Patterns of Evolutionary Constraints in Intronic and Intergenic DNA of Drosophila

Citation for published version:
https://doi.org/10.1101/gr.1329204

Digital Object Identifier (DOI):
10.1101/gr.1329204

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Published In:
Genome Research

Publisher Rights Statement:
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.1329204.

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
Patterns of Evolutionary Constraints in Intronic and Intergenic DNA of Drosophila

Daniel L. Halligan, Adam Eyre-Walker, Peter Andolfatto, and Peter D. Keightley

We develop methods to infer levels of evolutionary constraints in the genome by comparing rates of nucleotide substitution in noncoding DNA with rates predicted from rates of synonymous site evolution in adjacent genes or other putatively neutrally evolving sites, while accounting for differences in base composition. We apply the methods to estimate levels of constraint in noncoding DNA of Drosophila. In introns, constraint (the estimated fraction of mutations that are selectively eliminated) is absolute at the 5’ and 3’ splice junction dinucleotides, and averages 72% in base pairs 3–6 at the 5’-end. Constraint at the 5’ base pairs 3–6 is significantly lower in the lineage leading to Drosophila melanogaster than in Drosophila simulans, a finding that agrees with other features of genome evolution in Drosophila and indicates that the effect of selection on intron function has been weaker in the melanogaster lineage. Elsewhere in intron sequences, the rate of nucleotide substitution is significantly higher than at synonymous sites. By using intronic sites outside splice control regions as a putative neutrally evolving standard, constraint in the 500 bp of intergenic DNA upstream and downstream regions of protein-coding genes averages ~44%. Although the estimated level of constraint in intergenic regions close to genes is only about one-half of that of amino acid sites, selection against single-nucleotide mutations in intergenic DNA makes a substantial contribution to the mutation load in Drosophila.

[Supplemental material is available online at www.genome.org. The sequence data from this study have been submitted to GenBank under accession nos. AY459538–AY459582.]

Understanding the functional significance of intronic and intergenic noncoding DNA sequences is one of the major challenges in genomics research at present. If functional elements of the genome are close to adaptive optima owing to past directional selection, these sequences are expected to show evidence of purifying selection. This manifests itself as a lower rate of between-species nucleotide substitution when comparisons are made with evolutionary rates in neutrally evolving DNA segments having similar base composition and mutation rates. The level of functional conservation in the genome is important in determining the genome-wide mutation load due to the selective elimination of deleterious mutations (Kondrashov 1995), and this affects several important evolutionary issues (Charlesworth and Charlesworth 1998). Although it is well established that most protein-coding sequences are strongly constrained, that is, that most amino acid altering mutations are deleterious and become selectively eliminated (e.g., Li 1997), functional conservation in noncoding DNA has been much less well studied and is subject to controversy. Although some introns contain regulatory elements, several comparative studies suggest that introns evolve largely free from selective constraints (Gilbert 1978; Li and Graur 1991; Li 1997). However, recent genome-wide interspecific comparisons imply that intron sequences are subject to significant evolutionary pressures (Jareborg et al. 1999; Shabalina and Kondrashov 1999; Bergman and Kreitman 2001; Shabalina et al. 2001), and comparisons of known or putative regulatory elements (Ludwig and Kreitman 1995; Glazko et al. 2003) have also revealed substantial constraints, but the overall level of conservation and the distribution of conserved elements in intergenic regions of the genome is still largely unknown.

Present methods to quantify functional constraints in DNA sequences mostly depend on comparative genomics approaches. They relate to a method for inferring the genome-wide deleterious mutation rate based on sequence divergence (Kondrashov and Crow 1993). Shabalina and Kondrashov (1999) proposed that the proportion of bases that are subject to strong purifying selection can be quantified by comparing the genomes of distantly related species. It is assumed that homologous segments lacking similarity are saturated with nucleotide and/or indel substitutions, and are evolving free from functional constraint, whereas segments showing similarity ("hits") are under strong functional constraint. Constraint is quantified as the fraction of conserved nucleotides in the hits, which is assumed to comprise bases under strong purifying selection. Potential difficulties with the approach are variation across the genome in the mutation rate, which could make nonfunctional elements appear functional (Clark 2001), and obtaining the correct (or most probable) sequence alignment. If the DNA sequence alignment method is heuristic, and, for example, genuine similarities are missed, then functional elements could appear nonfunctional.

