Uncovering the Footprint of Positive Selection on the X Chromosome of Drosophila melanogaster

Dorcas J. Orengo* and Montserrat Aguadé
Departament de Genètica, Facultat de Biologia, Universitat de Barcelona, Barcelona, Spain

*Corresponding author: E-mail: dorcasorengo@ub.edu.
Associate editor: Michael Nachman

Abstract

A usual approach to detect the spatial footprint left by recent adaptive events has been to follow up putative candidates emerging from multilocus scans of variation by sequencing additional fragments. We have used a similar experimental and analytical approach to study variation at 15 independently evolving and randomly chosen regions of the X chromosome of Drosophila melanogaster. These incompletely sequenced regions, each extending over ~40 kb, were subjected to two tests of positive selection that take into account the spatial distribution of nucleotide variation. Our analysis of variation at these genomic regions in a European population of D. melanogaster has allowed us to uncover a candidate region for positive selection and to empirically evaluate the comparative performance of the two tests of selection under a bottleneck scenario. Moreover, the boundaries here estimated for both the rate of adaptive substitution (\(d\)) and the average selection coefficient (\(s\)) would support previous estimates obtained by maximum likelihood that suggest rather strong but uncommon positive selection.

Key words: Drosophila melanogaster, nucleotide polymorphism, positive selection.

Introduction

Variation in DNA sequences is shaped over evolutionary time by diverse forces as genetic drift, demography, and selection. The comparison of sequences from extant individuals can thus be informative on the past action of such agents. The recent action of positive selection has a locus-specific and short-lived effect on linked nucleotide polymorphism (hitchhiking effect; Maynard Smith and Haigh 1974; Kaplan et al. 1989; Stephan et al. 1992; Kim and Stephan 2002). The increasing interest in identifying the molecular changes underlying adaptations in a particular lineage has fostered the search for the footprint of recent adaptive changes in DNA sequences. Also, it has led to the development of statistical tests that take into account the spatial distribution of nucleotide variation expected around the target of selection (Kim and Stephan 2002; Nielsen et al. 2005).

In Drosophila melanogaster, a usual experimental approach to detect the spatial footprint left by recent adaptive events has been to follow up putative candidates emerging from multilocus scans of variation by sequencing additional fragments (Beisswanger et al. 2006; Glinka et al. 2006; Orengo and Aguadé 2007). These incompletely sequenced regions have been then subjected to the Kim and Stephan test of positive selection (Kim and Stephan 2002) and subsequently to a goodness-of-fit (GOF) test (Jensen et al. 2005). Here, we have followed a similar experimental and analytical approach to study variation not at candidate regions but at 15 independently evolving and randomly chosen regions of the X chromosome. Indeed, similarly to the candidate regions, each of the 15 regions has been incompletely sequenced, and the sequences obtained have been subjected to the same two statistical tests. The 15 regions have also been subjected to a second test of positive selection that would be a priori more robust to demographic change and thus less prone to yield false positives (Nielsen et al. 2005). Our analysis of variation at the 15 genomic regions has allowed us 1) to empirically evaluate the performance of the Nielsen et al. method as compared with the Kim and Stephan method under a bottleneck scenario; 2) to uncover a candidate region for positive selection; and 3) to obtain boundaries for the rate and strength of adaptive evolution in this population.

Materials and Methods

Drosophila Strains and Sequences

The same 13 X-isochromosomal D. melanogaster lines from the Sant Sadurní d’Anoia (Catalonia, Spain) population analyzed in Orengo and Aguadé (2004) were used in the present study. Moreover, Drosophila simulans and Drosophila sechellia sequences for all new fragments studied were obtained directly from the genome sequence (Drosophila 12 Genomes Consortium 2007; http://flybase.org) or as a consensus sequence from five D. simulans genome sequences (Begun et al. 2007; http://flybase.org).

