On the Utility of Short Intron Sequences as a Reference for the Detection of Positive and Negative Selection in *Drosophila*

John Parsch,*[,1] Sergey Novozhilov,[1] Sarah S. Saminadin-Peter,†[,1] Karen M. Wong,[2] and Peter Andolfatto[2]

[1]Department of Biology II, University of Munich, Planegg-Martinsried, Germany

[2]Department of Ecology and Evolutionary Biology and the Lewis-Sigler Institute for Integrative Genomics, Princeton University

†Present address: Department of Systems Biology, Harvard Medical School, Boston, Massachusetts

***Corresponding author:** E-mail: parsch@bio.lmu.de.

Associate editor: Arndt von Haeseler

Research article

Abstract

The detection of selection, both positive and negative, acting on a DNA sequence or class of nucleotide sites requires comparison with a reference sequence that is unaffected by selection. In *Drosophila*, recent findings of widespread selective constraint, as well as adaptive evolution, in both coding and noncoding regions highlight the difficulties in choosing such a reference sequence. Here, we investigate the utility of short intron sequences as a reference for the detection of selection. For a set of 119 *Drosophila melanogaster* genes containing 195 short introns ($\leq$120 bp), we analyzed polymorphism and divergence at 1) 4-fold synonymous sites, 2) all sites of introns $\leq$120 bp, 3) all sites of introns $\leq$65 bp, 4) bases 8–30 of introns $\leq$120 bp, and 5) bases 8–30 of introns $\leq$65 bp. The last class of sites shows the highest levels of both interspecific divergence and intraspecific polymorphism, suggesting that these sites are under the least selective constraint. Bases 8–30 of introns $\leq$65 bp also have the lowest ratio of divergence to polymorphism, which may indicate that a small proportion of substitutions in the other classes of sites are the result of adaptive evolution. Although there is little signal of selection on the primary sequence of short introns, patterns of insertion–deletion polymorphism and divergence suggest that both positive and negative selection act to maintain an optimal intron length.

Key words: *Drosophila*, selective constraint, positive selection, population genetics, intron evolution.

Introduction

Unconstrained DNA sequences (i.e., those whose evolution is unaffected by selection) are useful for many molecular evolutionary and population genetic analyses, including phylogenetic reconstruction, demographic inference, and the detection of selection. For this last purpose, it is essential to have a selection-free reference that can be used as a baseline for determining the type and strength of selection acting on a particular target sequence (or collection of sites/sequences).

Traditionally, synonymous sites within protein-coding regions have been used as a reference for unconstrained evolution. Due to the degeneracy of the genetic code, mutations at these sites do not alter the amino acid sequence of the encoded protein and, thus, are expected to be invisible to selection. The presumed lack of constraint at synonymous sites forms the basis of many tests of adaptive protein evolution (e.g., McDonald and Kreitman 1991; Yang et al. 2000; Fay et al. 2001; Bustamante et al. 2002; Smith and Eyre-Walker 2002; Andolfatto 2007; Sawyer et al. 2007). However, several observations suggest that synonymous sites may be under at least weak selection. For example, almost all genomes investigated to date show some degree of codon bias in which the synonymous codons for a given amino acid are not used with equal frequency (reviewed by Hershberg and Petrov 2008). In many cases, this is thought to be a result of selection favoring codons that can be translated more rapidly or accurately (Akashi 1994, 1995; Carlini and Stephan 2003; Stoletzki and Eyre-Walker 2007). Additionally, synonymous sites may be under selection to maintain (or avoid) splicing enhancers (Parmley et al. 2006), messenger RNA secondary structures (Parsch et al. 1997; Baines et al. 2004; Stoletzki 2008), or particular short sequence motifs (Antezana and Kreitman 1999).

Noncoding DNA sequences, such as introns and intergenic regions, are also potential candidates for unconstrained sites. Because the vast majority of DNA in the genomes of higher eukaryotes does not encode proteins, it is often assumed that much of it is junk DNA that is under no selective constraint. However, it is possible that these sequences have functions that are unrelated to protein coding, and a number of recent findings suggest that noncoding DNA does not evolve free of constraint. For example, many eukaryotic genomes contain highly conserved noncoding sequences that are subject to purifying selection (Shabalina and Kondrashov 1999; Shabalina et al. 2001; Siepel et al. 2005; Drake et al. 2006; Asthana et al. 2007). In *Drosophila*, levels of divergence in intergenic regions and long introns indicate that $\geq$50% of sites are subject

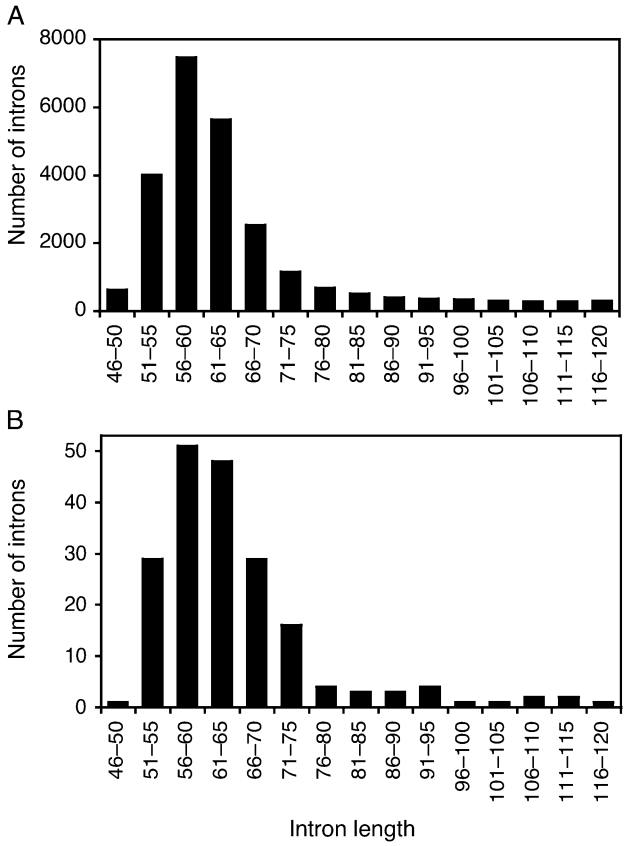

FIG. 1. Length distribution of short introns. (*A*) Whole-genome distribution of intron lengths in *Drosophila melanogaster*. Introns with length >120 bp are not shown, but none of the 5-bp bins >120 bp contain more than 214 introns. Genome-wide median intron length is 68 bp, with 46% of introns having length ≤65 bp. (*B*) Length distribution of introns used in the present study. The median intron length is 63 bp, with 66% of the introns having length ≤65 bp.

to purifying selection (Bergman and Kreitman 2001; Andolfatto 2005; Bachtrog and Andolfatto 2006; Halligan and Keightley 2006; *Drosophila 12* Genomes Consortium 2007), an inference that is supported by analyses of polymorphism patterns (Andolfatto 2005; Bachtrog and Andolfatto 2006; Casillas et al. 2007; Haddrill et al. 2008). In addition, introns and intergenic regions show evidence for a divergence excess relative to neutral expectations in *Drosophila melanogaster–D. simulans* comparisons (Kohn et al. 2004; Andolfatto 2005; Casillas et al. 2007; Haddrill et al. 2008), which is consistent with the action of recurrent positive selection.