A second general approach for quantifying evolutionary constraint also uses comparisons between DNA segments of re-
lated species, but uses sequences from species showing lower levels of divergence that are far from saturation. It is based on comparing the rate of evolution of a putative functional segment of noncoding DNA with the rate of evolution of a DNA segment or a category of nucleotide sites that is assumed to be evolving free from constraint (a neutral segment), that has the same mutation rate, and can therefore act as a standard. Constraint is the factor by which evolution is slowed down in the functional segment (Kimura 1983). Nucleotides are assumed to fall into two classes in the functional sequence: neutral, which evolve at the same rate as the neutral sequence; or strongly constrained, in which mutations are eliminated unconditionally by natural selection. The neutral segment should be adjacent to the functional segment, thereby making the assumption of equality of mutation rates defensible. A close to ideal situation would be to compare the rate of evolution of a pseudogene (assumed to completely lack function) to that of an adjoining noncoding DNA segment. Unfortunately, because in many taxa, including Drosophila, pseudogenes are uncommon (Petrov et al. 1996), an alternative category of neutrally evolving sequences is needed. A candidate for such a category is synonymous sites of genes, because changes in these do not lead to change in the amino acid sequence. In many taxa, including Drosophila, however, there is evidence of past selection acting on synonymous codon usage (Shields et al. 1988), and this could retard rates of evolution at synonymous sites (Li 1997, Chapter 7).

In this paper, our initial approach is to use synonymous sites of Drosophila protein-coding genes as a standard for estimation of constraint in adjacent noncoding DNA. Features of the dynamics of synonymous site evolution indicate that selection at synonymous sites is weak or, in some cases, absent in Drosophila. In the lineage leading to Drosophila melanogaster from its common ancestor with Drosophila simulans, there has been a surge in the rate of preferred to unpreferred synonymous substitutions (Akashi 1995, 1996; McVean and Vieira 2001). This surge in the rate of substitution indicates a genome-wide relaxation of selection at synonymous sites, possibly because of demographic changes that have changed the efficiency of natural selection. In D. melanogaster, a population genetics analysis of the pattern of synonymous site divergence indicates that selection has been relaxed to the point of being completely absent (McVean and Vieira 2001). Further evidence for low levels of selection presently acting on D. melanogaster synonymous codon usage comes from an analysis of the frequency spectrum of segregating synonymous sites (Akashi 1999). A weakening of selection to approximately one-half of that in the ancestral species is estimated to have occurred in the D. simulans lineage (McVean and Vieira 2001). Furthermore, weak selection of the magnitude thought to be acting on synonymous codon usage in Drosophila (Akashi 1995, 1996) is predicted to have only a small effect on substitution rates (Eyre-Walker and Bulmer 1995). Recently, an apparent excess of preferred to unpreferred synonymous site substitutions has been reported in the Xdh gene of several Drosophila species (Begun and Whitley 2002). Possible explanations for this observation are an evolutionary shift in base composition towards A/T nucleotides in these in many Drosophila lineages (the explanation favored by Begun and Whitley 2002; see also Duret et al. 2002), a general weakening of selection in Drosophila lineages, or an artifact of parsimony if nucleotide mutation rates are sufficiently variable.

There are two additional difficulties in interpreting comparisons between synonymous site or other putatively neutral site divergence and nucleotide divergence in noncoding DNA. First, differences in the rate of substitution can be induced by differences in base composition; this stems from variation in average mutation rates between different kinds of nucleotides. We address this by comparing expected and observed numbers of substitutions; expected numbers in a noncoding segment are predicted on the basis of substitution rates at synonymous or other putatively neutral sites of adjacent genes, after the compositional effect has been accounted for.

A second potential problem in analyzing evolutionary rates in noncoding DNA concerns inference of the correct sequence alignment. Consider two alternative plausible alignments of a pair of sequences containing at least one gap:

<table>
<thead>
<tr>
<th>Alignment 1</th>
<th>Alignment 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>ATGCATGCG</td>
<td>ATGCACTGCG</td>
</tr>
<tr>
<td>AT–GAGCA</td>
<td>AT–CA–GCA</td>
</tr>
</tbody>
</table>

If alignment 1 were taken as the true alignment, the fraction of nucleotide differences (k = 3/7) would be radically different from taking alignment 2 (k = 1/7). The uncertainty is due to the unknown pattern of indels (gaps) between the sequences. In general, putative alignments containing too many gaps relative to the true alignment tend to have too few nucleotide substitutions or vice versa, and the bias can be serious. A solution to this problem has been proposed by Thorne et al. (1991, 1992), who developed an algorithm to compute probabilities of alternative alignments according to explicit models of indel evolution. Here, we use a Monte Carlo approach, MCALIGN, to tackle the problem of aligning noncoding DNA (P.D. Keightley and T. Johnson, unpubl.). Noncoding DNA sequences are aligned according to a model of indel evolution that is parameterized by relative rates of indels and nucleotide substitutions in noncoding DNA of closely related Drosophila species, and the most probable alignment is used in the subsequent analysis.