Region and Fragment Selection

In order to study the pattern of variation in randomly chosen and independently evolving regions of the X chromosome, 15 previously sequenced fragments (Orengo and Aguadé 2004) were used for their extended analysis. These fragments (hereafter named focal fragments) can be considered to have evolved independently because they are all located along the portion of the X chromosome with normal-to-high recombination rate (i.e., 2 cM/Mb) and they are all separated by more than 350 kb. Moreover, their
choice was not based on previous knowledge of their level and pattern of variation, and they can indeed be considered a random sample of the 109 previously sequenced fragments concerning functional class (intergenic vs. intronic), scaled nucleotide diversity (π/K), and Tajima’s D statistic (Orengo and Aguadé 2004). Release 4.3 of the D. melanogaster genome sequence (http://flybase.org) was used to select four additional fragments (and to design the corresponding amplification and sequencing primers) for each region with the same criteria as in Orengo and Aguadé (2004). Each region consists therefore in five approximately 1-kg–long fragments spaced approximately 10 kb apart (fig. 1). Each region is named from its focal fragment number in Orengo and Aguadé (2004). Fragments in each region are named with this number followed by either “_0” (focal fragment) or a letter (a to d from proximal to distal to the telomere). When results led to sequence additional fragments, their distance in kb to the focal fragment was used to name them.

DNA Sequencing and Sequence Analysis

Genomic DNA was obtained using a quick DNA extraction procedure (protocol 48 in Ashburner 1998). After purification of polymerase chain reaction products as described in Dean et al. (2003), both strands were sequenced with the ABI PRISM version 3.1 kit (Applied Biosystems, Foster City, CA) following the manufacturer’s instructions. Sequencing reaction products were ethanol precipitated and later separated on an ABI PRISM 3730 sequencer (PerkinElmer, Norwalk, CT). Sequences were assembled and multiplied aligned using the DNASTAR (Madison, WI) software package. Chromatograms were in all cases visually inspected and all polymorphic sites checked both in each line and across lines. The newly obtained sequences have been deposited in the EMBL/GenBank Data Library under accession numbers AM933850-AM934624 and FM211276-FM211352.

The multiple alignments were edited with the MacClade version 3.06 program (Maddison WP and Maddison DR 1992). The DnaSP version 4.10.9 program (Rozas et al. 2003) was used to estimate nucleotide diversity (π; Nei 1987), Tajima’s D statistic (Tajima 1989), and nucleotide divergence (K; Jukes and Cantor 1969). A spreadsheet program was used to calculate the D/D_{min} statistic values (Schaeffer 2002; Schmid et al. 2005).

Analysis of Selection and Sweep Localization

The composite likelihood ratio (CLR) method of Kim and Stephan (2002) for detecting positive selection along a recombining chromosome was applied to each region composed of five fragments that span approximately 40 kb. When information was missing for a particular strain and fragment, analyses were performed considering that the polymorphic sites at these missing sequences had the ancestral state. This is a conservative assumption, as shown by Orengo and Aguadé (2007).

The clsw program was used to obtain the likelihood ratio (LR) of the observed data set for test B (Kim and Stephan 2002), under the assumption of recombination. Availability of the D. simulans and D. sechellia sequences allowed using the LR1 option of the program, which assumes that the derived state of a segregating site is known. In the few cases where the derived state could not be unambiguously established, the nucleotide with the highest frequency was assumed to be the ancestral one.

Statistical significance was obtained from computer simulations. The ms program (Hudson 2002) was used to generate neutral genealogies of the region under study. Simulations were performed under recombination and conditioned on theta. The population recombination parameter used (C = 2Nc, given the lack of recombination in Drosophila males) was obtained for each region using the Hey and Kliman (2002) estimates of the recombination rate (c). Simulations were performed both under stationarity (standard neutral model, hereafter SNM) and under the demographic scenario proposed by Li and Stephan (2006), which considers a recent bottleneck in the derived European population and an older expansion in the ancestral African population. The simulated data sets were run through clsw in order to obtain the corresponding P values of the observed data set.

We also used the parametric approach of Nielsen et al. (2005) to detect sweeps from single nucleotide polymorphism (SNP) data. Since the derived state of segregating sites could be inferred (see above), the unfolded option from the SweepFinder program was used. The background frequency spectrum used in this method was obtained from the 107 polymorphic loci (1030 polymorphic sites) in the X-chromosome scan of Orengo and Aguadé (2004). For each of the 15 five-fragment regions, the LR was calculated for a different grid size to obtain estimates every 100 sites. Although this test seems to be robust to demographic assumptions (Nielsen et al. 2005), P values were estimated by comparing the LR value from the real data set with the distribution obtained from 1,000 simulated data sets under the SNM and the bottleneck scenario.

