Purifying Selection Maintains Highly Conserved Noncoding Sequences in Drosophila

Sònia Casillas,*† Antonio Barbadilla,† and Casey M. Bergman*

*Faculty of Life Sciences, University of Manchester, Michael Smith Building, Manchester M13 9PT, UK; and †Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, 08193 Bellaterra (Barcelona), Spain

The majority of metazoan genomes consist of nonprotein-coding regions, although the functional significance of most noncoding DNA sequences remains unknown. Highly conserved noncoding sequences (CNSs) have proven to be reliable indicators of functionally constrained sequences such as cis-regulatory elements and noncoding RNA genes. However, CNSs may arise from nonselective evolutionary processes such as genomic regions with extremely low mutation rates known as mutation “cold spots.” Here we combine comparative genomic data from recently completed insect genome projects with population genetic data in Drosophila melanogaster to test predictions of the mutational cold spot model of CNS evolution in the genus Drosophila. We find that point mutations in intronic and intergenic CNSs exhibit a significant reduction in levels of divergence relative to levels of polymorphism, as well as a significant excess of rare derived alleles, compared with either the nonconserved spacer regions between CNSs or with 4-fold silent sites in coding regions. Controlling for the effects of purifying selection, we find no evidence of positive selection acting on Drosophila CNSs, although we do find evidence for the action of recurrent positive selection in the spacer regions between CNSs. We estimate that ~85% of sites in Drosophila CNSs are under constraint with selection coefficients (N_e s) on the order of 10–100, and thus, the estimated strength and number of sites under purifying selection is greater for Drosophila CNSs relative to those in the human genome. These patterns of nonneutral molecular evolution are incompatible with the mutational cold spot hypothesis to explain the existence of CNSs in Drosophila and, coupled with similar findings in mammals, argue against the general likelihood that CNSs are generated by mutational cold spots in any metazoan genome.

Introduction

A largely unexplained structural feature common to metazoan genomes is the presence of vast amount of nonprotein-coding DNA (Britten and Davidson 1969; Taft and Mattick 2003). For example, over 75% of the euchromatic portion of the Drosophila melanogaster genome sequence is found in noncoding intronic and intergenic regions (Misra et al. 2002). Unlike large mammalian genomes, the overwhelming majority of noncoding DNA in the D. melanogaster genome sequence is unique, with less than 6% of the genome confidently identified as repetitive transposable element sequences (Quesneville et al. 2005; Bergman et al. 2006). Of these 80+ Mb of unique noncoding sequences in D. melanogaster, a minimum of ~20–30% has been shown to be highly conserved among other insect species (Bergman and Kreitman 2001; Bergman et al. 2002; Siepel et al. 2005). These highly conserved noncoding sequences (CNSs) are typically interpreted as representing the signature of functionally constrained elements maintained by purifying selection and have been successfully used to guide the prediction of cis-regulatory sequences (Bergman et al. 2002; Costas et al. 2004) and functional noncoding RNAs (Enright et al. 2003; Lai et al. 2003).

Despite the widespread assumption that CNSs are maintained by the action of purifying selection, the first results on the length distribution of CNSs reported in Bergman and Kreitman (2001) were shown to be compatible with a nonselective mode of evolution that invokes only mutation and genetic drift (Clark 2001), albeit assuming extremely low mutation rates that vary over spatial scales on the order of tens of base pairs. This lead Clark (2001) to propose the “mutational cold spot” hypothesis to explain the existence of CNSs as an alternative to the functional constraint hypothesis (see also Shabalina and Kondrashov 1999). Subsequently, arguments against the mutational cold spot hypothesis were levied on the grounds of the nonrandom spatial distribution of CNSs in Drosophila (Bergman et al. 2002) and Caenorhabditis (Webb et al. 2002), a pattern which would further require a nonrandom organization of mutational cold spots to explain the pattern of CNS evolution. Although no molecular mechanism has been shown to produce such localized mutation cold spots, local variation in mutation rates remains a formal possibility that must be investigated more thoroughly to demonstrate the general property that CNSs are indeed selectively constrained.