**RESULTS**

**Indel Frequencies in Noncoding DNA**

Between D. simulans and D. sechellia and Parameterization of Alignment Models

To investigate the frequency distribution of indels, and to parameterize models of indel evolution suitable for aligning Drosophila noncoding DNA, we compiled intronic and intergenic DNA sequences from homologous loci of D. simulans and D. sechellia (Table 1). These species were chosen because they are part of the melanogaster subgroup, and are sufficiently closely related as to make alignment of noncoding DNA by standard heuristic methods virtually unambiguous. The frequency distribution of indel length in three DNA categories is shown in Figure 1 (two long intronic indels of length 29 and 37 are omitted to aid clarity). Distributions of indel length are not dissimilar to geometric distributions, as has been suggested previously (Gu and Li 1995). Numbers of substitutions do not differ significantly between DNA categories: A likelihood ratio test for heterogeneity among indel rates relative to substitution rates is nonsignificant (Table 1;
Evolution of Noncoding DNA

Evolutionary Conservation of Intergenic DNA Sequences of Drosophila

We computed estimates of the level of constraint in D. melanogaster/simulans intron sequences by the two lineage approach, using synonymous sites as the putatively neutral standard, as described in Methods, under the assumption that the equilibrium G + C content \( f_G \) is equal to the G + C content of intron sequences in our data set (0.37). Separate estimates were made for complete intron sequences, and for intron sequences stripped of putative 5’- and 3’-splice control sequences (Table 2). Estimates of constraint in intron sequences are negative in sequences either including \((P = 0.1)\) or excluding \((P = 0.007)\) splice control sequences. Negative estimates of constraint imply that fourfold sites evolve more slowly than intronic sites (particularly those sites that are outside splice control regions), after differences in base composition have been accounted for. We investigated the slightly higher constraint in sequences including putative splice control sequences by calculating constraints for groups of bases at the 5’- and 3’-ends of intron sequences (Table 3). In the sequences analyzed, conservation is absolute for the 5’-splice-junction GT dinucleotide, and constraint is also strong for base pairs 3–6 at the 5’-end; all these bases are important in delimiting the 5’-end of introns (Sharp 1994). With the exception of the invariant 3’-splice-junction AT dinucleotide, constraint is non-significant at the 3’-end, a somewhat surprising result, considering that the consensus for nucleotides 5–16 is a run of pyrimidines showing moderate conservation across eukaryotic lineages (Sharp 1994). Constraint in intronic splice sequences, calculated using non-splice-control intronic sequences as the putatively neutral standard, gives similar results to those shown in Table 3 (i.e., 5’ bp 3–6: \( C = 0.713 \), SEM = 0.0484; 3’ bp 3–16: \( C = 0.0517 \), SEM = 0.0746).

We also calculated constraint specific to the D. melanogaster and D. simulans lineages by the three-lineage approach described in Methods, using Drosophila yakuba sequences as the outgroup. The results are consistent for those obtained for the two-sequence method, and suggest weak constraint in intron sequences lacking putative splice sites in both species. The level of constraint is significantly lower for bases 3–6 of the 5’-end in melanogaster than simulans \((P = 0.048)\); bootstrap analysis, Table 4), lending support to the idea of a lower intensity of selection in the lineage leading to D. melanogaster. The constraint difference at the 3’-end (Table 4) is nonsignificant \((P > 0.8)\).

Table 2. Observed and Expected Numbers of Nucleotide Substitutions Along With Estimates for Constraint in Noncoding DNA Sequences of Three Categories, Calculated by the Two Lineage Approach Using Four-Fold Sites of Homologous Genes from D. simulans and D. melanogaster as a Standard

