LETTERS

# Adaptive evolution of non-coding DNA in *Drosophila*

Peter Andolfatto[1]

**A large fraction of eukaryotic genomes consists of DNA that is not translated into protein sequence, and little is known about its functional significance. Here I show that several classes of non-coding DNA in *Drosophila* are evolving considerably slower than synonymous sites, and yet show an excess of between-species divergence relative to polymorphism when compared with synonymous sites. The former is a hallmark of selective constraint, but the latter is a signature of adaptive evolution, resembling general patterns of protein evolution in *Drosophila*[1,2]. I estimate that about 40–70% of nucleotides in intergenic regions, untranslated portions of mature mRNAs (UTRs) and most intronic DNA are evolutionarily constrained relative to synonymous sites. However, I also use an extension to the McDonald–Kreitman test[3] to show that a substantial fraction of the nucleotide divergence in these regions was driven to fixation by positive selection (about 20% for most intronic and intergenic DNA, and 60% for UTRs). On the basis of these observations, I suggest that a large fraction of the non-translated genome is functionally important and subject to both purifying selection and adaptive evolution. These results imply that, although positive selection is clearly an important facet of protein evolution, adaptive changes to non-coding DNA might have been considerably more common in the evolution of *D. melanogaster*.**

The high degree of protein sequence similarity between phenotypically diverged species has led some to propose that regulatory evolution may be of considerably more importance than protein evolution[4,5]. Although most of the typical eukaryotic genome is comprised of non-coding DNA, comparatively little is known about the evolutionary forces acting on it. Some unknown fraction of the non-translated genome is presumed to be crucial for the regulation of gene expression. Most of our direct knowledge regarding the evolution of regulatory elements comes from a handful of direct functional studies[5,6]. A second, indirect approach is based on comparative genomics[7]. The rationale for this second approach is that if newly arising mutations are typically detrimental to gene function, functionally important parts of the genome are expected to evolve more slowly than those lacking function[8–11].

There are some limitations to the comparative genomics approach. First, a given genomic region might be conserved owing simply to a lower mutation rate[12]. Second, known regulatory elements do not seem to be particularly well conserved as a class, at least in *Drosophila*[10]. This finding suggests that taking an approach based on sequence conservation alone may lead to a biased view of regulatory evolution. Functionality of DNA sequences implies that they can be subject to both negative and positive selection. If a significant fraction of divergence between species observed in non-coding DNA is positively selected rather than selectively neutral or constrained, this could lead to underestimates of the functional importance of non-coding DNA and cause researchers to overlook the contribution of arguably the most interesting class of mutations in genome evolution—those reflecting adaptive differences between populations and species.

These limitations can be overcome by combining comparative genomic analyses with population-level variability data[1–3,13]. To assess the mode of selection acting on non-coding DNA, I have analysed new and previously published polymorphism data for 35 coding fragments (average length 667 base pairs (bp)) and 153 non-coding fragments (average length 426 bp) scattered across the X chromosome of *D. melanogaster* (see Supplementary Materials 1). To estimate levels of between-species divergence, I have compared *D. melanogaster* with its closely related sibling species, *D. simulans*.

On the basis of the current *Drosophila* genome annotation (release 4), I separated the surveyed fragments into several categories that are likely to differ in the intensity and mode of selection acting on them (see Table 1). It is apparent that most non-coding DNA evolves considerably slower than synonymous sites (that is, sites in protein-coding sequences at which mutations do not result in amino acid substitutions; Table 1). This is the case for introns and UTRs (see also refs 14–16), as well as intergenic DNA, much of which is far from the closest known gene (see Supplementary Materials 1). I estimate levels of constraint in *Drosophila* non-coding DNA to be 40% for introns, 50% for intergenic regions (IGRs), and 60% for UTRs (Table 2). These are all considerably higher than previous estimates from a variety of species comparisons[11,15–18]. The non-coding DNA surveyed is also generally less polymorphic than synonymous sites in *D. melanogaster* (Table 1; $p < 10^{-10}$, Wilcoxon two-sample test for UTRs and introns+IGRs versus synonymous sites). Thus, both polymorphism and divergence in non-coding DNA are significantly reduced relative to synonymous sites in *D. melanogaster*.

