ values among different species (Fig. 2a), mainly by the high number of gene from *D. melanogaster* and *D. simulans*. We investigated this bias analysing the Dataset 2. The results of the correlation in this dataset was $r_{Spearman} = -0.151$, $P = 0.046$, with confidence intervals at 95% of −0.273 to −0.024 (obtained by bootstrapping of the correlation). Furthermore, although not significant, the correlation between mean values of $\pi_s$ and ENC by species shows the same tendency (Fig. 2b). ENC could be biased by the base composition (Heger & Ponting, 2007), however, Vicario *et al.* (2007) found a high correlation (lower than −0.8) between ENC and CIA (Codon Adaptation Index), which is based in the use of a set of preferred codons and is not biased by the base composition. This positive association between codon bias and $\pi_s$, is indicative of that the selection for preferred codons depends on population size, which agrees with Vicario *et al.*'s (2007) hypothesis that differences in codon bias among *Drosophila* species could be explained by differences in effective population size.

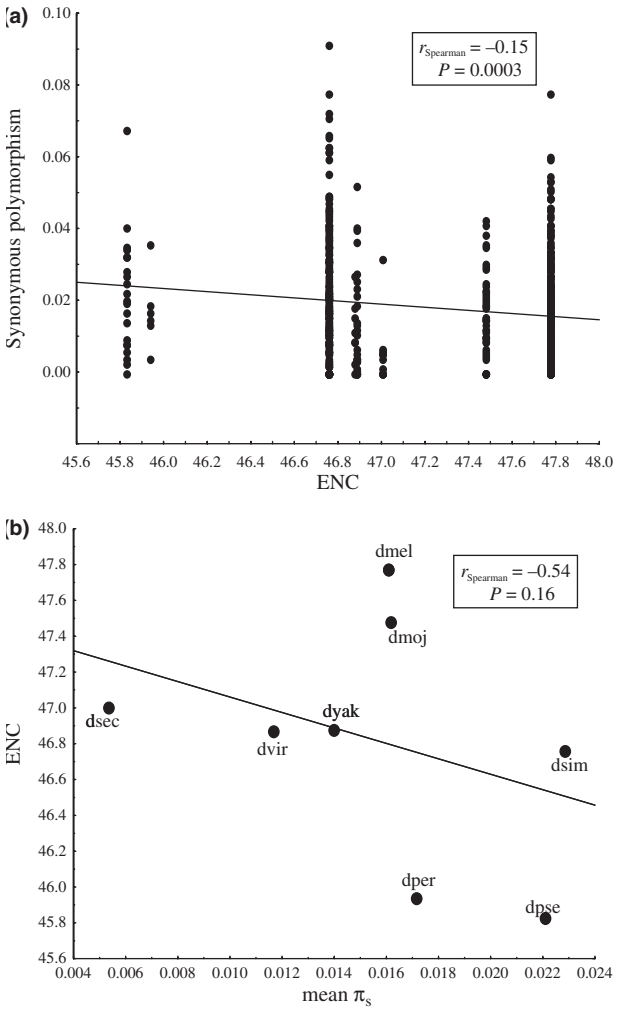


**Fig. 2** Scatterplot of the correlation between codon bias [whole-genome effective number of codons (ENC) and synonymous polymorphim $\pi_s$.

*Purifying selection*

We analysed the differences in purifying selection among *Drosophila* species in two ways: (1) analysing the depen-

dence between synonymous and nonsynonymous polymorphism and (2) analysing the differences in the content of repetitive sequences in the genomes.

For a given gene, the levels of synonymous and nonsynonymous polymorphism will be equal if both kinds of changes are neutral. Inversely, if nonsynonymous changes undergo purifying selection, the dependence between synonymous and nonsynonymous polymorphism will be low or null. We have calculated the slope of the regression between synonymous and nonsynonymous polymorphism for the different genes of each species as a measure of constraint on nonsynonymous polymorphism in a species (a slope = 1 implies absence of constrains compared with synonymous changes). The regression analyses between $\pi_n$ and $\pi_s$ by species are shown in the supplementary Table S6. Each slope was correlated with the mean level of synonymous polymorphism of its corresponding species for all the species. This correlation was negative and significant ($r_{pearson} = -0.652$, $r^2 = 0.43$, $P = 0.008$, $n = 15$, Fig. 3), indicating that higher levels of synonymous polymorphism (or larger effective population size) imply more constraints (low slope values) on nonsynonymous variation. Several correlations between $\pi_n$ and $\pi_s$ by species were not significant (Table S6). However, without the data of the nonsignificant slopes, the correlation between the slopes and mean $\pi_s$ was still significant ($r_{pearson} = -0.78$, $P = 0.036$, $n = 7$).