A general GOF test (Jensen et al. 2005) was performed for those regions where significant P values were obtained by both CLR methods under the bottleneck scenario. The GOF values obtained for the empirical data set were
compared with those obtained from 1,000 data sets simulated under a selective sweep model. These simulations were performed with the ssw program (Kim and Stephan 2002) using the number of segregating sites (S) and the maximum-likelihood estimates of the location of the advantageous mutation (X) and the strength of selection ($\alpha = 1.5Ns$) obtained from the empirical data set by the Kim and Stephan method. These simulations also allowed us to calculate the confidence intervals (CIs) of $X$ and $s$ by parametric bootstrapping (Jensen et al. 2008).

Exploring the Selection Parameter Space

Computer simulations were performed to estimate the probability of detecting $x$ recent selective events under our experimental design (i.e., considering the 15 five-fragment regions studied) and under different selection scenarios as characterized by the rate of adaptive substitution ($\delta$) and by their average selection coefficient ($s$). The parameter space was coarsely explored (through 36 $\delta$ and 10 $s$ estimates) in an effort to cover the range of previous estimates. Ten thousand replicates were obtained for each parameter combination. The selective events that under a particular rate $\delta$ would have affected the X chromosome during the last 15,000 years were distributed randomly across this chromosome ($\sim 22.22$ Mb long; Flybase release 4.3) by sampling from a uniform distribution. The stretch affected by each selective sweep ($2d$, where $d$ is the distance to the target of selection) was estimated from the selection coefficient $s$ using the ($c/s$) log $N$ approximation of Maynard Smith and Haigh (1974), where the recombination rate $c$ was considered to be equal to 2 cm/Mb and the effective population size $N$ equal to $10^6$. Two common requirements were established for any sweep to be considered detectable: 1) the target of selection had to be within the b–c section of any of the 15 regions actually analyzed (fig. 1) and 2) the target distance to the first (or last) nucleotide of fragment a (or d) had to be at least 30% of $d$. Moreover, for the smallest sweeps, at least one of the sequenced fragments had to be within the central stretch corresponding to 80% of the affected region ($2d$).

For each parameter combination, three probabilities were obtained: $P(x < n)$, $P(x = n)$, and $P(x > n)$, where $n$ is the number of regions that in the empirical data set were affected by a selective sweep. Two likelihood ratios were obtained: $LR_A = P(x < n)/P(x \geq n)$ and $LR_B = P(x \leq n)/P(x > n)$, which were used to initially establish the upper and lower bounds of the adaptive substitution rate $\delta$. Moreover, the likelihood surface of $P(x = n)$ was used to narrow down this range and to establish the range of average selection coefficients ($s$) consistent with our observation.

Results and Discussion

Nucleotide Polymorphism and Tests of Positive Selection

Figure 1 shows the location across the X chromosome of the 15 five-fragment regions studied in D. melanogaster. The average distance between consecutive focal fragments was 1.15 Mb, with only two cases where the distance was below 500 kb. Each region covers approximately 40 kb (from 38.8 to 45.1 kb) of chromosome X, of which an average of 4.2 kb was sequenced on both strands. A total of 780 SNPs were found in the 62,452 nucleotide sites sequenced. Detailed information on nucleotide polymorphism and divergence for each fragment is provided in supplementary table S1, Supplementary Material online. Average estimates of polymorphism ($\pi = 0.0040$) and divergence ($K = 0.055$) in these fragments were similar to those previously detected at 109 short X-chromosome fragments ($\pi = 0.0038$, $K = 0.051$; Oreno and Aguadé 2004).