<table>
<thead>
<tr>
<th>DNA category</th>
<th>Data set</th>
<th>Number of loci</th>
<th>Base pairs per locus</th>
<th>Observed (SEM)</th>
<th>Expected (SEM)</th>
<th>Constraint (SEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intron</td>
<td>Complete</td>
<td>91</td>
<td>228</td>
<td>16.23 (1.69)</td>
<td>14.86 (1.56)</td>
<td>−0.094 (0.059)</td>
</tr>
<tr>
<td>Intron</td>
<td>Splice sequences omitted</td>
<td>91</td>
<td>190</td>
<td>13.72 (1.63)</td>
<td>11.75 (1.45)</td>
<td>−0.17 (0.069)</td>
</tr>
<tr>
<td>5’ intergenic</td>
<td>bp 1–100</td>
<td>77</td>
<td>99</td>
<td>5.43 (0.36)</td>
<td>8.29 (0.46)</td>
<td>0.34 (0.033)</td>
</tr>
<tr>
<td>5’ intergenic</td>
<td>bp 101–200</td>
<td>73</td>
<td>98</td>
<td>5.58 (0.45)</td>
<td>7.97 (0.46)</td>
<td>0.30 (0.056)</td>
</tr>
<tr>
<td>5’ intergenic</td>
<td>bp 201–300</td>
<td>69</td>
<td>188</td>
<td>6.51 (0.53)</td>
<td>7.82 (0.49)</td>
<td>0.16 (0.079)</td>
</tr>
<tr>
<td>5’ intergenic</td>
<td>bp 301–500</td>
<td>66</td>
<td>188</td>
<td>13.37 (1.06)</td>
<td>13.86 (0.86)</td>
<td>0.03 (0.088)</td>
</tr>
<tr>
<td>3’ intergenic</td>
<td>bp 1–100</td>
<td>42</td>
<td>95</td>
<td>5.78 (0.69)</td>
<td>7.29 (0.54)</td>
<td>0.21 (0.086)</td>
</tr>
<tr>
<td>3’ intergenic</td>
<td>bp 101–200</td>
<td>31</td>
<td>88</td>
<td>5.35 (0.77)</td>
<td>6.58 (0.83)</td>
<td>0.17 (0.14)</td>
</tr>
<tr>
<td>3’ intergenic</td>
<td>bp 201–300</td>
<td>21</td>
<td>92</td>
<td>5.91 (0.79)</td>
<td>6.65 (1.06)</td>
<td>0.093 (0.17)</td>
</tr>
<tr>
<td>3’ intergenic</td>
<td>bp 301–500</td>
<td>18</td>
<td>150</td>
<td>7.10 (1.30)</td>
<td>11.60 (2.13)</td>
<td>0.36 (0.19)</td>
</tr>
</tbody>
</table>
findings agree with Duret and Mouchiroud (2000), who found a negative correlation between gene expression level and substitution rate in 5′- and 3′-untranslated regions of mammals, indicating the operation of purifying selection.

**DISCUSSION**

In contrast to intron sequences in mammals and several other taxa (International Human Genome Sequencing Consortium 2001; Mouse Genome Sequencing Consortium 2002), intron sequences tend to be rather short in *Drosophila*, with a peak length of only ~60 bp (Adams et al. 2000), and several studies have revealed precursor mRNA secondary structure in intronic sequences (Stephan and Kirby 1993; Kirby et al. 1995; Leicht et al. 1995). We therefore expected that constraints would be easily observed in *Drosophila*, if introns commonly contain gene expression control sequences.

The analysis did not bear this expectation out. The results are consistent with somewhat faster evolution at most intronic sites than fourfold sites, which themselves are thought to be under weak selection. Under the assumption of a nonequilibrium model of sequence evolution, our analysis indicates that intronic sequences outside splice control sequences evolve ~17% faster than fourfold sites of adjacent genes. Similar findings have recently been reported in primates (Chen and Li 2001; Hellmann et al. 2003) and rodents (Keightley and Gaffney 2003), although a different study in primates did not reveal the pattern (Subramanian and Kumar 2003). We examined the robustness of the result by changing \( f_e \), the equilibrium GC content. Values of \( f_e \) below 0.53 give higher rates of evolution at intronic sites than fourfold sites, and for values below 0.41 the difference in rates (as measured by constraint) is significant at \( P < 0.05 \). In introns, moderate to strong constraint was only detected between *melanogaster* and *simulans* at the dixone-intron boundaries and at 5′ nucleotides 3–6; these latter nucleotides also show a notable difference in constraint between the two lineages, possibly brought about by a difference in the long-term effectiveness of selection between the species (Aquadro et al. 1988; Akashi 1995; Moriyama and Powell 1996; Andolfatto 2001; Eyre-Walker et al. 2002). This pattern of constraint close to intron boundaries implies that mutations at these sites are slightly deleterious (Ohta 1992), and is therefore indirect evidence that the remaining sequences are genuinely evolving free from selective constraints. There is little difference between expected and observed numbers of substitutions between *melanogaster* and *simulans* (Table 4), an observation that is also consistent with models of weak selection, because selection coefficients of the order of the reciprocal of \( N_e \) are predicted to have little influence on substitution rates. The present findings are in broad agreement with McVean and Vieira (2001), who found that predicted rates of substitution were similar to observed rates in *Drosophila*. Our results concord with observations of the density of nucleotide polymorphisms in human introns as a function of distance from the 5′- or 3′-end (F.A. Kondrashov pers. comm.); there was no evidence for selection operating beyond about nucleotide 10 from the 5′- or 3′-end.