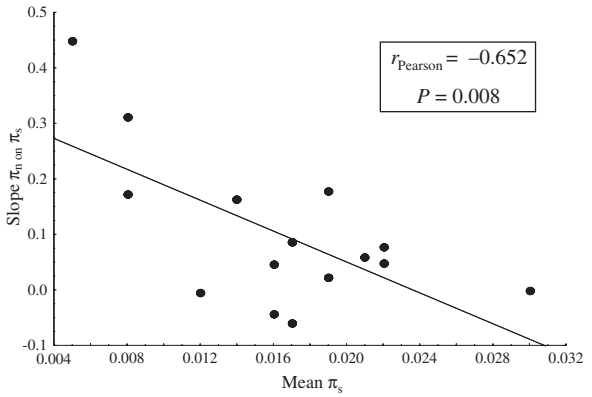


**Fig. 3** Scatterplot of the relationships between mean $\pi_s$ and the slope of the regression analysis between the dependent variable $\pi_n$ and independent variable $\pi_s$ calculated for each species (Table S6).

Using recently published genome data from the 12 species of *Drosophila*, we tested Lynch & Conery's (2003) hypothesis that effective population size affects the rate of genome evolution within a single genus. Correlations between percentage of repetitive sequences in the genome and $\pi_s$ were negative and significant (Table 4, Fig. 4a–d). The correlations between the mean values of $\pi_s$ and two measures of percentage of repetitive sequences in the genomes showed the same negative tendency (Fig. 4c,d). The correlation between the mean values of $\pi_s$ and %repeat1 (see Materials and methods) by species was significant ($r_{Spearman} = -0.738$, $P = 0.036$, $N = 8$; Fig. 4c). Moreover, the correlation between mean values of $\pi_s$ and %repeat2 (see Materials and methods) by species was significant when data from *D. sechellia* was excluded, which seem be an outlier ($r_{Spearman} = -0.892$, $P = 0.0068$, $N = 7$; Fig. 4d). The two measures of percentage of repetitive sequences in the genomes considered here are different each other. This variation is noticeable mainly in dmel, dsec and dvir (Fig. 4a–d). The percentage of repeats calculated based on the number of the base pairs that are repeat as determined by RepeatMasker (%repeat1; see Materials and methods) seems to be underestimated for dvir and overestimated for dmel and dsec. It is possibly due to RepeatMasker database being enriched for transposable elements of *D. melanogater*. Similar results were obtained analysing the Datastet 2 (Table 4).

### *Positive selection*

We performed an exploratory analysis to find differences in the proportion of adaptive substitutions for orthologous genes between pairs of species with evidence of differences in the levels of $\pi_s$ detected by the Tukey

**Table 4** Correlation and regression analyses between $\pi_s$ and percentage of repetitive sequences in the genomes. In the regression analysis, the independent variable was the natural logarithm of $\pi_s$ ($N = 540$), and significance was calculated based on the $F$ statistic of the ANOVA of the regression. Significant values are bold faced. The confidence intervals at 95% were calculated by bootstrapping of the correlations.

| | All data | | | Dataset 2 | | |
|---|---|---|---|---|---|---|
| | $r_{Spearman}$ ($P$) $N = 594$ | $r^2$ | $F_{1,538}$ ($P$) | $r_{Spearman}$ (CI 95%) ($P$) $N = 175$ | $r^2$ | $F_{1,172}$ ($P$) |
| %repeat1† | **−0.175 (< 0.001)** | **0.059** | **33.85 (< 0.001)** | **−0.239 (−0.338 to −0.061) (0.001)** | **0.048** | **8.83 (0.003)** |
| %repeat2‡ | **−0.154 (< 0.001)** | **0.025** | **13.91 (0.001)** | **−0.148 (−0.29 to −0.031) (0.049)** | 0.016 | 2.49 (0.061*) |

*The regression is significant without data of dsec ($F_{(1,147)} = 4.58$; $P = 0.033$; see Fig. 4d)

†Percentage of base pairs in the total genome that are repeats, as detected by repeatmasker (UCSC, Karolchik *et al.*, 2003).