The 15 five-fragment regions were subjected individually to the Kim and Stephan test and also to the Nielsen et al. test. Both tests yielded rather similar results under the SNM with 9 and 10 regions, respectively, exhibiting significant results (table 1). This similarity was rather unexpected given that the Nielsen et al. test, unlike the Kim and Stephan test, seemed to be robust to demographic assumptions (Nielsen et al. 2005). The observed LR values were then tested against the null distribution built under the more realistic scenario of a bottleneck (Li and Stephan 2006) with two different purposes. Indeed, for the subset of regions that conformed to SNM predictions, this extended analysis aimed to confirm that the SNM results were robust to demographic uncertainty. In contrast, for those regions with significant LR values under the SNM, the extended analysis aimed to uncover putative false positives. In the first case, the LR values obtained for each region by both methods remained nonsignificant (results not shown), indicating that performing the analysis under the SNM was sufficient to identify true negatives. In the second case, the previously detected significance of observed LR values (for 9 and 10 regions, respectively) was lost for an important fraction of regions. Indeed, only the LR value at regions 35 and 43 for the Kim and Stephan test was significant and at regions 35, 43, and 49 for the Nielsen et al. test (table 1). Both tests yielded a similar and large number of false positives when the demographic history of the studied population was ignored, that is, when simulations under the SNM were used to establish the significance of the LR value obtained from the data. This observation is consistent with previous work on the Kim and Stephan test (Kim and Stephan 2002; Jensen et al. 2005; Thornton and Jensen 2007). It is, however, in contrast with the robustness of the Nielsen et al. test to demographic scenarios other than those considered here (Nielsen et al. 2005; but see also Williamson et al. 2007; Pavlidis et al. 2008). Further exploration of the robustness of the Nielsen et al. test both under bottleneck and other demographic scenarios seems necessary, and any significant result of this test under the SNM should be therefore taken with caution. When both tests were considered (eight regions with significant LR values under the SNM), only for two of the regions (35 and 43) did the test results suggest the recent action of positive selection. In order to discard any possible false positive, the GOF test was applied to both regions. Only the
Table 1. Composite Likelihood Ratio Tests for Positive Selection.

<table>
<thead>
<tr>
<th>Locus</th>
<th>L</th>
<th>S</th>
<th>LR(KS)</th>
<th>SNM</th>
<th>Bott</th>
<th>LR(SF)</th>
<th>SNM</th>
<th>Bott</th>
</tr>
</thead>
<tbody>
<tr>
<td>12</td>
<td>38,808</td>
<td>33</td>
<td>4.69</td>
<td>NS</td>
<td></td>
<td>1.52</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>18</td>
<td>40,105</td>
<td>64</td>
<td>2.95</td>
<td>NS</td>
<td></td>
<td>2.08</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>30</td>
<td>39,723</td>
<td>61</td>
<td>10.94</td>
<td>&lt;0.001*</td>
<td>NS</td>
<td>5.07</td>
<td>0.014*</td>
<td>NS</td>
</tr>
<tr>
<td>35</td>
<td>43,828</td>
<td>63</td>
<td>23.06</td>
<td>&lt;0.001*</td>
<td>0.001*</td>
<td>29.49</td>
<td>&lt;0.001*</td>
<td>&lt;0.001*</td>
</tr>
<tr>
<td>36</td>
<td>43,203</td>
<td>48</td>
<td>4.97</td>
<td>NS</td>
<td></td>
<td>1.59</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>43</td>
<td>43,343</td>
<td>23</td>
<td>9.53</td>
<td>&lt;0.001*</td>
<td>0.011*</td>
<td>10.98</td>
<td>&lt;0.001*</td>
<td>0.010*</td>
</tr>
<tr>
<td>46</td>
<td>41,610</td>
<td>51</td>
<td>13.59</td>
<td>&lt;0.001*</td>
<td>0.033</td>
<td>13.88</td>
<td>&lt;0.001*</td>
<td>0.033</td>
</tr>
<tr>
<td>49</td>
<td>43,858</td>
<td>20</td>
<td>5.64</td>
<td>0.009*</td>
<td>NS</td>
<td>9.13</td>
<td>&lt;0.001*</td>
<td>0.010*</td>
</tr>
<tr>
<td>51</td>
<td>45,091</td>
<td>58</td>
<td>6.20</td>
<td>0.018*</td>
<td>NS</td>
<td>8.30</td>
<td>&lt;0.001*</td>
<td>NS</td>
</tr>
<tr>
<td>59</td>
<td>45,007</td>
<td>59</td>
<td>11.06</td>
<td>0.001*</td>
<td>NS</td>
<td>2.03</td>
<td>NS</td>
<td></td>
</tr>
<tr>
<td>66</td>
<td>41,838</td>
<td>43</td>
<td>13.51</td>
<td>&lt;0.001*</td>
<td>0.039</td>
<td>11.65</td>
<td>0.001*</td>
<td>NS</td>
</tr>
<tr>
<td>71</td>
<td>41,501</td>
<td>60</td>
<td>2.22</td>
<td>NS</td>
<td></td>
<td>3.98</td>
<td>0.033*</td>
<td>NS</td>
</tr>
<tr>
<td>73</td>
<td>40,089</td>
<td>56</td>
<td>8.10</td>
<td>0.001*</td>
<td>NS</td>
<td>9.04</td>
<td>0.001*</td>
<td>NS</td>
</tr>
<tr>
<td>84</td>
<td>43,604</td>
<td>87</td>
<td>5.14</td>
<td>NS</td>
<td></td>
<td>5.19</td>
<td>0.023*</td>
<td>NS</td>
</tr>
<tr>
<td>97</td>
<td>44,745</td>
<td>51</td>
<td>3.04</td>
<td>NS</td>
<td></td>
<td>4.02</td>
<td>0.044</td>
<td></td>
</tr>
</tbody>
</table>