Another class of sites that has been proposed to evolve free of selection is short intron sequences. In *Drosophila*, there is an asymmetrical distribution of intron lengths (fig. 1*A*), with a high proportion of "short" introns falling into a narrow size range (~40–120 bp) and the remaining "long" introns following a more uniform size distribution (hundreds to thousands of base pairs), although the boundary between short and long introns is not discrete (Mount et al. 1992; Deutsch and Long 1999; Comeron and Kreitman 2000). Short introns appear to be under less evolutionary constraint than long introns (Parsch 2003), and several authors have reported a negative correlation between intron length and interspecific divergence (Haddrill, Charlesworth, et al. 2005; Marais et al. 2005; Halligan and Keightley 2006). Within short introns, the least constrained sites are those falling between the 5′ and 3′ regions of the intron that function in splice site recognition (Halligan and Keightley 2006). Halligan and Keightley (2006) found the fastest evolving (and presumably least constrained) intron sites to be bases 8–30 of introns ≤65 bp, which show greater divergence between species than 4-fold degenerate synonymous sites. These findings suggest that short intron sequences (or portions thereof) may be the most appropriate reference for the inference of selection.

To further explore this possibility, we examined DNA sequence polymorphism and divergence at 119 *Drosophila* genes containing 195 short introns (mean intron length = 65 bp, range = 46–120 bp). We find that both intraspecific polymorphism and interspecific divergence are slightly higher at intronic sites than at 4-fold degenerate synonymous sites, which is consistent with short introns being under less selective constraint. Despite this, we find that bases 8–30 of introns ≤65 bp show the lowest ratio of divergence to polymorphism and, when used as a reference, yield slightly higher estimates of the prevalence of positive selection than synonymous sites in the *D. melanogaster* lineage. Although there appears to be little selective constraint on the sequence of short introns, patterns of insertion–deletion (indel) polymorphism and divergence suggest that the length of these introns is subject to both positive and negative selection.

Materials and Methods

Sequence Data

The data set consists of 119 protein-coding genes, all of which contain at least one short intron (here defined as ≤120 bp and measured as the alignment length of all *D. melanogaster* sequences after removal of gaps). This includes 53 autosomal genes previously reported by Pröschel et al. (2006), 18 X-linked genes reported by Baines et al. (2008), 12 X-linked genes reported by Andolfatto (2007), 1 X-linked gene reported by Glinka et al. (2003), and 35 genes (12 autosomal and 23 X-linked) that are new to this analysis. In addition to at least one short intron, 16 of the 119 genes also contained one or more introns greater than 120 bp in length. These long introns (in total 17) were excluded from the analysis. A complete list of genes and their GenBank/EMBL accession numbers are provided in supplementary table S1 (Supplementary Material online). Alignments of all loci are provided in supplementary figure S1 (Supplementary Material online). Because polymorphism is reduced and nonsynonymous divergence is elevated in chromosomal regions with very low recombination, we excluded genes from regions with recombination rates below 0.2 cM/Mb (Glinka et al. 2003; Haddrill et al. 2007). There was not a significant correlation between recombination rate and nucleotide polymorphism (Pearson's R = 0.11,

$P = 0.31$) or nonsynonymous divergence (Pearson's $R = -0.16$, $P = 0.14$) in the genes used in our analysis. To minimize the sampling variance of synonymous and nonsynonymous sites across loci, we chose genes from a relatively narrow size range, with the mean number of coding sites per gene being 746 (standard deviation = 289).

For all the above genes, polymorphism statistics were calculated using a sample of *D. melanogaster* alleles from Zimbabwe, Africa. The number of sequenced alleles ranged from 7 to 12 per gene, with the majority having a sample size of 12 (mean = 11.2). For the genes new to this analysis, sequences were generated by direct sequencing of polymerase chain reaction products (both strands) using BigDye chemistry and a 3730 automated sequencer (Applied Biosystems, Foster City, CA). Sequencing was performed on the same sets of Zimbabwe strains used in previous studies (Pröschel et al. 2006; Andolfatto 2007; Baines et al. 2008).

Divergence statistics were calculated using a single allele from *D. simulans*. When available (47 of 119 genes), we used the previously published sequence from an inbred *D. simulans* strain from Chapel Hill, NC (Meiklejohn et al. 2004). The reference sequence from the *D. simulans* genome project (*Drosophila* 12 Genomes Consortium 2007) was used for 63 of the 119 genes. For the remaining nine genes, the complete *D. simulans* sequence was not available, and the reference sequence of its sister species, *D. sechellia*, was used in its place. The reference *D. yakuba* sequence was used as an outgroup for ancestral reconstruction. All alignments were performed using MUSCLE (http://www.drive5.com/muscle) with manual adjustments and, in the case of coding regions, adjustments to preserve reading frame.

Analysis of Nucleotide Polymorphism and Divergence

Interspecific nucleotide divergence (K) was calculated as the average pairwise divergence between all *D. melanogaster* sequences and a single *D. simulans* (or *D. sechellia*) outgroup sequence, with Jukes–Cantor correction for multiple hits. When multiple differences were present within a single codon of a protein-coding region, we chose the mutation pathway that minimized the number of nonsynonymous changes (Nei and Gojobori 1986). For intron sequences, the invariant GT and AG dinucleotides at the 5′ and 3′ splice junctions, respectively, were excluded before calculating divergence. However, these sites were included when calculating total intron length and when extracting bases 8–30 of introns. Only aligned nucleotide positions (without gaps) were used for counting bases 8–30. To polarize nucleotide divergence to either the *D. melanogaster* or the *D. simulans* lineage, the ancestral sequence of the two species was reconstructed by maximum likelihood as implemented in the *codeml* (for coding regions, free ratio model [model = 1]) and *baseml* (for noncoding regions) programs of PAML (Yang 2007) using the *D. yakuba* sequence as the outgroup. Selective constraint (C) was estimated from divergence data using the formula $C = 1 - (K_{test}/K_{ref})$, where the subscripts indicate the test and the reference sites, respectively (Kondrashov and Crow 1993; Eyre-Walker and Keightley 1999).