‡Repeat coverage in percentages, calculated as the fraction of scaffolds > 200 kb covered by repeats (Drosophila 12 Genomes Consortium, 2007).

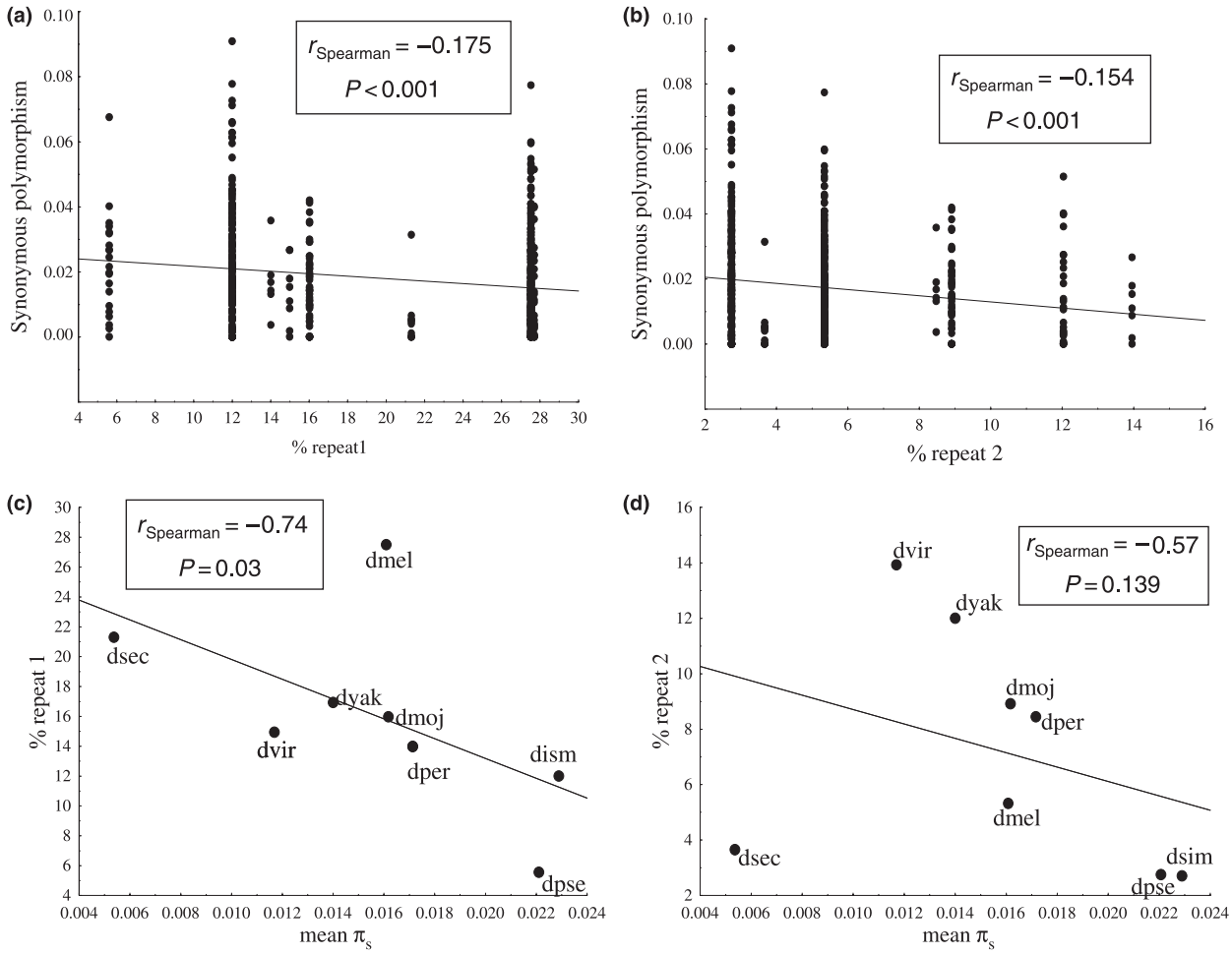


**Fig. 4** Scatterplot of the relationships between $\pi_s$ (a and b), mean $\pi_s$ (c and d) and the percentage of repetitive sequences estimated from genomes of *Drosophila* species. The (a) and (b) percentage of the base pairs in the total genome that are repeats, as detected by RepeatMasker in the web site UCSC (Karolchik *et al.*, 2003). The (c) and (d) repeat coverage in percentages, calculated as the fraction of scaffolds > 200 kb covered by repeats (Drosophila Genomes Consortium 2007).

and/or paired *t*-test (Table 2). We calculated an average $\alpha$ ($\overline{\alpha}$) and an average number of adaptive substitutions ($\bar{a}$) following Smith and Eyre-Walker (Smith & Eyre-Walker, 2002; see Materials and methods). The calculated $\overline{\alpha}$ values are presented in the Table 5. In all comparisons the results showed that species with significantly higher levels of synonymous polymorphism have higher proportion of adaptive substitutions in orthologous genes, except in the comparison dmir–dmel (Tables 2 and 5). All estimated $\overline{\alpha}$ are within the confidence intervals estimated by bootstrapping for the values of $\alpha$ by gene (eqn 2 of Smith & Eyre-Walker, 2002; see Material and methods). The confidence intervals were very wide in many of the cases where the number of genes analysed is low (Tables S7 and S8). Thus, the estimated $\overline{\alpha}$ are representative of the genes analysed, but not of the species. Moreover, the highest estimates of $\overline{\alpha}$ in the comparisons between *melanogaster* species group and *obscura* species group could be biased by the long divergence time from the outgroup, dmoj (Table 5). The differences in $\overline{\alpha}$ between pairs of species are not greatly affected by single genes (Table S8) except in the comparison of dmel–dmir. In this dataset, the $\overline{\alpha}$ values of dmel