Our results contrast with recent estimates of the levels of constraint in introns and intergenic DNA in *Drosophila* (Bergman and Kreitman 2001) and *Caenorhabditis* (Shabalina and Kondrashov 1999), in which constraint was calculated from the fraction of conserved nucleotides in alignable blocks of DNA between distantly related species. Surprisingly, frequencies of conserved blocks in introns and intergenic DNA were similar to each other (of the order of 20%). However, variability in the mutation rate from region to region (Clark 2001) could give the false impression of evolutionary conservation in a segment that is evolving at the neutral rate. Furthermore, alignment of noncoding DNA is problematic with widely diverged species. Any noncoding DNA alignment that is not based on a model of indel evolution is likely to be biased (Thornc et al. 1991), and estimates of numbers of nucleotide substitutions may either be too high or too low depending on whether the alignment algorithm inserts too few or too many indels. Estimates of the proportion of conserved blocks in noncoding regions between mouse and human (Jareborg et al. 1999) are also likely to be susceptible to such biases.

The data in Table 3 indicate that the number of constrained nucleotides per intron is ~4.1. If there are 41,000 introns in the *Drosophila* genome (Adams et al. 2000), the predicted number of constrained nucleotides in introns is therefore only 0.17 Mb. The level of constraint at amino acid sites of *Drosophila* genes has been estimated to be ~84% (Eyre-Walker et al. 2002), implying that the total number of constrained amino acid sites in the *Drosophila* genome is ~16 Mb (~14,000 protein-coding genes), comprising an average of 591 codons [Adams et al. 2000], and about three-quarters of sites in coding DNA lead to an amino acid change if mutated. The number of constrained nucleotides in introns is therefore relatively small in relation to the protein-coding segment of the genome. However, the number of constrained nucleotides in intergenic DNA could potentially be of the same order as in coding DNA. For example, if we assume the average constraint values calculated relative to intronic sequences, we

---

**Table 3.** Estimates of Constraint in Runs of Nucleotides Close to the 5′- or 3′-Ends of Introns, Calculated Using the Two Lineage Method, Applied to Sequences from *D. melanogaster* and *D. simulans*.

<table>
<thead>
<tr>
<th>Intron end</th>
<th>Base pairs</th>
<th>Observed (SEM)</th>
<th>Expected (SEM)</th>
<th>Constraint (SEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5′</td>
<td>1–2</td>
<td>0 (0.0)</td>
<td>0.294 (0.021)</td>
<td>1.00 (0.00)</td>
</tr>
<tr>
<td></td>
<td>3–4</td>
<td>0.145 (0.037)</td>
<td>0.277 (0.021)</td>
<td>0.47 (0.13)</td>
</tr>
<tr>
<td></td>
<td>5–6</td>
<td>0.055 (0.024)</td>
<td>0.295 (0.022)</td>
<td>0.82 (0.077)</td>
</tr>
<tr>
<td>3′</td>
<td>1–2</td>
<td>0 (0.0)</td>
<td>0.292 (0.021)</td>
<td>1.00 (0.00)</td>
</tr>
<tr>
<td></td>
<td>3–16</td>
<td>2.248 (0.202)</td>
<td>1.912 (0.142)</td>
<td>–0.178 (0.098)</td>
</tr>
</tbody>
</table>

---

**Table 4.** Estimates of the Level of Constraint in Introns of 38 Loci From *D. melanogaster* and *D. simulans*, Computed Using the Three-Lineage Method.