Note.—Locus, number of the focal fragment according to Orengo and Aguadé (2004); L, length spanning the S sequenced fragments; S, number of polymorphic sites in the S fragments; LR(KS), composite likelihood ratio estimated by the method of Kim and Stephan (2002); SNM, probability of the observed LR under the standard neutral model; Bott, probability of the observed LR under the bottleneck as in Li and Stephan (2006); LR(SF), composite likelihood ratio estimated by the method of Nielsen et al. (2005); NS, not significant.

*Significant regions at the 5% false discovery rate (Benjamini and Hochberg 1995); these regions are also significant when applying the stringent sequential Bonferroni correction (Rice 1989) at the 5% level for tests under SNM and at the 9% level for those under Bott.

Exploration of the Selection Parameter Space

Two neutrality tests that use information on the frequency spectrum of polymorphic variants and on their spatial distribution in a recombining chromosome were used in the present study to analyze variation at 15 independently evolving regions of the X chromosome spanning ~40 kb each. The analysis was performed in a sequential way, that is, only regions exhibiting a significant result at each step were further analyzed to separate false from true positives. From this analysis, only region 43 emerged as reflecting a recent adaptive change. This observation can be used to coarsely explore the parameter space in an effort to establish boundaries for both the rate of adaptive substitution ($\delta$) and the average selection coefficient ($s$). As described in Materials and Methods, a simplified and therefore approximate approach was used for that purpose. Figure 2 shows the likelihood surfaces corresponding to the two ratios $LR_A$ and $LR_R$ and to $P(x = 1)$. An $LR_A$ value equal to 1 was used to establish the lower limit of $\delta (1.6 \times 10^{-11})$ given that values of $LR_A$ higher than 1 correspond to situations where it would be more likely that our experimental design had not detected any region affected by a sweep. On the other hand, $LR_R = 1$ was used to establish the upper limit of $\delta (5.2 \times 10^{-11})$ given that values of $LR_R$ less than 1 correspond to situations where it would be more likely that our experimental design had uncovered more than one affected region. Finally, the likelihood surface for $P(x = 1)$ was used to narrow down these limits and to establish boundaries for the most probable values of $s$. This procedure suggested $\delta$ values between $2 \times 10^{-11}$ and $3.8 \times 10^{-11}$ and $s$ values between 0.002 and 0.0055. The boundaries for the rate of adaptive evolution obtained using our approximate approach fall within the CI ($2.6 \times 10^{-11}$ to $6.46 \times 10^{-11}$) estimated by Li and Stephan (2006) for another European population using a rigorous maximum-likelihood approach. Moreover, the limits here estimated for the strength of selection are compatible with those estimated by these authors. Our results would therefore support these and other similar estimates for D. melanogaster (Li and Stephan 2006; Jensen et al. 2008) suggesting rather strong but uncommon positive selection as opposed to estimates suggesting weak but frequent positive selection (Andolfatto 2007). It should be however considered that the experimental and analytical

Table 2. Composite Likelihood Ratio and Goodness-of-Fit Tests for Positive Selection.