Reduced levels of polymorphism and divergence in non-coding DNA resemble general patterns of protein evolution[19] and suggest that non-coding DNA is either functionally constrained or is subject to a lower mutation rate than synonymous sites. One way to distinguish between these two models is to consider the distribution of polymorphism frequencies. Negative selection acting on polymorphic variants will keep them at lower frequencies in a population than expected if they were neutral[20]. Consistent with this prediction, the distribution of polymorphism frequencies at both non-coding DNA and amino acid sites is skewed towards rare frequencies relative to synonymous polymorphisms (as indicated by a more negative Tajima's *D* value[20], Fig. 1). The distribution of Tajima's *D* values for non-synonymous sites among loci is negatively skewed relative to synonymous sites, suggesting that amino acid polymorphisms are subject to purifying selection (Fig. 1; $p = 0.002$, Wilcoxon two-sample test versus synonymous sites). Here I show that this same pattern extends to polymorphisms in non-coding DNA (Fig. 1; Wilcoxon test versus synonymous sites: pooled non-coding, $p = 0.0001$; UTRs, $p < 0.0001$; introns, $p = 0.001$; IGRs, $p = 0.005$). This finding, together with the observed reduction in polymorphism and divergence, implies that mutations in non-coding DNA are subject, on average, to stronger negative selection than synonymous sites (see also Supplementary Materials 2).

Does selective constraint alone account for patterns of non-coding DNA evolution? McDonald and Kreitman[3] have proposed a frame-

[1]Section of Ecology, Behavior and Evolution, Division of Biological Sciences, University of California San Diego, La Jolla, California 92093, USA.

**Table 1 | Polymorphism and divergence in coding and non-coding DNA of *D. melanogaster***

| Mutation class | No. of regions | Mean $\pi$* | Mean $D_{xy}$† | $D$‡ | $P$§ | $p$‖ | $P'$¶ | $p$# |
|---|---|---|---|---|---|---|---|---|
| Synonymous | 35 | 2.87 | 13.59 | 604 | 502 | – | 323 | – |
| Non-synonymous | 35 | 0.18 | 1.72 | 260 | 115 | $<10^{-6}$ | 52 | $<10^{-9}$ |
| Non-coding | 153 | 1.06 | 5.94 | 3,168 | 2,386 | 0.14 | 1,295 | $<10^{-3}$ |
| UTRs | 31 | 0.54 | 4.54 | 471 | 246 | $<10^{-5}$ | 107 | $<10^{-11}$ |
| 5′UTRs | 18 | 0.61 | 5.41 | 328 | 160 | $<10^{-5}$ | 71 | $<10^{-9}$ |
| 3′UTRs | 13 | 0.45 | 3.35 | 143 | 86 | 0.034 | 36 | $<10^{-4}$ |
| Introns | 72 | 1.25 | 6.71 | 1,564 | 1,221 | 0.39 | 675 | 0.010 |
| IGRs | 50 | 1.11 | 5.72 | 1,133 | 919 | >0.5 | 513 | 0.059 |
| pIGRs | 20 | 1.29 | 6.58 | 500 | 400 | >0.5 | 237 | 0.25 |
| dIGRs | 30 | 0.99 | 5.18 | 633 | 519 | >0.5 | 276 | 0.041 |
| Introns+IGR | 122 | 1.19 | 6.25 | 2,697 | 2,140 | 0.50 | 1,188 | 0.013 |

Mutation classes: synonymous sites, non-synonymous sites, untranslated transcribed regions (UTRs), intergenic regions within 2 kb of a gene (pIGRs), intergenic regions more than 4 kb away from a gene (dIGRs).
*$\pi$ is the weighted average within-species pairwise diversity per 100 sites.
†$D_{xy}$ is the weighted average pairwise divergence per 100 sites between *D. melanogaster* and *D. simulans*, corrected for multiple hits (Jukes–Cantor). $D_{xy}$ at fourfold degenerate synonymous sites is 12.0%.
‡$D$ is the estimated number of fixed differences between species using a Jukes–Cantor correction for multiple hits (see Methods).
§$P$ is the number of intraspecific polymorphisms.
‖McDonald–Kreitman test of probability using all polymorphisms.
¶$P'$ is the number of intraspecific polymorphisms excluding singletons.
#McDonald–Kreitman test of probability excluding singleton polymorphisms. Probabilities are from two-tailed Fisher's exact tests and assume sites are independent. These are likely to be only slight underestimates given probable levels of intragenic recombination (see Supplementary Materials 2).