<table>
<thead>
<tr>
<th>Lineage</th>
<th>Data set</th>
<th>bp Per locus</th>
<th>Observed (SEM)</th>
<th>Expected (SEM)</th>
<th>Constraint (SEM)</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>melanogaster</em></td>
<td>Complete</td>
<td>204</td>
<td>7.05 (1.07)</td>
<td>7.17 (1.15)</td>
<td>0.010 (0.11)</td>
</tr>
<tr>
<td><em>simulans</em></td>
<td></td>
<td></td>
<td>6.49 (1.04)</td>
<td>7.82 (1.59)</td>
<td>0.15 (0.14)</td>
</tr>
<tr>
<td><em>melanogaster</em></td>
<td>Splice sequences omitted</td>
<td>161</td>
<td>5.87 (0.99)</td>
<td>5.45 (0.96)</td>
<td>–0.087 (0.13)</td>
</tr>
<tr>
<td><em>simulans</em></td>
<td></td>
<td></td>
<td>5.39 (0.95)</td>
<td>5.90 (1.36)</td>
<td>0.065 (0.15)</td>
</tr>
<tr>
<td><em>melanogaster</em></td>
<td>5′ bases 3–6</td>
<td>7.8</td>
<td>0.184 (0.074)</td>
<td>0.326 (0.045)</td>
<td>0.44 (0.21)</td>
</tr>
<tr>
<td><em>simulans</em></td>
<td></td>
<td></td>
<td>0.052 (0.036)</td>
<td>0.370 (0.057)</td>
<td>0.86 (0.10)</td>
</tr>
<tr>
<td><em>melanogaster</em></td>
<td>3′ bases 3–16</td>
<td>27.3</td>
<td>1.003 (0.200)</td>
<td>1.112 (0.196)</td>
<td>0.084 (0.19)</td>
</tr>
<tr>
<td><em>simulans</em></td>
<td></td>
<td></td>
<td>1.034 (0.200)</td>
<td>1.205 (0.201)</td>
<td>0.12 (0.21)</td>
</tr>
</tbody>
</table>
obtain 14,000 genes $\times 1000 \text{ bp} \times 0.44 = 6.2 \text{ Mb}$. This is a minimum estimate, whose value could be much larger if there are appreciable functional constraints deep in the intergenic DNA.

**METHODS**

**Data**

Homologous gene sequences (partial or complete) from *D. simulans* and *D. melanogaster*, and, where available *D. yakuba*, were compiled from GenBank. Genes were selected if they contained at least one intron, or at least 60 bp of intergenic DNA upstream or downstream for the start or stop codon. Coding sequences were aligned using CLUSTAL (Thompson et al. 1994) and corrected manually. Noncoding sequences were aligned using MALIGN (P.D. Keightley and T. Johnson, unpubl.), a procedure that attempts to find the most probable alignment according to specific models of indel evolution (see below). The parameters of the alignment model were derived from data on relative rates of indels and nucleotide substitutions between *D. simulans* and *D. sechellia* (Table 1), and the frequency distribution of indel lengths (Fig. 1). Noncoding DNA alignments of *D. simulans* and *D. sechellia* are virtually unambiguous, by alignment with any standard heuristic alignment method. The most probable alignment of the *D. simulans*/melanogaster/yakuba sequences were used in subsequent analysis. Intergenic DNA was categorized either as 5′ or 3′. In cases in which genes are so close together in the genome that this categorization was ambiguous, stretches of DNA were arbitrarily assigned to the 5′ category, although they could have been considered to belong to the 3′-segment of an adjacent gene. Intergenic DNA includes any DNA that is 5′ or 3′ from the start or stop codon, and therefore contains elements of transcribed untranslated DNA.

Introns were either analyzed as complete sequences or partial sequences after removal of putative splice control sequences. Coding sequences after removal of putative splice control sequences.

**Sequencing of Additional Drosophila simulans**

**Intergenic Sequences**

We obtained additional intergenic DNA sequences from *Drosophila simulans* by sequencing the 5′-flanking regions of genes for which the orthologous coding sequences were available for both *simulans* and *melanogaster* on GenBank. Genes for which there was only a short length of available coding sequence in *simulans* were excluded (we used an arbitrary cutoff of 200 bp), and we did not sequence upstream DNA from previously sequenced *simulans* loci. Primers for sequencing were designed (using Primer Premier 5.00; Premier Biosoft International) to be ~650 to 700 bp apart, based on the *melanogaster* sequence. Upstream primers were usually designed from the noncoding *melanogaster* sequence (where possible an upstream coding sequence was used), and downstream primers were designed using the *simulans* coding sequence.

Genomic DNA for PCR reactions was prepared (Gentra Systems, Research Triangle Park) from a single partially inbred male *Drosophila simulans* fly collected in Aswan, Egypt in 2001. A single male fly was used as a source of DNA in all cases to reduce sequencing problems associated with heterogeneity in template DNA. A combination of standard PCR and asymmetric PCR (Miller et al. 2003) was used to amplify the appropriate section of DNA, if the primers failed to amplify the appropriate section of DNA, the primers were redesigned. If the appropriate DNA segment still could not be amplified after the primers had been redesigned three times, investigation of the gene was terminated. In 18 cases out of 63, we could not obtain sufficient amplification of the appropriate section of DNA.