Intraspecific nucleotide polymorphism was calculated in two ways: as the average number of pairwise differences among *D. melanogaster* alleles (π) with Jukes–Cantor correction and as Watterson's (1975) estimator of diversity (θ). To calculate θ, we used the number of segregating mutations as there were some rare cases in which three different nucleotides were segregating at a single position in the alignment. Differences between the two above estimators of nucleotide diversity were quantified by Tajima's (1989) D statistic. Because Tajima's D is sensitive to the sample size and the number of segregating sites, we also determined D', which is the ratio of D to the absolute value of its theoretical minimum (Schaeffer 2002). The derived allele frequency (DAF) was calculated as the average frequency of derived mutations over all sites, with the derived state being determined by parsimony using *D. simulans* as the outgroup. Sites for which the derived state could not be determined unambiguously were excluded from the analysis. Selective constraint was estimated from polymorphism data using the formula $C = 1 - (\pi_{test}/\pi_{ref})$.

A bootstrap analysis was performed by randomly resampling (with replacement) the loci in our original data set, with the condition that the proportion of genes containing an intron $\leq$65 bp remained constant. For each of 10,000 replicates, we calculated mean values of π and θ, as well as the summed ratio of polymorphic-to-divergent sites (D/P), for all classes of sites. We then compared these values between bases 8–30 of introns $\leq$65 bp and all other classes of sites to determine the proportion of replicates in which bases 8–30 of introns $\leq$65 bp had a higher π or θ (or lower D/P) than the other site classes.

The proportion of fixed differences between species attributable to positive selection (α) was determined using two multilocus implementations of the test of McDonald and Kreitman (1991): the method of Fay et al. (2001) that uses the sums of all polymorphic and divergent sites over all loci and the method of Bierne and Eyre-Walker (2004) that performs a maximum likelihood estimation of α using paired polymorphism and divergence counts within each locus. Because segregating deleterious mutations may lead to the underestimation of α, we excluded all low-frequency polymorphisms (frequency $\leq$15%) from our analysis (Charlesworth and Eyre-Walker 2008; Parsch et al. 2009). For both of the above methods, α, its 95% confidence interval, and the probability of $\alpha > 0$ were calculated using the program DoFE (http://www.lifesci.sussex.ac.uk/home/Adam_Eyre-Walker/Website/Software.html).

Analysis of Indels

Indels that were polymorphic within *D. melanogaster* were classified as either insertions or deletions by comparison with the *D. simulans* sequence under the assumption of parsimony. In rare cases where there were multiple overlapping indels and the classification was ambiguous (typically involving repetitive sequences), the indel was excluded from the analysis. To classify divergent indels and assign them to either the *D. melanogaster* or the *D. simulans* lineage, we used *D. yakuba* as the outgroup and

Table 1. Summary of Interspecific Divergence.

Class	Sites	K_{ms}[a]	K_m[b]	K_m/K_{ms}	C_m[c]
Nonsynonymous	70,668	1.9	0.9	0.47	86.0
Synonymous (4-fold)	14,009	10.6	6.0	0.56	7.6
Intron	11,531	11.4	5.5	0.49	14.1
Intron ≤65	6,917	11.0	5.8	0.53	10.2
Intron 8–30	4,485	12.3	6.2	0.51	3.2
Intron ≤65 8–30	2,967	12.3	6.4	0.53	—

[a] Mean divergence (per 100 sites) between all *Drosophila melanogaster* sequences and *D. simulans*.
[b] Mean divergence (per 100 sites) on the *D. melanogaster* lineage only.
[c] Constraint; the percentage of sites subject to purifying selection on the *D. melanogaster* lineage relative to bases 8–30 of introns ≤65 bp, as estimated from interspecific divergence.

Table 2. Summary of Intraspecific Polymorphism.

Class	π[a]	θ[b]	Taj D[c]	Taj D'[d]	DAF	C[e]
Nonsynonymous	0.15	0.19	−0.64	−0.42	0.226	93.0
Synonymous (4-fold)	1.78	1.85	−0.18	−0.11	0.313	18.3
Intron	1.61	1.66	−0.11	−0.08	0.317	25.8
Intron ≤65	1.84	1.90	−0.09	−0.05	0.321	15.4
Intron 8–30	1.90	2.01	−0.15	−0.11	0.295	12.7
Intron ≤65 8–30	2.17	2.32	−0.14	−0.09	0.306	—

[a] Mean pairwise nucleotide diversity (per 100 sites).
[b] Watterson's (1975) estimator of nucleotide diversity (per 100 sites).
[c] Tajima's (1989) D statistic.
[d] Tajima's D relative to its theoretical minimum given the sample size and number of segregating sites (Schaeffer 2002).
[e] Constraint; the percentage of sites subject to purifying selection relative to bases 8–30 of introns ≤65 bp, as estimated from intraspecific polymorphism.

followed the parsimony approach of Presgraves (2006) (see his fig. 1). Indels that could not be assigned unambiguously to a lineage were excluded from the analysis.

Results and Discussion

Selective Constraint: Interspecific Divergence

Our data set consists of 119 protein-coding genes, each containing at least one short intron (≤120 bp). Introns greater than 120 bp in length (long introns) were excluded from the analysis. In total, there are 195 short introns, 129 of which are ≤65 bp (fig. 1*B*). This allows us to analyze several classes of putatively unconstrained sites: 1) 4-fold synonymous sites, 2) all intronic sites, 3) all sites of introns ≤65 bp, 4) bases 8–30 of all introns, and 5) bases 8–30 of introns ≤65 bp. The last class of sites was suggested by Halligan and Keightley (2006) to be the least selectively constrained.