varied widely (from 0.03 to 0.98), depending on the genes included (Table S8). The differences in the values of average number of adaptive substitutions ($\overline{a} = \overline{\alpha} * D_n$) are significant in all the cases, except in the comparisons dmel-dpse and dmel-dmir (Table 5). Thus, species with higher levels of synonymous polymorphism seem to exhibit larger amounts of adaptive substitutions.

The differences found in $\overline{\alpha}$ between *D. melanogaster* and *D. simulans* might be affected by the recent demographic histories of these two species (Bierne & Eyre-Walker, 2004; Welch, 2006), as changes in population size might affect the proportion of mutations that are effectively neutral (McDonald & Kreitman, 1991). To test this, we

**Table 5** Differences in the proportion of adaptive substitutions among species of *Drosophila* with evidences of differences in the levels of $\pi_s$ from Table 2. $\chi^2$ test calculated using the number of adaptive substitutions $\bar{a}$, calculated as $\bar{\alpha} * D_n$ following Smith & Eyre-Walker (2002).

| Species group | sp1–sp2 | *N* genes | $\bar{\alpha}^{a\dagger}$ sp1–sp2 | $D_n$‡ sp1–sp2 | ā† sp1–sp2 | $\chi^2$ | *P* | Species div |
|---|---|---|---|---|---|---|---|---|
| *melanogaster* | dmel–dsec | 10 | 0.56–0.25 | 277–345 | 155–86 | 42.27 | **< 0.00001** | dyak |
| | dsim–dsec | 10 | 0.55–0.25 | 431–345 | 237–86 | 76.87 | **< 0.00001** | dyak |
| | dmau–dsec | 10 | 0.47–0.25 | 342–345 | 161–86 | 25.98 | **< 0.00001** | dyak |
| | dmel–dsim | 88 | 0.30–0.43 | 2858–3202 | 857–1377 | 98.44 | **< 0.00001** | dyak |
| | dyak–dsim | 8 | 0.20–0.35 | 346–283 | 69–99 | 27.98 | **< 0.00001** | dpse |
| *obscura* | dpse–dmir | 8 | 0.54–0.30 | 95–204 | 51–61 | 16.02 | **0.00006** | dmel |
| | | (6*) | (0.40*–0.04*) | (84*–44*) | (34–2*) | | | |
| *obscura vs melanogaster* | dmel–dpse | 5 | 0.76–0.85 | 99–103 | 75–88 | 2.92 | 0.087 | dmoj |
| | dsim–dmir | 8 | 0.87–0.37 | 284–237 | 247–89 | 144.11 | **< 0.00001** | dmoj |
| | dmel–dmir | 6 | 0.58–0.62 | 85–100 | 49–62 | 0.351 | 0.553 | dmoj |