Purified PCR products were sequenced on both strands using an ABI prism BigDye terminator cycle sequencing kit (Applied Biosystems) and run on an Applied Biosystems 3730 DNA Analyzer (Applied Biosystems). Sequences from each strand for each gene were then assembled using Sequencer 3.0 software (Gene Codes), and alignments were checked manually. The GenBank accession numbers are AY459568–AY459582.

**Likelihood Ratio Test for Variation in Rates of Indels Relative to Nucleotide Substitutions**

The test was constructed under the assumption that sequences are sufficiently closely related such that multiple hits can be ignored, and that the number of indels is linearly related to the number of nucleotide substitutions. Assume that there are $n$ categories of noncoding DNA (say, $n = 3$ with 1 = intronic, 2 = 5′-intergenic, and 3 = 3′-intergenic). Let $\alpha_i$ be a parameter for the fraction of nucleotide differences between sequences of category $i$, and $\theta_i$ be a compound parameter for the fraction of indels differentiating sequences in category $i$. Under the assumption of independent binomially distributed nucleotide substitution and indel numbers, the likelihood of observing $n_i$ substitutions and $m_i$ indels is

$$L_i = k^n_i (1 - k)^{m_i} (\theta_i)^{n_i} (1 - \theta_i)^{m_i},$$

where $k_i$ is the number of base pairs in the sequence (excluding indels) and $m_i$ is the number of sites not occupied by an indel. The likelihood of the observations of three categories of DNA is

$$L = L_1 \times L_2 \times L_3,$$

and the two models are compared according to

$$\theta_1 = \theta_2 = \theta_3 (\text{full model}), \text{and } \theta_1 = \theta_2 = \theta_3 (\text{reduced model}).$$

The likelihood with respect to $k$ and $\theta$ was maximized numerically.

**Alignment of Noncoding DNA Sequences According to Models of Indel Evolution**

Alignment was carried out by a procedure MALIGN, available at PDK’s Web page. The procedure uses a Monte Carlo algorithm to search for the most probable alignment of a pair of sequences or of three sequences that includes an outgroup, based on a model of indel evolution. The parameters of the model are $\theta$, the rate of indels relative to nucleotide substitutions, and a vector parameter $w$ specifying the frequency distribution of indel lengths. Because $\theta$ is a parameter of the model, estimated alignments containing large (small) numbers of nucleotide differences tend to have large (small) numbers of indels, a pattern supported by mouse–human sequence alignments (Hardison et al. 2003). In aligning pairs of sequences, the Jukes-Cantor method is used to correct for multiple nucleotide substitutions. For three sequences, parsimony is used to assign substitutions and indel events to the ingroup or the outgroup, and the probability of the alignment is the product of probabilities for the ingroup and outgroup.

The model parameters $\theta$ and $w$ come from external data, in the present case from alignments of noncoding DNA of *D. simulans* and *D. sechellia*. Values of $\theta$ from Table 1 were used to parameterize three alternative models, for aligning intronic, 3′-intergenic, or 5′-intergenic DNA. The vector parameter $w$ was assumed to be the same for each model, and was derived from the frequency distribution of indels in introns, after some smoothing of the distribution had been applied.

**Two-Lineage Approach to Compute Constraint in Noncoding DNA**

Following distance-based methods for calculating constraint in coding DNA (Eyre-Walker and Keightley 1999), the present method uses rates of substitution at fourfold sites or other putative neutral sites of a gene to predict expected numbers of substitutions in an adjacent noncoding DNA segment, such as an intron or flanking sequence, assuming equal rates of mutation in the sequences. The method takes into account differences in base composition. The expected numbers of substitutions ($E$) are compared with the observed numbers ($O$) to calculate constraint ($C$). For example, if $E = O$, the constraint in the noncoding segment is
zero; if \( O = 0 \), constraint takes the value of 1. The method is only applicable to closely related species for which multiple hits can be safely ignored.