Mean values of interspecific divergence between *D. melanogaster* and *D. simulans* are shown in Table 1. Divergence is significantly lower at nonsynonymous sites than at all other classes of sites (Mann–Whitney test, $P < 0.001$ in all cases) but does not differ significantly among any of the putatively unconstrained classes of sites by the Mann–Whitney test. However, a bootstrap analysis indicates that divergence is consistently higher for bases 8–30 of introns ≤65 bp than for other classes of sites (fig. 2). This is in agreement with Halligan and Keightley (2006) and suggests that at least some 4-fold synonymous and intronic sites are subject to purifying selection. A similar pattern is seen when divergence is measured only on the *D. melanogaster* lineage (table 1). Within our set of short introns, there was no correlation between intron length and interspecific divergence (Pearson's $R = -0.064$, $P = 0.48$).

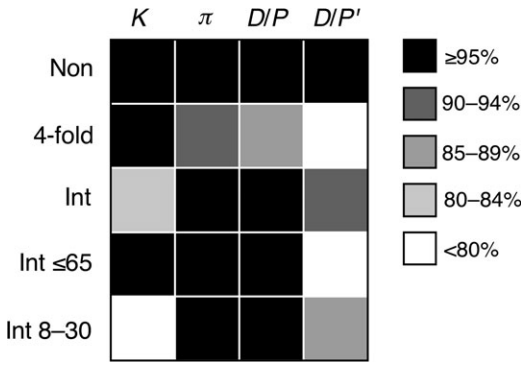

Fig. 2. Results of bootstrap analysis. For K and π, shading indicates the proportion of 10,000 bootstrap replicates in which bases 8–30 of introns ≤65 bp had a higher mean value than the class of sites given at the left. For D/P and D/P', shading indicates the proportion of replicates in which bases 8–30 of introns ≤65 bp had a lower summed ratio than the class of sites given at the left. Non, nonsynonymous sites; 4-fold, 4-fold synonymous sites; Int, all sites of introns ≤120 bp; Int ≤65, all sites of introns ≤65 bp; Int 8–30, sites 8–30 of all introns ≤120 bp.

Using bases 8–30 of introns ≤65 bp as the reference, we estimated the selective constraint on the other classes of sites on the *D. melanogaster* lineage (table 1). For example, 86% of substitutions in nonsynonymous sites in our sample appear to be subject to purifying selection (table 1), which matches the previous genome-wide estimate (Halligan and Keightley 2006). The different classes of putatively unconstrained sites show levels of constraint ranging from 3% to 14%, with the greatest constraint on intronic sites (table 1). Our estimate of constraint at 4-fold synonymous sites on the *D. melanogaster* lineage is 7.6%, which is slightly lower than the 12.6% estimated by Halligan and Keightley (2006) from the total divergence between *D. melanogaster* and *D. simulans*. This is consistent with there being weaker selection on synonymous codon usage in *D. melanogaster* than in *D. simulans* (Akashi 1995). Note that these estimates of constraint on synonymous sites assume equal mutation rates, which may not be the case if noncoding DNA is more highly mutable (Andolfatto 2005).

Selective Constraint: Intraspecific Polymorphism

In addition to interspecific divergence data, we also have intraspecific polymorphism data for all the genes in our sample, which allows us to examine selective constraint on the various classes of sites at the population level. For this, we used a population of *D. melanogaster* from Zimbabwe, Africa, that is less affected by the recent demographic changes that other worldwide populations have experienced (Dieringer et al. 2005; Haddrill, Thornton, et al. 2005; Stephan and Li 2006).

Overall, nucleotide sequence polymorphism follows the same general pattern as divergence, with the highest levels of polymorphism at bases 8–30 of introns ≤65 bp (table 2),

Table 3. Numbers of Polymorphic and Divergent Sites.

Class	D[a]	P[b]	P'[c]	D/P	D/P'
Nonsynonymous	1,225	392	162	3.13	7.56
Synonymous (4-fold)	1,159	758	422	1.53	2.75
Intron	1,021	545	309	1.87	3.30
Intron ≤65	574	387	212	1.48	2.71
Intron 8–30	411	262	137	1.57	3.01
Intron ≤65 8–30	264	201	98	1.31	2.69

[a] Number of fixed differences between *Drosophila melanogaster* and *D. simulans*.
[b] Number of polymorphic sites within *D. melanogaster*.
[c] Number of polymorphic sites after excluding low-frequency (≤15%) polymorphisms.

which also suggests that these sites are under the least selective constraint. Nucleotide diversity (π) is significantly lower at nonsynonymous sites than at all other classes of sites (Mann–Whitney test, $P < 0.001$ in all cases) but does not differ significantly among any of the other classes of sites. There is no correlation between intron length and π (Pearson's $R = -0.046$, $P = 0.61$). However, mean values of π are consistently higher for bases 8–30 of introns <65 bp than for other sites in a bootstrap analysis (fig. 2). Using bases 8–30 of introns ≤65 bp as the reference, we estimated the selective constraint on the other classes of sites based on levels of intraspecific polymorphism (table 2). In all cases, the constraint estimated from polymorphism was greater than that estimated from divergence (table 1). This may be because positively selected mutations appear disproportionately in divergence and lead to reduced estimates of constraint.

At the population level, selective constraint may also be detected by an excess of low-frequency polymorphisms that leads to negative values of Tajima's D. We observe negative values of Tajima's D at all classes of sites, but the only clear outlier is the nonsynonymous sites, which show a large excess of low-frequency variants (table 2). At nonsynonymous sites, Tajima's D' (the ratio of Tajima's D to the absolute value of its theoretical minimum; Schaeffer 2002) is −0.42 and is significantly less than that at the other classes of sites (Mann–Whitney test, $P < 0.002$ in all cases). This indicates that purifying selection keeps nonsynonymous mutations at low frequency in the population. For the other classes of sites, Tajima's D' falls between −0.11 and −0.05 and does not differ significantly among classes. The overall excess of rare alleles observed for all sites may be the result of an ancient population expansion in the African population (Stephan and Li 2006), although a past bottleneck (Haddrill, Thornton, et al. 2005), direct effects of purifying selection (Haddrill et al. 2008), or indirect effects of linked positive selection (Andolfatto 2007) are also possibilities.

Using *D. simulans* as an outgroup to polarize segregating mutations as either ancestral or derived, we calculated the average DAF over all sites. We find that nonsynonymous sites show significantly lower DAF than all other classes of sites (Mann–Whitney test, $P < 0.05$ in all cases). There are no significant differences in DAF among the other classes of sites.