*Genes from the Neo X chromosome were excluded (see supporting Information)
†$\bar{\alpha}$ and $\bar{a}$ calculated following Smith and Eyre-Walker (Smith & Eyre-Walker, 2002); see Materials and methods).
‡The divergence values were calculated from a species external to the tested lineage.

calculated the $\bar{\alpha}$ for 21 orthologous genes sampled from African populations of *D. melanogaster* and *D. simulans*. The $\bar{\alpha}$ values ($\bar{\alpha}$A_dmel = 0.33; $\bar{\alpha}$A_dsim = 0.53) were similar to that obtained for non-African populations of both species (Table 5). This result indicates that demographic histories are not the main factor influencing $\bar{\alpha}$ in these two species. Moreover, Fay *et al.* (2002) found that $\bar{\alpha}$ is affected by the presence of low frequency variants in *D. melanogaster*. We did not find an effect of low frequency variants (< 5%) in the estimations of $\bar{\alpha}$ for dmel and dsim.

## Discussion

### Synonymous polymorphism levels are different between *Drosophila* species

We found significant differences in the levels of synonymous polymorphism among *Drosophila* species. For those species whose polymorphism had already been studied, our estimates of synonymous polymorphism were quite similar to those previously reported (Moriyama & Powell, 1996; Kliman *et al.*, 2000; Bachtrog & Andolfatto, 2006; Begun *et al.*, 2007; Shapiro *et al.*, 2007). These differences among species are not explained by differences in selective constraint of the sampled genes, or phylogenetic relationships among species. Our analysis of the mean values of Tajima's *D* suggests that the bulk of the differences found in the levels of synonymous polymorphism are not caused by recent demographic events.

Differences in the effective population size among *Drosophila* species have been proposed repeatedly in numerous studies. Aquadro (1992) and Akashi (1996) proposed differences in effective population size between *D. melanogaster* and *D. simulans*. However, whether these differences are due to recent $N_e$ vs. $N_e$ in the distant past of these two species has been widely disputed (Capy & Gibert, 2004; Mousset & Derome, 2004; Nolte & Schlotterer, 2008). *Drosophila sechellia* is a species endemic to an island, and a small effective population size had been proposed for this species (Kliman *et al.*, 2000). Moreover, differences in effective population size have been proposed for the sister species *D. miranda* and *D. pseudoobscura* (Yi *et al.*, 2003). Finally, a recent study of genome wide levels of codon bias among *Drosophila* species suggests that the differences found in levels of codon bias could be attributed to differences in effective population size (Vicario *et al.*, 2007). Our results support the existence of true differences in the levels of synonymous polymorphism among *Drosophila* species, which are likely differences in effective population size.

### Selection efficiency depends on effective population size

Codon bias has been widely detected in *Drosophila* species (Akashi, 1995, 1997; McVean & Vieira, 2001; Heger & Ponting, 2007; Singh *et al.*, 2007; Vicario *et al.*, 2007). The positive correlation found between synonymous polymorphism and the whole-genome codon bias estimates of *Drosophila* species can be interpreted in terms of nearly neutral theory: selection for preferred codons is more effective in species with larger effective population size.

Levels of nonsynonymous polymorphism are constrained by purifying selection. Thus, the nonsynonymous mutations that are segregating in the populations are possibly nearly deleterious (Ohta, 1976; Fay *et al.*, 2001; Eyre-Walker *et al.*, 2002). The slope of the regression between $\pi_n$ and $\pi_s$ take higher values in species with small effective population size (lower levels of mean $\pi_s$), indicating that a higher proportion of nonsynonymous polymorphism seem to behave as neutral in these species. Thus, the efficiency of purifying selection is higher in

species with larger effective population size in agreement with the expectations of the nearly neutral theory.