work to distinguish neutrality (and variation in mutation rate) from negative and positive selection in the genome. Their approach compares levels of polymorphism within and divergence between species for a putatively selected class of sites in the genome to a neutral standard. If reduced levels of polymorphism and divergence in non-coding DNA can be explained by a lower mutation rate, the ratio of polymorphism to divergence should be similar to that for synonymous sites. Positive selection will increase divergence relative to polymorphism at selected sites, whereas negative selection is expected to result in the opposite pattern[21]. Although this framework was originally designed to detect selection within protein-coding genes, it can be generalized to consider arbitrary classes of putatively selected sites sampled from multiple genomic regions, including non-coding DNA (see Supplementary Materials 2). Using all polymorphisms, there is a significant excess of divergence for amino acid replacement sites ($p = 5 \times 10^{-7}$) and for UTRs ($p = 3 \times 10^{-6}$, two-tailed Fisher's exact test) but not at other subclasses of non-coding DNA (Table 1). This preliminary analysis suggests that, similar to the pattern observed for amino acid substitutions[1,2], a significant proportion of nucleotide divergence at UTRs was also driven to fixation by positive selection.

The presence of weakly negatively selected variants in polymorphism can mask the signature of adaptive evolution in the genome[1,22], making the McDonald–Kreitman test very conservative. As I have shown above that polymorphic variants in non-coding DNA are subject to stronger selective constraint than synonymous sites (Table 1 and Fig. 1), negatively selected variants contributing to polymorphism in non-coding DNA are likely to be a factor limiting power to detect positive selection. This problem can be partially overcome by considering only those mutations that are not rare in a sample from both the neutral and putatively selected classes (see ref. 23 and Supplementary Materials 2). Applying this approach reveals a significant excess of divergence in UTRs and in most other classes of non-coding DNA relative to synonymous sites (Table 1; UTRs, $p = 5 \times 10^{-12}$; introns, $p = 0.01$; dIGRs, $p = 0.04$; introns+IGRs, $p = 0.01$). A Hudson–Kreitman–Aguadé (HKA) test[24] also provides statistical support for a reduced ratio of polymorphism to divergence for non-coding DNA relative to synonymous sites (UTRs, $p < 10^{-3}$; pooled introns and IGRs, $p = 0.02$; see Supplementary Materials 2). Together, these results show that a significant fraction of the divergence in UTRs, introns and intergenic DNA was probably driven to fixation by positive selection.

To quantify the intensity and the relative importance of positive selection in shaping the evolution of non-coding DNA, I apply two extensions of the McDonald–Kreitman approach[2,13]. First I estimate $\alpha$, defined as the proportion of the divergence between species that was driven by positive selection[2]. I estimate that about 20% of the nucleotide divergence in introns and intergenic DNA was driven to fixation by positive selection, and about 60% for UTRs (Fig. 2a and Table 2). Using a hierarchical bayesian framework[13], I estimate the

**Table 2 | Functionally relevant nucleotides in non-coding DNA**

| Class | $C$ (%)* | $\alpha$(%)† | $p$ ($\alpha \leq 0$)‡ | FRN (%)§ |
|---|---|---|---|---|
| UTRs | 60.4 | 57.5 | $<10^{-3}$ | 83.2 |
| 5′UTRs | 52.9 | 60.8 | $<10^{-3}$ | 80.9 |
| 3′UTRs | 70.7 | 52.9 | $<10^{-3}$ | 86.2 |
| Introns | 39.5 | 19.3 | 0.007 | 51.2 |
| IGRs | 49.3 | 15.3 | 0.036 | 57.1 |
| pIGRs | 40.6 | 11.4 | 0.165 | 47.4 |
| dIGRs | 54.6 | 18.5 | 0.019 | 63.0 |
| Introns + IGR | 44.2 | 17.6 | 0.013 | 54.0 |