Recently, the mutational cold spot hypothesis has been tested directly using population genetic data from single-nucleotide polymorphism (SNP) projects in the human genome (Keightley, Kryukov et al. 2005; Drake et al. 2006). Both the mutational cold spot hypothesis and the functional constraint hypothesis predict that levels of within-species variation in CNSs should be reduced relative to flanking non-CNS sequences. This is because either low mutation rates or the elimination of deleterious alleles by purifying selection can reduce the number of segregating polymorphisms observed in CNSs. In contrast to levels of variation, the 2 hypotheses make distinct predictions about the distribution of allele frequencies in CNS and non-CNS regions. Under the mutational cold spot hypothesis, CNS and non-CNS regions should not differ from one another in their allele frequency spectra because their dynamics are both governed only by mutation and genetic drift. In contrast, the functional constraint hypothesis predicts that CNSs should harbor more rare derived alleles relative to non-CNS regions if CNSs are maintained by purifying...
selection that confines weakly deleterious alleles within low population frequencies. Tests of these alternative models based on differences in derived allele frequency (DAF) distributions have concluded that CNSs in the human genome exhibit an excess of rare derived alleles, consistent with functional constraint acting to maintain CNSs in mammals (Keightley, Kryukov et al. 2005; Drake et al. 2006). Similar results have also been reported for SNPs in predicted microRNA binding sites in the 3’ untranslated regions (UTRs) of human genes (Chen and Rajewsky 2006). Moreover, analysis of the distribution of fitness effects on mutations in mammalian CNSs has revealed the action of weak purifying selection (Kryukov et al. 2005), a mode of evolution which would allow deleterious SNPs to be observed in nature but prevent most of them from reaching high population frequency or going to fixation.

Here we investigated the evolutionary forces governing the evolution of CNSs in the Drosophila genome, using recently available comparative genomic data from insect genome projects (Holt et al. 2002; Richards et al. 2005; The Honeybee Genome Sequencing Consortium 2006; http://rana.lbl.gov/drosophila) combined with population genetic data from published nucleotide polymorphism studies in noncoding regions of the X chromosome in D. melanogaster (Glinka et al. 2003; Orengo and Aguade 2004; Ometto, Glinka et al. 2005). We extend previous within-species analysis of evolutionary dynamics in CNSs by testing a second prediction of the mutational cold spot hypothesis, based on a modified version of the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991), similar to that used to test differences between transcription factor binding sites and spacer regions in cis-regulatory sequences (Jenkins et al. 1995; Ludwig and Kreitman 1995). Specifically, we test whether the ratio of polymorphism within D. melanogaster relative to divergence with its sister species Drosophila simulans is the same in CNS and non-CNS regions, as is expected under a model of strictly neutral evolution required by the mutational cold spot hypothesis. By using resequencing data rather than data from SNP studies as used in the studies in mammals cited above, we have access to more comprehensive, unbiased population genetic data from contiguous genomic sequences. This permits us to test the mutational cold spot hypothesis more rigorously using both the MK test as well as the DAF test and further allows us to investigate the impact of both single nucleotide as well as insertion/deletion (indel) mutations on the evolution of CNSs.

Materials and Methods
Compilation of Published Sequence Data

We have used population genetic data from genome scans of noncoding regions homogeneously distributed across the X chromosome of D. melanogaster (Glinka et al. 2003; Orengo and Aguade 2004; Ometto, Glinka et al. 2005). These data include ~12 alleles per locus from each of 3 distinct data sets: AFR is an African sample from Lake Kariba, Zimbabwe; EUR1 is a European sample from Leiden, The Netherlands; and EUR2 is another European sample from Sant Sadurni d’Anoia, Catalonia, Spain. We used a modified version of Pipeline Diversity Analysis (PDA) (Casillas and Barbadilla 2006) to automatically retrieve noncoding sequences from Genbank, classify and align them by population and locus (see below), and estimate diversity measures. Data for coding regions to calculate levels of silent site polymorphism and divergence was obtained from Andolfatto (2005).

We selected one allele from each data set and used BLAT (Kent 2002) to map each alignment to the D. melanogaster genome sequence, which also provided an additional allele to each locus that was used only for reference purposes but which was never included in the polymorphism analyses, thus conserving the unique origin of the sequences in each population data set. Based on preliminary results indicating unusual properties of indels in the D. simulans composite whole-genome shotgun assembly, we used the orthologous D. simulans allele provided by the authors in the original papers. For loci where this sequence was not available, we obtained the orthologous D. simulans regions using whole-genome alignments from the VISTA browser (Couronne et al. 2003) using the genomic coordinates of D. melanogaster. D. simulans alleles were used to polarize polymorphic SNPs and indel polymorphisms (IPs) and to define single-nucleotide fixed differences (SNFs) and indel fixed differences (IFs).