In a pairwise comparison, it is not possible to determine the direction of a particular substitution (i.e., whether a C→T difference is caused by a C→T or a T→C substitution). However, it is possible to infer the proportion of changes that are in a particular direction if we assume or know the equilibrium base composition. Let us group Cs and Ts together, and As and Ts together. Let \( f_x \) be the equilibrium G+C content of the sequence; this is the G+C content that the sequence will eventually reach, and let \( z \) be a mutation rate parameter such that the rate at which A or T sites change to G or C is \( f_x \), and the rate at which G or C sites change to A or T is \( z f_x \). We can then use the present and equilibrium G+C content to infer the proportion of observed AT→GC differences that go in a particular direction (this category of differences involves the following pairwise differences: \( A \rightarrow G, A \rightarrow C, \) \( T \rightarrow G, \) and \( T \rightarrow C \)). However, with only two species, we cannot infer whether an observed G+C difference is caused by a G-to-C mutation or C-to-G mutation (this would require a parsimony approach). Similarly, we cannot assign polarity to any observed A→T differences. We can therefore only calculate four different kinds of nucleotide substitutions at fourfold and twofold degenerate synonymous sites of a gene, and to use these rates to calculate expected numbers of substitutions in an adjacent noncoding DNA segment. The numbers of substitutions at synonymous and noncoding sites are estimated using parsimony.

The basis of the approach is to calculate rates for all possible kinds of nucleotide substitutions at fourfold and twofold degenerate synonymous sites of a gene, and to use these rates to calculate expected numbers of substitutions in an adjacent noncoding DNA segment. The numbers of substitutions at synonymous and noncoding sites are estimated using parsimony. The rates for the 12 possible kinds of synonymous substitution in one of the branches, \( N_e(K_{A\rightarrow T}) \), and so on, are computed by dividing weighted averages of the frequencies of differences at fourfold, and, where applicable, twofold degenerate sites. The arrangement of neutral evolution at synonymous sites and equal mutation rates in the coding and noncoding DNA, the expected number of substitutions in the noncoding segment associated with gene \( i \) is, therefore,

\[
E_i = N_e(K_{A\rightarrow T} + K_{A\rightarrow C} + K_{G\rightarrow T} + N_f(K_{T\rightarrow A} + \ldots) + N_e(K_{C\rightarrow A} + \ldots))
\]

where \( N_f \) and \( N_e \) are the numbers of \( A, T, C, \) and \( G \) nucleotides, respectively, in the noncoding sequence. In cases in which all three base pairs differ, averages of \( K_{A\rightarrow T} + K_{A\rightarrow C} \) and so on, are calculated, weighted by the probabilities of alternative ancestral states, on the assumption that the relative lengths of the branch from the \emph{melanogaster}/\emph{simulans} common ancestor (a) to \emph{simulans} and \emph{melanogaster} are 0.2, and the relative length of the branch from a to \emph{yakuba} is 0.6. This model gives probabilities for the ancestral state being the \emph{melanogaster} or \emph{simulans} base of 0.462, and for the \emph{yakuba} base of 0.0769. \( O \) is the observed number of substitutions in the aligned noncoding sequence associated with gene \( i \). The average constraint is calculated by equation 8.

**ACKNOWLEDGMENTS**

We thank Brian Charlesworth, Toby Johnson, Penny Haddrill, and two reviewers for helpful comments and suggestions.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

**REFERENCES**


Akashi, H. 1995. Inferring weak selection from patterns of polymorphism and divergence at silent sites in \emph{Drosophila} DNA. \emph{Genetics} \textbf{139}:1067–1076.

———. 1996. Molecular evolution between \emph{Drosophila melanogaster} and \emph{D. simulans}: Reduced codon bias, faster rates of amino acid substitution, and larger proteins in \emph{D. simulans}. \emph{Genetics} \textbf{144}:1297–1307.

———. 1999. Inferring the fitness effects of DNA mutations from polymorphism and divergence data: Statistical power to detect directional selection under stationarity and free recombination. \emph{Genetics} \textbf{151}:221–238.


Aquadro, C.F., Lado, K.M., and Noon, W.A. 1988. The \emph{Xdh} region of \emph{Drosophila melanogaster} and \emph{Drosophila simulans}. I. Contrasting levels of naturally-occurring DNA restriction map variation and divergence. \emph{Genetics} \textbf{119}:875–888.

Begun, D.J. and Whitely, P. 2002. Molecular population genetics of \emph{Xdh} and the evolution of base composition in \emph{Drosophila}. \emph{Genetics} \textbf{162}:1725–1735.


WEB SITE REFERENCES

http://homepages.ed.ac.uk/eang33/; executables, source code, and user instructions for MCALIGN.

http://homepages.ed.ac.uk/eang33/; aligned sequence data files.

Received March 11, 2003; accepted in revised form December 1, 2003.