*Constraint ($C$) is estimated relative to fourfold degenerate synonymous sites.
†$\alpha$ is the estimated fraction of divergence driven by positive selection.
‡Probabilities ($\alpha \leq 0$) have been adjusted for effects of linkage within loci (see Supplementary Materials 2.5).
§FRN is the inferred fraction of functionally relevant nucleotides given levels of constraint and $\alpha$ (that is, FRN $\approx C + (1 - C)\alpha$).

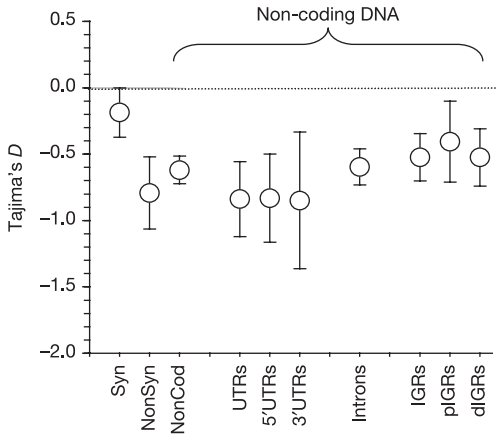


**Figure 1 | Mean Tajima's *D* values for coding and non-coding DNA.** Means across loci are given with bars indicating two standard errors. The expectation of $D$ under the neutral model is shown as a dotted line. Syn, synonymous sites; NonSyn, non-synonymous sites; NonCod, pooled non-coding DNA.

selection intensity on non-coding DNA (including UTRs, introns and IGRs) to be positive and significantly different from zero in most cases (Fig. 2b; Supplementary Materials 3). As this bayesian approach assumes that segregating and fixed variants are subject to the same direction and intensity of selection, it is likely to underestimate the magnitude of $2N_e s$ (the intensity of selection) for nucleotide substitutions fixed by positive selection (see Supplementary Materials 2).

Evidence that a significant fraction of non-coding DNA is functionally important is emerging from a variety of comparative genomic studies. However, my finding of a large fraction of positively selected divergence implies that 'evolutionary constraint' will substantially underestimate the fraction of functionally relevant nucleotides because it ignores the contribution of positively selected mutations to divergence. For the example of UTRs, I estimate evolutionary constraint to be 60%. However, as 58% of the observed divergence was positively selected, this implies that 83% of nucleotides in UTRs are in fact functionally relevant. Likewise, the fraction of functionally relevant nucleotides in introns and IGRs is likely to be about 10–20% higher than suggested by levels of constraint alone (Table 2).

How frequent is adaptation in the *Drosophila* genome? Rough calculations (see Supplementary Materials 4) suggest that there has been about one adaptive amino acid substitution every 20 years since the split of *D. melanogaster* and *D. simulans* (see also ref. 2). Although this is substantial, consider that the total number of sites contained in introns, intergenic regions and UTRs far outweighs the number of codons in the *Drosophila* genome[25]. I estimate that UTRs alone contribute as much to adaptive divergence between species as do amino acid changes, and the summed contribution of non-coding DNA to adaptive divergence could easily be an order of magnitude larger. These findings support previous intuitions[4,5] about the great importance of regulatory changes in evolution.