<table>
<thead>
<tr>
<th>Locus</th>
<th>L</th>
<th>S</th>
<th>LR(KS)</th>
<th>X</th>
<th>LR(SF)</th>
<th>GOF</th>
</tr>
</thead>
<tbody>
<tr>
<td>35</td>
<td>43,828</td>
<td>63</td>
<td>11,274.90</td>
<td>14,519</td>
<td>23.06</td>
<td>138.55 (0.104)</td>
</tr>
<tr>
<td>46</td>
<td>41,610</td>
<td>51</td>
<td>13.59</td>
<td>0.001*</td>
<td>0.033</td>
<td>13.88</td>
</tr>
<tr>
<td>49</td>
<td>43,858</td>
<td>20</td>
<td>5.64</td>
<td>0.009*</td>
<td>NS</td>
<td>9.13</td>
</tr>
<tr>
<td>51</td>
<td>45,091</td>
<td>58</td>
<td>6.20</td>
<td>0.018*</td>
<td>NS</td>
<td>8.30</td>
</tr>
<tr>
<td>59</td>
<td>45,007</td>
<td>59</td>
<td>11.06</td>
<td>0.001*</td>
<td>NS</td>
<td>2.03</td>
</tr>
<tr>
<td>66</td>
<td>41,838</td>
<td>43</td>
<td>13.51</td>
<td>&lt;0.001*</td>
<td>0.039</td>
<td>11.65</td>
</tr>
<tr>
<td>71</td>
<td>41,501</td>
<td>60</td>
<td>2.22</td>
<td>NS</td>
<td></td>
<td>3.98</td>
</tr>
<tr>
<td>73</td>
<td>40,089</td>
<td>56</td>
<td>8.10</td>
<td>0.001*</td>
<td>NS</td>
<td>9.04</td>
</tr>
<tr>
<td>84</td>
<td>43,604</td>
<td>87</td>
<td>5.14</td>
<td>NS</td>
<td></td>
<td>5.19</td>
</tr>
<tr>
<td>97</td>
<td>44,745</td>
<td>51</td>
<td>3.04</td>
<td>NS</td>
<td></td>
<td>4.02</td>
</tr>
</tbody>
</table>

Note.—Locus, number of the focal fragment according to Orengo and Aguadé (2004); L, length spanning the sequenced fragments; S, number of polymorphic sites; $x$, population selection parameter; X, location of advantageous mutation; LR(KS), composite likelihood ratio estimated by the method of Kim and Stephan (2002); GOF, goodness-of-fit test by Jensen et al. (2005); P values associated to the GOF test are given in parentheses.

a Six-fragment region.

b Ten-fragment region.
approaches used in the present study cannot uncover weak sweeps, which might lead to overestimate the strength of selection and underestimate the rate of adaptive evolution.

Further Characterization of the Candidate Region under Selection

According to the Kim and Stephan method, the putative target of selection for region 43 would be located near the distal end of the region surveyed, leading to a poor delimitation of the effect of the sweep on that distal part. In an effort to better delimit the sweep effect at this region, variation was surveyed at an additional fragment (+43; fig. 3), which would also allow establishing how robust the inference of positive selection based on the five-fragment region was. The scaled nucleotide diversity (π/K) and the $D/D_{\text{min}}$ value at fragment +43 were similar to those at fragment 43b and much higher than at intervening fragments 43_0, 43c, and 43d (supplementary table S2, Supplementary Material online). For the six-fragment region, like for the five-fragment region, the Kim and Stephan test of positive selection yielded significant LR values under the SNM and under the bottleneck scenario (results not shown). Likewise, the GOF test would also support the sweep hypothesis (table 2). The estimated location of the putative target of selection from the five-fragment and six-fragment data differed only by 163 nucleotides, although the associated CIs were in both cases rather large as expected from the low coverage of the studied region (Jensen et al. 2008). The analysis was therefore further extended by additional sequencing between fragments 43_0 and +43. Indeed, four additional fragments (+5, +24, +27, and +36) were sequenced and fragments c and d were widely extended. The effort concentrated specially around the putative target of selection and the nearby exon of gene CG34339, where an almost continuous 13-kb stretch was sequenced. Only small parts of this region could not be sequenced due to the presence of long homonucleotide runs. A total of 71 nucleotide polymorphic sites (43 with singletons) were detected in the 10 fragments extending over 66.7 kb (supplementary fig. S1, Supplementary Material online). Similarly to the maximum-likelihood analysis of the 5- and 6-fragment data, application of the Kim and Stephan method to the 10-fragment data yielded significant LR values both under the SNM and under the bottleneck scenario (results not shown). Moreover, the GOF test would also support the sweep hypothesis (table 2).