Lynch & Conery (2003) showed that many features of complex genomes could be initiated as a nonadaptive process due to reductions in effective population size. We have tested their hypothesis using the data from the sequenced genomes of *Drosophila* and our estimates of $4N_es$ ($\pi_s$). We found that differences in the percentage of repetitive sequences, which is indicative of the content of transposable elements in the genomes, could be explained by differences in effective population size in *Drosophila* species. According to Lynch & Conery's (2003) interpretation, in species with low effective population size, selection is ineffective against mildly deleterious insertions, which increase genome size. Bosco *et al.* (2007) showed that differences in genome size among species are mainly due to differences in repetitive DNA satellite contained in the heterochromatin. The two estimates of percentage of repetitive sequences used in this study are based on the euchromatic portion of the genomes and may be underestimates. Thus, our results agree with the hypothesis of Lynch & Conery (2003) and demonstrate that genetic drift could play a crucial role in genome evolution of *Drosophila* species. Genome size seems to increase by insertions of sequences that are weakely deleterious, which are effectively eliminated by natural selection in species with large effective population size, but they may fix in organism with low effective population size (Lynch & Conery, 2003; Yi & Streelman, 2005).

According with the nearly neutral theory, we expect to find more evidence of positive selection in a species with larger $N_e$. Bakewell *et al.* (2007) found evidence that this is true when analysing the genome of chimpanzees and humans. We find that this is also true in *Drosophila*. Our results are based on comparisons of a small number of genes and must be considered suggestive rather than conclusive. However, for the better represented species in our study (*D. melanogater* and *D. simulans*), our estimates of $\bar{\alpha}$ from non-African populations are similar to those obtained by other authors using more data and more sophisticated methods (Bierne & Eyre-Walker, 2004; Welch, 2006; Begun *et al.*, 2007; Shapiro *et al.*, 2007). Thus, our results support the hypothesis that species with larger effective population size undergo higher levels of adaptive selection.

## Conclusions

This study presents a multispecies analysis of levels of synonymous polymorphism among *Drosophila* species. The whole evidence suggests that effective population size is the main explanatory factor of levels of polymorphism. This conclusion has implications on the molecular evolution of *Drosophila* species, as is supported by the analyses testing differences in selection efficiency: (1) differences in whole-genome estimates of codon bias among *Drosophila* species are positively correlated with differences in levels of synonymous polymorphism, (2) the efficiency of purifying selection is higher in species with higher effective population size; and the putative nonadaptive fixation of sequences in the genomes is negatively correlated with levels of synonymous polymorphism of the species, and (3) Species with lower levels of synonymous polymorphism seem to have smaller proportions of adaptive substitutions in orthologous genes.

## Acknowledgments

The authors thank Josefa Gonzalez, Ruth Hershberg and Raquel Egea for corrections and comments on this manuscript, and Dmitri Petrov and two anonymous reviewers for their useful comments about this study. This work was funded by the Ministerio de Educaciòn y Ciencia (Grant BFU-2006-08640). N.P. was supported by a PIF grant from the Departament de Genètica i Microbiologia of the Universitat Autònoma de Barcelona.

## References

Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at ''silent'' sites in Drosophila DNA. *Genetics* **139**: 1067–1076.

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

Akashi, H. 1997. Codon bias evolution in *Drosophila*. Population genetics of mutation-selection drift. *Gene* **205**: 269–278.

Al-Shahrour, F., Minguez, P., Vaquerizas, J.M., Conde, L. & Dopazo, J. 2005. BABELOMICS: a suite of web tools for functional annotation and analysis of groups of genes in high-throughput experiments. *Nucleic Acids Res.* **33**: W460–W464.

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

Aquadro, C.F. 1992. Why is the genome variable? Insights from *Drosophila*. *Trends Genet.* **8**: 355–362.

Bachtrog, D. & Andolfatto, P. 2006. Selection, recombination and demographic history in *Drosophila miranda*. *Genetics* **174**: 2045–2059.

Bakewell, M.A., Shi, P. & Zhang, J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl Acad. Sci. USA* **104**: 7489–7494.

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

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.W., Poh, Y.P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N., Pachter, L., Myers, E. & Langley, C.H. 2007. Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS Biol.* **5**: e310.