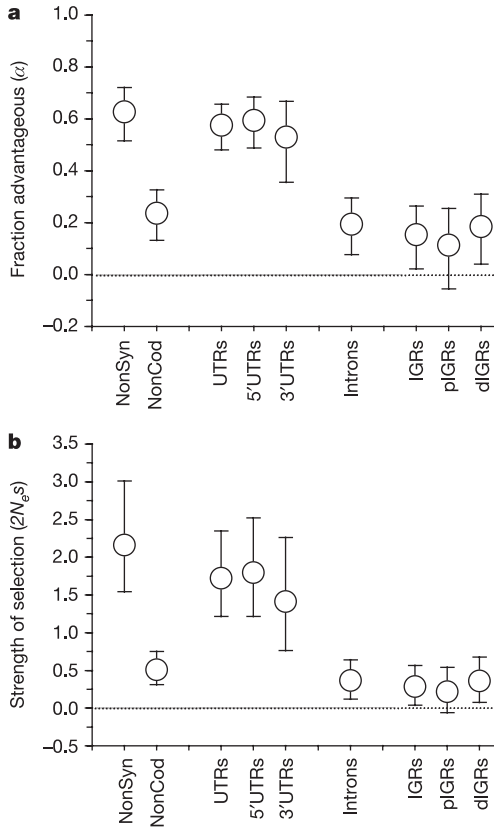


**Figure 2 | Quantifying adaptive divergence and selection intensity.** **a**, Estimates of $\alpha$, the fraction of nucleotide divergence driven by positive selection. Error bars indicate 90% confidence limits determined by a non-parametric bootstrapping. Estimated probabilities that $\alpha \geq 0$ corrected for partial linkage are given in Table 2. **b**, Estimates of the intensity of selection ($2N_e s$) acting on non-synonymous and non-coding DNA sites. Error bars indicate 90% confidence limits determined by simulation (see Methods). Singleton polymorphisms were excluded in estimates of $\alpha$ and $2N_e s$ (see Supplementary Materials 3). Abbreviations as in Fig. 1.

## METHODS

**Data.** All loci used in this study, previously published or newly collected, are X-linked genomic fragments, with a sample size of 12 *D. melanogaster* alleles sampled from a population in Zimbabwe, and a single *D. simulans* sequence. For coding DNA (synonymous and non-synonymous sites), I collected polymorphism and divergence in 31 coding regions selected randomly with respect to gene function, and 51 non-coding regions (27 intergenic and 24 untranslated transcribed regions). Information about these 82 loci and primers used can be found in Supplementary Materials 1. I used polymerase chain reaction (PCR) to amplify 700–800-bp regions from genomic DNA extracted from single male flies, removed primers and nucleotides using exonuclease I and shrimp alkaline phosphatase, and sequenced the cleaned product on both strands using Big-Dye (Version 3, Applied Biosystems). Sequences were collected on an ABI 3730 capillary sequencer and were aligned and edited using the program Sequencher (Gene Codes).

To the 82 regions surveyed above, I added previously published data for loci that had the same sample size ($n = 12$ flies) and were surveyed in similar samples from Zimbabwe[26,27]. A number of the previously published loci[26] had to be functionally reassigned when compared to Release 4 of the annotated *D. melanogaster* genome (http://flybase.bio.indiana.edu/annot/dmel-release4.html). I excluded any loci in regions of reduced recombination (see below). Previously published loci fitting these requirements were processed into 106 fragments (4 coding, 7 UTR, 23 intergenic and 72 intron). Thus, the total number of regions surveyed in this analysis is 188. Alignments for each locus are available upon request. A reciprocal best-hit BLAST protocol was used to confirm that the regions compared between *D. melanogaster* and *D. simulans* are indeed orthologous. Extra gaps were introduced into some alignments in regions that were particularly difficult to align. This procedure is likely to upwardly bias estimates of constraint, but is conservative with respect to detecting positive selection.

**Analyses.** The estimated number of synonymous sites, non-synonymous sites, average pairwise diversity ($\pi$), average pairwise divergence ($D_{xy}$), as well as counts of the number of polymorphic sites ($P$) were performed using DnaSP software (version 4; http://www.ub.es/dnasp/) and Perl code written by P.A. The number of divergent sites ($D$) was estimated as $D_{xy} - \pi$ using a Jukes–Cantor correction for multiple hits. Multiply hit sites were included in the analysis but insertion–deletion polymorphisms and mutations overlapping alignment gaps were excluded. Derived mutations were polarized using a single *D. simulans* sequence and assuming standard parsimony criteria. Tajima's $D$ value[20] was estimated from the number of polymorphisms and $\pi$.