The location of the putative target of selection (X) obtained from the 10-fragment data was more similar to that obtained from the 6-fragment data (corresponding to the same 67-kb region) than from the 5-fragment data (table 2). Also, the beneficial mutation strength (s) estimated from the 10-fragment data (~0.003) was closer to the 6-fragment estimate (~0.006) than to the 5-fragment estimate (~0.011). The 95% CIs for the X and s values estimated from the 10-fragment data were much narrower than those estimated from the 6-fragment data: (22,734–52,721) versus (35,377–44,003) for X, and (0.00026–0.01358) versus (0.00077–0.00513) for s. Indeed, the accuracy of the estimated values increases when analyzing long fragments of continuous sequence (Jensen et al. 2008) as in our extended analysis of 10 fragments.

![Joint likelihood surface for the rate of adaptive substitution (δ) and the average selection coefficient (s): (A) ratio of the probabilities of detecting zero and at least one selective event (LRₐ); (B) ratio of the probabilities of detecting at most one, and more than one selective event (LR₉); (C) probability of detecting one selective event as observed. The solid bars in (A) and (B) indicate the lower and upper bounds of δ shown in (A) and (B), respectively, whereas the horizontal solid lines correspond to the narrower range established from the likelihood surface for $P(x = 1)$, and the vertical solid lines indicate the lower and upper bounds for s.](image-url)
includes an ~13-kb nearly continuous sequenced stretch. Although the predicted target of selection would be in the third intron at 4.5 kb from exon 3 of gene CG34339, it could lie—according to the estimated CI—anywhere in the 5′ part of this long intron (fig. 3). Selection might therefore have acted at any putative regulatory motif within this intronic region (Haddrill et al. 2005), as previously suggested for the first intron of the ph-p gene of D. melanogaster (Beisswanger and Stephan 2008). Unfortunately, neither the function of gene CG34339 nor the associated biological processes are known, which precludes any functional inference concerning the recent adaptive change suggested by our analysis.

In summary, our analysis of nucleotide variation at 15 independently evolving and randomly chosen regions of the X chromosome revealed that the Nielsen et al. maximum-likelihood method, similarly to the Kim and Stephan method, is not robust to the particular bottleneck scenario considered here. Moreover, the boundaries for the rate of adaptive substitution and the average selection coefficient established from our observation of a single candidate region would support previous maximum-likelihood estimates that suggest rather strong and uncommon positive selection. Finally, the status of the candidate region was confirmed through its extended analysis, which allowed narrowing down the putative target of selection.

**Fig. 3.** Analysis of variation at region 43. Top panel, gray vertical bars indicate the location of the 10 sequenced fragments in the CG34339 gene and striped bars show the five originally sequenced fragments. Middle panel, distribution of the scaled nucleotide diversity ($\pi/K$; solid line) and the ratio of Tajima’s $D$ to its theoretical minimum value ($D/D_{\min}$; dashed line). Bottom panel, distribution of the Kim and Stephan LR statistic along the 10-fragment region; location of the predicted target of selection corresponding to the 6-fragment region (diamond) and the 10-fragment region (triangle) are shown with their corresponding 95% CIs.
to an ~9-kb stretch in an intron of gene CG34339. This region would be an important addition to the relatively few detected regions in *D. melanogaster* with evidence suggesting the recent fixation of an advantageous mutation (Harr et al. 2002; Bauer DuMont and Aquadro 2005; Glinka et al. 2006; Pool et al. 2006; Jensen et al. 2007; Orengo and Aguadé 2007; Beisswanger and Stephan 2008; Svetec et al. 2009).

**Acknowledgments**

We thank David Salguero for excellent technical assistance and Serveis Científico-Tècnics from Universitat de Barcelona for automated sequencing facilities. We also thank David Alvarez and Filipe G. Vieira for help with computer program implementation and Yuseob Kim for his program to graphically represent the output of the Kim and Stephan analysis. Special thanks are given to Aleksandra Gondek for her contribution, as a predoctoral Marie Curie fellow, at the initial stage of this project and to Carmen Segarra and anonymous reviewers for comments on the manuscript. This work was supported by grants QLRT-2001-00004 from the European Community, BFU2007-63228 from Comisión Interdepartamental de Ciencia y Tecnología, Spain; and 2005SGR-00166 from Comissió Interdepartamental de Recerca i Innovació Tecnològica, Generalitat de Catalunya, Spain.

**References**