Table 4. Estimates of the Proportion of Adaptive Substitutions at Nonsynonymous Sites Using Different Classes of Sites as the Reference.

Class	α_{FWW}[a] (95% CI)	α_{BEW}[b] (95% CI)
Synonymous (all)	63.3 (53.4–71.4)	48.3 (39.4–60.1)
Synonymous (4-fold)	63.7 (53.4–70.7)	51.8 (38.1–62.0)
Intron	56.3 (40.9–66.7)	45.5 (27.4–57.4)
Intron ≤65	64.1 (45.4–75.4)	55.2 (35.2–66.5)[c]
Intron 8–30	60.3 (42.9–71.5)	50.6 (31.1–63.8)
Intron ≤65 8–30	64.4 (40.9–77.0)	55.4 (24.5–68.9)[c]

NOTE.—CI, confidence interval.
[a] Percent of nonsynonymous substitutions fixed by positive selection as estimated by the method of Fay et al. (2001).
[b] Percent of nonsynonymous substitutions fixed by positive selection as estimated by the method of Bierne and Eyre-Walker (2004).
[c] Estimated from the 86 genes containing at least one ≤65 bp intron. All other estimates are from the full set of 119 genes.

Detection of Positive Selection: Polymorphism and Divergence

Having the combination of polymorphism and divergence data allows us to test not only for negative selection but also for positive selection. To do this, we used two multilocus implementations of the McDonald–Kreitman test to estimate the proportion of fixed differences between species that can be attributed to positive selection (α). Nonsynonymous sites were used as the test sites, and the other classes of sites were used as the reference. The first method compares ratios of divergence with polymorphism between the test sites and the reference sites using the summed values of polymorphic and divergent sites over all loci (Fay et al. 2001). This method has the advantage of being applicable to all loci, even those that do not contain short introns. However, it has the drawback that it may, in certain situations, lead to biased estimates of adaptive divergence (Smith and Eyre-Walker 2002; Welch 2006; Shapiro et al. 2007) or overconfidence about uncertainty in estimates (Andolfatto 2005, 2008). With the approach of Fay et al. (2001), α can be calculated as $1 - [(D_{ref}/P_{ref})/(D_{test}/P_{test})]$, where D and P represent the summed numbers of divergent and polymorphic sites, respectively, and the subscripts indicate the reference and the test sites, respectively. Positive selection ($\alpha > 0$) is detected when $D_{ref}/P_{ref} < D_{test}/P_{test}$. Table 3 shows the D/P ratio for the different classes of sites. The lowest ratio is found for bases 8–30 of introns ≤65 bp, which is consistent across bootstrap replicates (fig. 2). Concordantly, the highest value of α for nonsynonymous sites is obtained when bases 8–30 of introns ≤65 bp are used as the reference (table 4). However, the signal of positive selection at nonsynonymous sites is strong (α ranging from 56% to 64%) regardless of which sites are used as the reference, and there is considerable overlap in the α estimates generated by the different site classes, including the use of all synonymous sites as the common practice (table 4). Using bases 8–30 of introns ≤65 bp as the reference, we can also estimate α for the other classes of sites. These estimates range from 0.5% to 18.5%, with the highest α observed for all intronic sites (fig. 3). However, the 95% confidence intervals of α include zero for every class of sites (fig. 3).

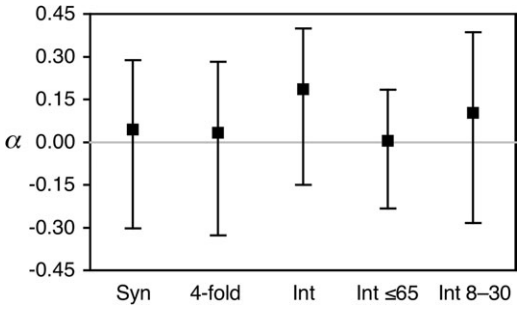

Fig. 3. The proportion of interspecific divergence at different classes of sites that can be attributed to positive selection (α) when bases 8–30 of introns $\leq$65 bp are used as the reference. Values were calculated using the method of Fay et al. (2001), with low-frequency polymorphisms ($\leq$15%) excluded. In all cases, the mean value of α (solid boxes) is greater than zero, although zero falls within the 95% confidence interval (error bars). Syn, all synonymous sites; 4-fold, 4-fold synonymous sites; Int, all sites of introns $\leq$120 bp; Int $\leq$65, all sites of introns $\leq$65 bp; Int 8–30, sites 8–30 of all introns $\leq$120 bp.

We also estimated α for nonsynonymous sites using the maximum likelihood method of Bierne and Eyre-Walker (2004). Like the original McDonald–Kreitman test, this method pairs the test and reference sites within each locus and, thus, avoids some of the potential problems of the method of Fay et al. (2001). However, it has the limitation that it cannot be applied to all loci. For example, if one wishes to use intronic sites as the reference, only genes containing introns (and of the correct size) can be tested. The α estimates for nonsynonymous sites produced by this method ranged from 45% to 55%, with the highest value produced when bases 8–30 of introns $\leq$65 bp was used as the reference (table 4).

A concern with using McDonald–Kreitman–based approaches to infer adaptive evolution is that the combination of weak purifying selection and changing population size may lead to overestimates of α (Eyre-Walker 2002). This occurs when slightly deleterious mutations contribute more to divergence than to polymorphism because the effective population size has increased since the time of species divergence. We tried to minimize this effect by using an African population of *D. melanogaster* that does not show evidence of a recent bottleneck like non-African populations do (Haddrill, Thornton, et al. 2005; Stephan and Li 2006). However, we cannot rule out population size changes over a longer timescale. In any case, the above problem would occur when using any class of sites as the neutral reference. It has been proposed that weak selection at synonymous sites could counteract the effect of weak selection at nonsynonymous sites in tests of adaptive evolution (Eyre-Walker 2002). However, we find no evidence from the frequencies of segregating mutations that weak purifying selection is more prevalent at synonymous sites than at the other classes of reference sites (table 2).