In this study, I assume that synonymous sites are more neutral than putatively selected classes of sites (see Supplementary Materials 2.2). I separated non-coding DNA into subclasses that I expected *a priori* to experience different selection pressures: 5′ and 3′ untranslated transcribed regions (UTRs), introns, intergenic regions within 2 kilobases (kb) of a gene (proximal intergenic regions, pIGRs), and intergenic regions further than 4 kb from the nearest gene (distal intergenic regions, dIGRs). My sample of intron fragments is biased towards introns larger than the median intron size (86 bp) (ref. 28), making estimates of constraint higher than expected with a random sample of introns[14]. However, 95% of intronic DNA is contained within introns longer than the median size[28], and thus my estimate reflects levels of constraint for most intronic DNA in the genome.

For comparisons of polymorphism and divergence between synonymous sites and non-coding DNA, it was necessary to pool sites in each class. I estimate evolutionary constraint relative to fourfold degenerate synonymous sites using the approach in ref. 15, except that I pooled classes of sites and used a Jukes–Cantor correction for multiple hits[19]. Given differences in base composition between coding and non-coding regions, I investigated possible differences in mutations rates owing to the 16 possible adjacent-base contexts of nucleotides (suggested by A. Kondrashov). There was no significant effect of adjacent-base context on rates of divergence (see Supplementary Materials 5).

I estimate the proportion of divergence driven by positive selection[1,2] as $\alpha = 1 - (D_S P_X / D_X P_S)$, where S denotes synonymous (that is, putatively neutral) sites, X denotes putatively selected sites, and $D = \sum_{i=1}^{n} D_i$ and $P = \sum_{i=1}^{n} P_i$,

where $D_i$ and $P_i$ are the number of divergent and polymorphic variants at locus $i$, respectively, and $n$ is the number of loci of class S or X. Confidence limits on $\alpha$ were estimated using a standard non-parametric bootstrapping procedure, assuming sites are independent. The issue of non-independence of sites within surveyed fragments is addressed in Supplementary Materials 2.5. For consistency, $\alpha$ was estimated for non-synonymous sites in the same way. The intensity of selection ($2N_e s$) was estimated on putatively selected classes (pooling sites as above) using a hierarchical bayesian method (http://cbsuapps.tc.cornell.edu)[13]. To avoid problems associated with large-scale variation in recombination rates, I restricted my survey of loci to regions of the X chromosome that have the highest rates of recombination[29] (see Supplementary Fig. 1.1).