Bierne, N. & Eyre-Walker, A. 2004. The genomic rate of adaptive amino acid substitution in *Drosophila*. *Mol. Biol. Evol.* **21**: 1350–1360.

Bosco, G., Campbell, P., Leiva-Neto, J.T. & Markow, T.A. 2007. Analysis of *Drosophila* species genome size and satellite DNA

content reveals significant differences among strains as well as between species. *Genetics* **177**: 1277–1290.
Caceres, M., Barbadilla, A. & Ruiz, A. 1999. Recombination rate predicts inversion size in Diptera. *Genetics* **153**: 251–259.
Capy, P. & Gibert, P. 2004. *Drosophila melanogaster*, *Drosophila simulans*: so similar yet so different. *Genetica* **120**: 5–16.
Casillas, S., Petit, N. & Barbadilla, A. 2005. DPDB: a database for the storage, representation and analysis of polymorphism in the *Drosophila* genus. *Bioinformatics* **21**(Suppl 2): ii26–ii30.
Casillas, S., Egea, R., Petit, N., Bergman, C.M. & Barbadilla, A. 2007. Drosophila Polymorphism DataBase (DPDB): a portal for nucleotide polymorphism in *Drosophila*. *Fly* **1**: 205–211.
Couronne, O., Poliakov, A., Bray, N., Ishkhanov, T., Ryaboy, D., Rubin, E., Pachter, L. & Dubchak, I. 2003. Strategies and tools for whole-genome alignments. *Genome Res.* **13**: 73–80.
Drosophila 12 Genomes Consortium [128 co-authors] 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **450**: 203–218.
Edgar, R.C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**: 1792–1797.
Egea, R., Casillas, S. & Barbadilla, A. 2008. Standard and generalized McDonald–Kreitman test: a website to detect selection by comparing different classes of DNA sites. *Nucleic Acids Res.* **36**: W157–W162.
Eyre-Walker, A. 2002. Changing effective population size and the McDonald–Kreitman test. *Genetics* **162**: 2017–2024.
Eyre-Walker, A., Keightley, P.D., Smith, N.G. & Gaffney, D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* **19**: 2142–2149.
Fay, J.C., Wyckoff, G.J. & Wu, C.I. 2001. Positive and negative selection on the human genome. *Genetics* **158**: 1227–1234.
Fay, J.C., Wyckoff, G.J. & Wu, C.I. 2002. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415**: 1024–1026.
Felsenstein, J. 1985. Phylogenies and the comparative method. *Am. Nat.* **125**: 1–15.
Haddrill, P.R., Halligan, D.L., Tomaras, D. & Charlesworth, B. 2007. Reduced efficacy of selection in regions of the Drosophila genome that lack crossing over. *Genome Biol.* **8**: R18.
Heger, A. & Ponting, C.P. 2007. Variable strength of translational selection among 12 *Drosophila* species. *Genetics* **177**: 1337–1348.
Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., Weber, R.J., Haussler, D. & Kent, W.J. 2003. The UCSC Genome Browser Database. *Nucleic Acids Res.* **31**: 51–54.
Kimura, M. 1983. *The neutral theory of molecular evolution*. Cambridge University Press, Cambridge.
Kimura, M. 1991. The neutral theory of molecular evolution: a review of recent evidence. *Jpn. J. Genet.* **66**: 367–386.
Kimura, M., Maruyama, T. & Crow, J.F. 1963. The mutation load in small populations. *Genetics* **48**: 1303–1312.
Kliman, R.M., Andolfatto, P., Coyne, J.A., Depaulis, F., Kreitman, M., Berry, A.J., McCarter, J., Wakeley, J. & Hey, J. 2000. The population genetics of the origin and divergence of the *Drosophila simulans* complex species. *Genetics* **156**: 1913–1931.
Kosiol, C., Vinar, T., da Fonseca, R.R., Hubisz, M.J., Bustamante, C.D., Nielsen, R. & Siepel, A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet.* **4**: e1000144.
Lynch, M. & Conery, J.S. 2003. The origins of genome complexity. *Science* **302**: 1401–1404.
Macpherson, J.M., Sella, G., Davis, J.C. & Petrov, D.A. 2007. Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in *Drosophila*. *Genetics* **177**: 2083–2099.
McDonald, J.H. & Kreitman, M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* **351**: 652–654.