Comparison of Autosomal and X-Linked Loci

The above analyses used a pooled data set of 65 autosomal and 54 X-linked genes containing 116 and 79 short introns, respectively. Analyzing the autosomal and X-linked loci separately gives similar results, with the highest levels of both polymorphism and divergence at bases 8–30 of introns $\leq$65 bp (table 5). There is a nonsignificant trend of higher constraint and greater adaptive evolution at X-linked loci (Mann–Whitney test, $P > 0.05$ in all cases), which may be a result of more efficient selection (both positive and negative) acting on X-linked recessive mutations in hemizygous males. For autosomal loci, the ratio of polymorphism to divergence (after excluding low-frequency polymorphisms) at bases 8–30 of introns $\leq$65 bp is slightly higher than that of 4-fold synonymous sites or other intronic sites. This leads to slightly lower estimates of α at nonsynonymous sites when bases 8–30 of introns $\leq$65 bp are used as the reference by the method of Fay et al. (2001). The method of Bierne and Eyre-Walker (2004), however, produces the highest estimates of α at nonsynonymous sites when bases 8–30 of introns $\leq$65 bp are used as the reference (table 5).

Selective Constraint on Intron Length: Indel Polymorphism and Divergence

Although there appears to be little selective constraint on the primary sequence of short introns, several studies have suggested that intron length is subject to purifying selection (Stephan et al. 1994; Carvalho and Clark 1999; Yu et al.

Table 5. Comparison of Autosomal and X-Linked Loci.

Chromosome	Class	π	C_{pol}[a]	K_m	C_{div}[b]	D/P	D/P'	α_{FWW}	α_{BEW}
Autosome	**Nonsynonymous**	**0.15**	**92.8**	**0.87**	**86.4**	**2.38**	**5.72**	**—**	**—**
	Synonymous (4-fold)	**1.63**	**20.6**	**5.92**	**6.9**	**1.65**	**2.92**	**49.0**	**44.6**
	Intron	**1.68**	**18.4**	**5.62**	**11.7**	**1.94**	**3.31**	**42.1**	**46.6**
	Intron ≤65	**1.72**	**16.3**	**5.76**	**9.4**	**1.67**	**3.16**	**44.7**	**46.1**
	Intron 8–30	**2.03**	**1.5**	**6.19**	**2.7**	**1.61**	**3.05**	**46.7**	**48.3**
	Intron ≤65 8–30	**2.06**	**—**	**6.36**	**—**	**1.47**	**3.22**	**43.7**	**48.7**
X-linked	**Nonsynonymous**	**0.16**	**93.3**	**0.98**	**85.4**	**4.03**	**9.86**	**—**	**—**
	Synonymous (4-fold)	**1.76**	**25.7**	**6.00**	**10.5**	**1.39**	**2.54**	**74.3**	**59.5**
	Intron	**1.50**	**36.8**	**5.32**	**20.7**	**1.78**	**3.29**	**66.6**	**51.0**
	Intron ≤65	**1.91**	**19.8**	**5.57**	**16.9**	**1.24**	**2.15**	**78.2**	**66.9**
	Intron 8–30	**1.89**	**20.6**	**6.37**	**5.1**	**1.50**	**2.93**	**70.3**	**51.2**
	Intron ≤65 8–30	**2.38**	**—**	**6.71**	**—**	**1.08**	**2.00**	**79.7**	**67.4**

Note.—As in Table 4, the last two columns give the estimated proportion of adaptive substitutions at nonsynonymous sites when using the given sites as the reference.
[a] Constraint (%) estimated from polymorphism.
[b] Constraint (%) estimated from divergence.

Table 6. Indel Polymorphism and Divergence in Short Introns.

Species	Type	Chromosome	Deletion	Insertion	Deletion/Insertion[a]
Drosophila melanogaster	Polymorphism	X-linked	20	14	1.43
		Autosome	12	22	0.55
		All	32	36	0.89
D. melanogaster	Divergence	X-linked	11	18	0.61
		Autosome	22	18	1.22
		All	33	36	0.92
D. simulans	Divergence	X-linked	30	11	2.73**
		Autosome	22	5	4.40**
		All	52	16	3.25***

[a] Asterisks indicate significant departure from 1 (χ^2 test).
**$P < 0.01$, ***$P < 0.001$.

2002; Parsch 2003; Presgraves 2006; Casillas et al. 2007). To investigate this, we examined indel polymorphism and divergence in short introns. In total, 68 indels (32 deletions and 36 insertions) were segregating within *D. melanogaster*, indicating a slight insertion bias (table 6). This contrasts with the overall polymorphism deletion bias that has been reported for long introns and intergenic regions (Comeron and Kreitman 2000; Ometto et al. 2005). Part of this discrepancy may be attributable to the smaller target size in short introns for deletions that do not overlap with conserved splice sites (Ptak and Petrov 2002). It has also been proposed that short introns evolve in a compensatory fashion in which their length decreases due to the accumulation of slightly deleterious deletions until the intron reaches a minimal length required for efficient splicing, after which selection favors insertions (Parsch 2003). Consistent with this model, we find that insertions segregate at significantly higher frequencies than deletions (fig. 4). Deletions segregate at a lower frequency than derived single nucleotide polymorphisms (SNPs), whereas insertions segregate at a higher frequency (fig. 4). If we assume that the vast majority of SNPs within short introns are free of selective constraint, then this suggests that insertions are favored by selection, whereas deletions are disfavored. A similar result was reported by Presgraves (2006) for a different set of introns, although he found an insertion bias only on the X chromosome, whereas we find it on the autosomes (table 6). However, we observe a larger difference in the frequencies of insertions and deletions on the X chromosome than on the autosomes (fig. 4), which was also found by Presgraves (2006).

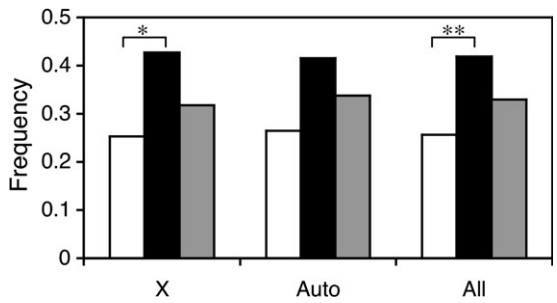

Fig. 4. Mean frequencies of deletions (open bars), insertions (solid bars), and derived SNPs (gray bars) within introns. Values are shown for X-linked loci, autosomal loci, and all loci combined. Brackets indicate significant differences in frequency between groups (Mann–Whitney test). *$P < 0.05$, **$P < 0.01$.