**Received 23 May; accepted 2 August 2005.**


1. Fay, J. C., Wyckoff, G. J. & Wu, C. I. Testing the neutral theory of molecular evolution with genomic data from *Drosophila*. *Nature* **415,** 1024–1026 (2002).
2. Smith, N. G. & Eyre-Walker, A. Adaptive protein evolution in *Drosophila*. *Nature* **415,** 1022–1024 (2002).
3. McDonald, J. & Kreitman, M. Adaptive protein evolution at the *Adh* locus in *Drosophila*. *Nature* **351,** 652–654 (1991).
4. King, M. C. & Wilson, A. C. Evolution at two levels in humans and chimpanzees. *Science* **188,** 107–116 (1975).
5. Carroll, S. B., Grenier, J. K. & Weatherbee, S. D. *From DNA to Diversity: Molecular Genetics and the Evolution of Animal Design* (Blackwell Science, Malden, Massachusetts, 2001).
6. Ludwig, M. *et al.* Functional evolution of a *cis*-regulatory module. *PLoS Biol.* **3,** e93 (2005).
7. Miller, W., Makova, K., Nekrutenko, A. & Hardison, R. Comparative genomics. *Annu. Rev. Genomics Hum. Genet.* **5,** 15–56 (2004).
8. Cliften, P. *et al.* Surveying *Saccharomyces* genomes to identify functional elements by comparative DNA sequence analysis. *Genome Res.* **11,** 1175–1186 (2001).
9. Gibbs, R. *et al.* Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
10. Richards, S. *et al.* Comparative genome sequencing of *Drosophila pseudoobscura*: chromosomal, gene, and *cis*-element evolution. *Genome Res.* **15,** 1–18 (2005).
11. Shabalina, S. & Kondrashov, A. Pattern of selective constraint in *C. elegans* and *C. briggsae* genomes. *Genet. Res.* **74,** 23–30 (1999).
12. Clark, A. The search for meaning in noncoding DNA. *Genome Res.* **11,** 1319–1320 (2001).
13. Bustamante, C. *et al.* The cost of inbreeding in *Arabidopsis*. *Nature* **416,** 531–534 (2002).
14. Haddrill, P. R., Halligan, D., Charlesworth, B. & Andolfatto, P. Patterns of intron sequence evolution in *Drosophila* are dependent upon length and GC content. *Genome Biol.* **6,** R67 (2005).
15. Halligan, D., Eyre-Walker, A., Andolfatto, P. & Keightley, P. Patterns of evolutionary constraints in intronic and intergenic DNA of *Drosophila*. *Genome Res.* **14,** 273–279 (2004).
16. Bachtrog, D. Sex chromosome evolution: molecular aspects of Y chromosome degeneration in *Drosophila*. *Genome Res.* **15,** 1393–1401 (2005).
17. Jareborg, N., Birney, E. & Durbin, R. Comparative analysis of noncoding regions of 77 orthologous mouse and human gene pairs. *Genome Res.* **9,** 815–824 (1999).
18. Bergman, C. & Kreitman, M. Analysis of conserved noncoding DNA in *Drosophila* reveals similar constraints in intergenic and intronic sequences. *Genome Res.* **11,** 1335–1345 (2001).
19. Li, W. *Molecular Evolution* (Sinauer Associates, Sunderland, Massachusetts, 1997).
20. Tajima, F. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123,** 585–595 (1989).
21. Kimura, M. *The Neutral Theory of Molecular Evolution* (Cambridge Univ. Press, Cambridge, 1983).
22. Charlesworth, B. The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genet. Res.* **63,** 213–227 (1994).
23. Templeton, A. Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. *Genetics* **144,** 1263–1270 (1996).
24. Hudson, R., Kreitman, M. & Aguadé, M. A test of neutral molecular evolution based on nucleotide data. *Genetics* **116,** 153–159 (1987).
25. Misra, S. *et al.* Annotation of the *Drosophila melanogaster* euchromatic genome: a systematic review. *Genome Biol.* **3,** research0083.1-0083.22 (2002).
26. Glinka, S., Ometto, L., Mousset, S., Stephan, W. & De Lorenzo, D. Demography and natural selection have shaped genetic variation in *Drosophila melanogaster*: a multi-locus approach. *Genetics* **165,** 1269–1278 (2003).
27. Haddrill, P. R., Thornton, K. R., Charlesworth, B. & Andolfatto, P. Multilocus patterns on nucleotide variability and the demographic and selection history of *Drosophila melanogaster* populations. *Genome Res.* **15,** 790–799 (2005).
28. Yu, J. *et al.* Minimal introns are not "junk". *Genome Res.* **12,** 1185–1189 (2002).
29. Charlesworth, B. Background selection and patterns of genetic diversity in *Drosophila melanogaster*. *Genet. Res.* **68,** 131–149 (1996).


**Supplementary Information** is linked to the online version of the paper at www.nature.com/nature.

**Acknowledgements** The author thanks D. Bachtrog for extensive comments on the manuscript and help with data quality issues, C. Bustamante and K. Thornton for providing code, and B. Ballard for Zimbabwe fly lines. P. Haddrill and K. Thornton assisted in designing primers for distal intergenic and coding regions, respectively. Thanks to B. Fischman for technical help, A. Betancourt, A. Kondrashov, A. Poon, D. Presgraves, M. Przeworski and S. Wright for critical comments on the manuscript, and L. Chao and J. Huelsenbeck for advice. Thanks also to the Washington University Genome Sequencing Center for providing unpublished *D. simulans* sequences. This work was funded in part by a research grant from the Biotechnology and Biological Sciences Research Council (UK) to P.A. The author is supported by an Alfred P. Sloan Fellowship in Molecular and Computational Biology.

**Author Information** Reprints and permissions information is available at npg.nature.com/reprintsandpermissions. The author declares no competing financial interests. Correspondence and requests for materials should be addressed to P.A. (pandolfatto@ucsd.edu).