McVean, G.A. & Vieira, J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in *Drosophila*. *Genetics* **157**: 245–257.
Moriyama, E.N. & Powell, J.R. 1996. Intraspecific nuclear DNA variation in *Drosophila*. *Mol. Biol. Evol.* **13**: 261–277.
Mousset, S. & Derome, N. 2004. Molecular polymorphism in *Drosophila melanogaster* and *D. simulans*: what have we learned from recent studies? *Genetica* **120**: 79–86.
Nei, M. & Gojobori, T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.* **3**: 418–426.
Nolte, V. & Schlotterer, C. 2008. African *Drosophila melanogaster* and *D. simulans* populations have similar levels of sequence variability, suggesting comparable effective population sizes. *Genetics* **178**: 405–412.
Ohta, T. 1976. Role of very slightly deleterious mutations in molecular evolution and polymorphism. *Theor. Popul. Biol.* **10**: 254–275.
Ohta, T. 2002. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99**: 16134–16137.
Rozas, J., Sanchez-DelBarrio, J.C., Messeguer, X. & Rozas, R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**: 2496–2497.
Shapiro, J.A., Huang, W., Zhang, C., Hubisz, M.J., Lu, J., Turissini, D.A., Fang, S., Wang, H.Y., Hudson, R.R., Nielsen, R., Chen, Z. & Wu, C.I. 2007. Adaptive genic evolution in the *Drosophila* genomes. *Proc. Natl Acad. Sci. USA* **104**: 2271–2276.
Singh, N.D., Bauer Dumont, V.L., Hubisz, M.J., Nielsen, R. & Aquadro, C.F. 2007. Patterns of mutation and selection at synonymous sites in *Drosophila*. *Mol. Biol. Evol.* **24**: 2687–2697.
Smith, N.G. & Eyre-Walker, A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* **415**: 1022–1024.
Sokal, R.R. & Rohlf, F.J. 1995. *Biometry: the principles and practice of statistics in biological research*, 3rd edn. W. H. Freeman and Co., New York.
Studer, R.A., Penel, S., Duret, L. & Robinson-Rechavi, M. 2008. Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes. *Genome Res.* **18**: 1393–1402.
Tajima, F. 1989a. The effect of change in population size on DNA polymorphism. *Genetics* **123**: 597–601.
Tajima, F. 1989b. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**: 585–595.
Vicario, S., Moriyama, E.N. & Powell, J.R. 2007. Codon usage in twelve species of *Drosophila*. *BMC Evol. Biol.* **7**: 226.
Welch, J.J. 2006. Estimating the genomewide rate of adaptive protein evolution in *Drosophila*. *Genetics* **173**: 821–837.
Wright, S. 1931. Evolution in Mendelian populations. *Genetics* **16**: 97–159.
Wright, F. 1990. The 'effective number of codons' used in a gene. *Gene* **87**: 23–29.
Yi, S. & Streelman, T. 2005. Genome size is negatively correlated with effective population size in ray-finned fish. *Trends Genet.* **21**: 643–646.

Yi, S., Bachtrog, D. & Charlesworth, B. 2003. A survey of chromosomal and nucleotide sequence variation in *Drosophila miranda*. *Genetics* **164**: 1369–1381.

## Supporting information

Additional supporting information may be found in the online version of this article:

**Table S1** DATASET get from DPDB after the filtering (see methods).

**Table S2** List of orthologous genes compared between species pairs by paired *t*-test.

**Table S3** Results of 10 analyses of variance of the model $\pi_s$ = species + GO.

**Table S4** Mean and standard deviation of the levels of synonymous polymorphism and selective constraints by functional categories.

**Table S5** Mean observed and expected values of Tajima's *D*.

**Table S6** Results of the regression analysis between the dependent variable nonsynonymous polymorphism and independent variable synonymous polymorphism for each species.

**Table S7** Average $\alpha$ and confidence intervals at 95% calculated by bootstrapping of $\alpha$ by gene.

**Table S8** Resamplig of the datasets and average $\alpha$ value by dataset.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

*Received 03 September 2008; revised 11 November 2008; accepted 13 November 2008*