In contrast to the polymorphism results, we observe a strong deletion bias for fixed differences between *D. melanogaster* and *D. simulans*. However, this deletion bias is limited to the *D. simulans* lineage (table 6). The *D. melanogaster* lineage shows a slight insertion bias, which is consistent with the polymorphism results and the compensatory model described above. It is difficult, however, to reconcile the *D. simulans* results with this model. A strong deletion bias in the *D. simulans* lineage was also seen by Presgraves (2006) and suggests that the two lineages may differ in mutational processes, splicing mechanisms, or strength of selection on noncoding sequence length. All the above results hold when single-base indels, which often involve simple sequence repeats, are excluded from the analysis (supplementary table S2, Supplementary Material online), suggesting that the observed patterns are not caused solely by slippage-like mutational processes.

Conclusions

We find that bases 8–30 of short introns show the highest average divergence between *D. melanogaster* and *D. simulans*. This is consistent with previous reports suggesting that these sites are the least constrained in the *Drosophila* genome (Halligan and Keightley 2006; *Drosophila* 12 Genomes Consortium 2007). Previous studies, however, were limited to comparisons of single sequences between species and, thus, could not disentangle the effects of negative and positive selection. For example, it is possible that the higher divergence observed at intron sites 8–30 is a result of increased adaptive evolution at these sites relative to synonymous sites. Our use of both polymorphism and divergence data allows us to rule out this possibility. The polymorphism data also allow us to get better estimates of constraint at all classes of sites because levels of constraint calculated only from divergence will be underestimates if some of the divergence is adaptive (Andolfatto 2005). Indeed, this appears to be the case as our polymorphism data give higher estimates of constraint than the divergence data (compare tables 1 and 2). Taken together, the high levels of polymorphism and divergence indicate

that bases 8–30 of introns ≤65 bp are the best choice as a reference for the detection of selective constraint.

The combination of polymorphism and divergence data also allows us to evaluate the utility of the intronic sites for tests of adaptive evolution, such as the McDonald–Kreitman test. We find that sites 8–30 of introns ≤65 bp are the best sites to use as a reference but do not differ very much from synonymous sites, which are often used for this purpose. This result is relevant to previous studies that have used polymorphism and divergence at synonymous sites to infer adaptive evolution of noncoding regions (e.g., Andolfatto 2005). For example, it has been suggested that nonneutral evolution of synonymous sites, such as purifying selection against unpreferred codon changes or balancing selection to maintain an overall optimal level of codon bias, could inflate the ratio of polymorphism to divergence at synonymous sites and lead to overestimates of α when they are used as the reference sites (Akashi 1995; see also http://www.f1000biology.com/article/id/1028663/dissent). Our results allow us to rule out these possibilities and suggest that, if anything, the estimates of adaptive evolution based on synonymous sites are slight underestimates. In the case of Andolfatto (2005), it should be noted that the vast majority of intronic sites that were analyzed came from long introns; only two introns were less than 65 bp (0.37% of intronic sites) and only four introns were less than 120 bp (0.91% of intronic sites). Thus, our finding that short introns (or portions thereof) show little signal of adaptive evolution does not contradict Andolfatto (2005) but rather suggests a difference between introns of different sizes. Indeed, when considering all sites of introns ≤120 bp, we obtain the largest point estimate of α (0.185; fig. 3), which is close to the value of 0.193 reported by Andolfatto (2005) for longer introns.

Despite the above advantages of short introns, in many cases, it may be preferable to use synonymous sites as the reference because they are present in all protein-coding genes, are relatively informative in comparison with short introns (or portions thereof), and are less prone to statistical errors in inferring positive selection at individual loci (Andolfatto 2008). It remains to be seen whether short introns are preferable to synonymous sites in species with more obvious evidence for selection on synonymous sites, such as *D. simulans* (Akashi 1995; Begun 1996; Haddrill et al. 2008).

Supplementary Material

Supplementary figure S1 and tables S1 and S2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank two anonymous reviewers for their constructive comments on the manuscript. This work was supported by *Deutsche Forschungsgemeinschaft* grant PA 903/4-1 to J.P. and US National Institutes of Health grant GM083228-01A2 to P.A.

References

Akashi H. 1994. Synonymous codon usage in *Drosophila melanogaster*: natural selection and translational accuracy. *Genetics* 136:927–935.

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

Andolfatto P. 2005. Adaptive evolution of non-coding DNA in Drosophila. *Nature* 437:1149–1152.

Andolfatto P. 2007. Hitchhiking effects of recurrent beneficial amino acid substitutions in the *Drosophila melanogaster* genome. *Genome Res.* 17:1755–1762.

Andolfatto P. 2008. Controlling type-I error of the McDonald-Kreitman test in genomewide scans for selection on noncoding DNA. *Genetics* 180:1767–1771.

Antezana MA, Kreitman M. 1999. The nonrandom location of synonymous codons suggests that reading frame-independent forces have patterned codon preferences. *J Mol Evol.* 49:36–43.

Asthana S, Noble WS, Kryukov G, Grant CE, Sunyaev S, Stamatoyannopoulos JA. 2007. Widely distributed noncoding purifying selection in the human genome. *Proc Natl Acad Sci U S A.* 104:12410–12415.

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

Baines JF, Parsch J, Stephan W. 2004. Pleiotropic effect of disrupting a conserved sequence involved in a long-range compensatory interaction in the Drosophila *Adh* gene. *Genetics* 66:237–242.

Baines JF, Sawyer SA, Hartl DL, Parsch J. 2008. Effects of X-linkage and sex-biased gene expression on the rate of adaptive protein evolution in *Drosophila*. *Mol Biol Evol.* 25:1639–1650.

Begun DJ. 1996. Population genetics of silent and replacement variation in *Drosophila simulans* and *D. melanogaster*: X/autosome differences? *Mol Biol Evol.* 13:1405–1407.

Bergman CM, Kreitman M. 2001. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* 11:1335–1345.

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

Bustamante CD, Nielsen R, Sawyer SA, Olsen KM, Purugganan MD, Hartl DL. 2002. The cost of inbreeding in *Arabidopsis*. *Nature* 416:531–534.

Carlini DB, Stephan W. 2003. *In vivo* introduction of unpreferred synonymous codons into the Drosophila *Adh* gene results in reduced levels of ADH protein. *Genetics* 163:239–243.

Carvalho AB, Clark AG. 1999. Intron size and natural selection. *Nature* 401:344.

Casillas S, Barbadilla A, Bergman CM. 2007. Purifying selection maintains highly conserved noncoding sequences in *Drosophila*. *Mol Biol Evol.* 24:2222–2234.

Charlesworth J, Eyre-Walker A. 2008. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol.* 25:1007–1015.

Comeron JM, Kreitman M. 2000. The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. *Genetics* 156: 1175–1190.

Deutsch M, Long M. 1999. Intron-exon structures of eukaryotic model organisms. *Nucleic Acids Res.* 27:3219–3228.

Dieringer D, Nolte V, Schlötterer C. 2005. Population structure in African *Drosophila melanogaster* revealed by microsatellite analysis. *Mol Ecol.* 14:563–573.

Drake JA, Bird C, Nemesh J, et al. (11 co-authors). 2006. Conserved noncoding sequences are selectively constrained and not mutation cold spots. *Nat Genet.* 38:223–227.

Drosophila 12 Genomes Consortium. 2007. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* 450:203–218.
Eyre-Walker A. 2002. Changing effective population size and the McDonald-Kreitman test. *Genetics* 162:2017–2024.
Eyre-Walker A, Keightley PD. 1999. High genomic deleterious mutation rates in hominids. *Nature* 397:344–347.
Fay JC, Wyckoff GJ, Wu CI. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.
Glinka S, Ometto L, Mousset S, Stephan W, De Lorenzo D. 2003. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* 165:1269–1278.
Haddrill PR, Bachtrog D, Andolfatto P. 2008. Positive and negative selection on noncoding DNA in *Drosophila simulans*. *Mol Biol Evol.* 25:1825–1834.
Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* 6:R67.
Haddrill PR, Halligan DL, Tomaras D, Charlesworth B. 2007. Reduced efficacy of selection in regions of the *Drosophila* genome that lack crossing over. *Genome Biol.* 8:R18.
Haddrill PR, Thornton KR, Charlesworth B, Andolfatto P. 2005. Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. *Genome Res.* 15:790–799.
Halligan DL, Keightley PD. 2006. Ubiquitous selective constraints in the *Drosophila* genome revealed by a genome-wide interspecies comparison. *Genome Res.* 16:875–884.
Hershberg R, Petrov DA. 2008. Selection on codon bias. *Annu Rev Genet.* 42:287–299.
Kohn MH, Fang S, Wu CI. 2004. Inference of positive and negative selection on the 5′ regulatory regions of *Drosophila* genes. *Mol Biol Evol.* 21:374–383.
Kondrashov AS, Crow JF. 1993. A molecular approach to estimating the human deleterious mutation rate. *Hum Mutat.* 2:229–234.
Marais G, Nouvellet P, Keightley PD, Charlesworth B. 2005. Intron size and exon evolution in Drosophila. *Genetics* 170:481–485.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* 351:652–654.
Meiklejohn CD, Kim Y, Hartl DL, Parsch J. 2004. Identification of a locus under complex positive selection in *Drosophila simulans* by haplotype mapping and composite-likelihood estimation. *Genetics* 168:265–279.
Mount SM, Burks C, Hertz G, Stormo GD, White O, Fields C. 1992. Splicing signals in *Drosophila*: intron size, information content, and consensus sequences. *Nucleic Acids Res.* 20:4255–4262.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol.* 3:418–426.
Ometto L, Stephan W, De Lorenzo D. 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* 169:1521–1527.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. *Mol Biol Evol.* 23:301–309.
Parsch J. 2003. Selective constraints on intron evolution in Drosophila. *Genetics* 165:1843–1851.
Parsch J, Tanda S, Stephan W. 1997. Site-directed mutations reveal long-range compensatory interactions in the *Adh* gene of *Drosophila melanogaster*. *Proc Natl Acad Sci U S A.* 94:928–933.
Parsch J, Zhang Z, Baines JF. 2009. The influence of demography and weak selection on the McDonald-Kreitman test: an empirical study in Drosophila. *Mol Biol Evol.* 26:691–698.
Presgraves DC. 2006. Intron length evolution in *Drosophila*. *Mol Biol Evol.* 23:2203–2213.
Pröschel M, Zhang Z, Parsch J. 2006. Widespread adaptive evolution of Drosophila genes with sex-biased expression. *Genetics* 174:893–900.
Ptak SE, Petrov DA. 2002. How intron splicing affects the deletion and insertion profile in *Drosophila melanogaster*. *Genetics* 162:1233–1244.
Sawyer SA, Parsch J, Zhang Z, Hartl DL. 2007. Prevalence of positive selection among nearly neutral amino acid replacements in *Drosophila*. *Proc Natl Acad Sci U S A.* 104:6504–6510.
Schaeffer SW. 2002. Molecular population genetics of sequence length diversity in the *Adh* region of *Drosophila pseudoobscura*. *Genet Res.* 80:163–175.
Shabalina SA, Kondrashov AS. 1999. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet Res.* 74:23–30.
Shabalina SA, Ogurtsov AY, Kondrashov VA, Kondrashov AS. 2001. Selective constraint in intergenic regions of human and mouse genomes. *Trends Genet.* 17:373–376.
Shapiro JA, Huang W, Zhang C, et al. (12 co-authors). 2007. Adaptive genic evolution in the *Drosophila* genomes. *Proc Natl Acad Sci U S A.* 104:2271–2276.
Siepel A, Bejerano G, Pedersen JS, et al. (16 co-authors). 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* 15:1034–1050.
Smith NG, Eyre-Walker A. 2002. Adaptive protein evolution in *Drosophila*. *Nature* 415:1022–1024.
Stephan W, Li H. 2006. The recent demographic and adaptive history of *Drosophila melanogaster*. *Heredity* 98:65–68.
Stephan W, Rodriguez VS, Zhou B, Parsch J. 1994. Molecular evolution of the metallothionein gene *Mtn* in the *melanogaster* species group: results from *Drosophila ananassae*. *Genetics* 138:135–143.
Stoletzki N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. *BMC Evol Biol.* 8:224.
Stoletzki N, Eyre-Walker A. 2007. Synonymous codon usage in *Escherichia coli*: selection for translational accuracy. *Mol Biol Evol.* 24:374–381.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.
Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–276.
Welch JJ. 2006. Estimating the genomewide rate of adaptive protein evolution in Drosophila. *Genetics* 173:821–837.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol.* 24:1586–1591.
Yang Z, Nielsen R, Goldman N, Pederson AM. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.
Yu J, Yang Z, Kibukawa M, Paddock M, Passey D, Wong GK. 2002. Minimal introns are not "junk." *Genome Res.* 12:1185–1189.