Multiple Sequence Alignment

We evaluated the performance of several programs to generate multiple alignments of alleles by correlating estimates of polymorphism and divergence obtained automatically here with estimates derived from manually curated alignments previously reported in the primary publications. Based on this analysis, we chose to use MUSCLE (Edgar 2004) for multiple sequence alignment, which yielded correlations for polymorphism with $r > 0.96$ for the 3 populations and correlations for divergence with $r > 0.82$ (Supplementary File 1, Supplementary Material online).

Masking Exons, Repeats, and Low-Quality Regions

Each alignment was visually inspected but left unmodified. Unreliable alignments were discarded (7 loci from AFR and EUR1 and 10 loci from EUR2), and regions at the ends of the alignments were trimmed because of differences in the extent of sequence available for each allele. Coding exons, UTR exons, interspersed, and low-complexity repetitive sequences were also masked using the annotations from the University of California Santa Cruz (UCSC) Genome Browser (Hinrichs et al. 2006).

Defining Evolutionary CNSs

CNSs were obtained from the phastConsElements9-way track for the D. melanogaster genome at the UCSC browser, which identifies highly conserved elements using a phylogenetic hidden Markov model (Siepel et al. 2005) applied to multiple alignment of 7 species of Drosophila, mosquito, and honeybee. The version of the genome assemblies used for the phastConsElements9way
track are *D. melanogaster* (dm2), *D. simulans* (droSim1), *Drosophila yakuba* (droYak1), *Drosophila ananassae* (droAna1), *Drosophila pseudoobscura* (dp2), *Drosophila virilis* (droVir1), *Drosophila mojavensis* (droMoj1), *Anopheles gambiae* (anoGam1) and *Apis mellifera* (api-Mel1). Similar results were obtained when CNSs were defined as conserved blocks of >20 bp and >90% identity in pairwise alignments of *D. melanogaster* with *D. pseudoobscura* or *D. virilis*, respectively, using the VISTA browser (Couronné et al. 2003).

### Identification of Point and Indel Mutations

SNPs and IPs in *D. melanogaster* were polarized using the orthologous region in *D. simulans* assuming standard parsimony criteria. Therefore, only those well-characterized SNPs and IPs that could be polarized were kept for the analyses. SNPs were discarded when: 1) more than 2 different alleles were segregating at a polymorphic site, 2) polarization was not possible because the corresponding site in the outgroup species was missing (gapped site) or unknown (N), or 3) the *D. simulans* allele did not match either *D. melanogaster* allele. Correspondingly, IPs were discarded when: 1) they were located at the boundaries of the alignments, 2) they could not be derived because 2 or more indels were overlapping (either at the polymorphism level and/or with the outgroup species), or 3) they were spanning different categories of sites (e.g., a single indel was laying part in a CNS and part in a non-CNS region).

We considered SNFs as those nucleotide sites that were identical in all the *D. melanogaster* sequences, but different in the outgroup species. Similarly, IPs were determined as those indels that were identical in all the *D. melanogaster* sequences, but different in the outgroup species (insertion in *D. melanogaster*/deletion in *D. simulans*, or vice versa). As above, SNFs were discarded when the corresponding site in the outgroup species was a gap or an N, and IPs were discarded when: 1) they were located at the boundaries of the alignments, 2) indels in the 2 species were overlapping, or 3) they were spanning different categories of sites.

### Multilocus MK Test for Noncoding DNA

For comparisons of polymorphism and divergence between non-CNSs and CNSs in order to detect selection, we applied a modification of the MK test (McDonald and Kreitman 1991) for noncoding DNA. In this test, non-CNS sites were used in place of synonymous coding sites, and CNS sites were tested for the action of natural selection in place of nonsynonymous coding sites. Tests were performed both for point mutations (SNPs and SNFs) and indels (IPs and IFs) in all 3 populations. In each case, sites were pooled across all loci, and the significances were tested according to a $\chi^2$ test with 1 degrees of freedom (df). For a single fragment, the assumption that CNS and non-CNS sites share the same genealogy with little or no recombination is as valid as similar modifications to the MK test used for cis-regulatory sequences (Jenkins et al. 1995; Ludwig and Kreitman 1995).

### DAF Test

DAF analyses were performed both for point mutations and indels. Frequency distributions were created for sets of SNPs and IPs based on the data set and whether they were within or outside of CNSs. Significance was assessed using the Kolmogorov–Smirnov tests (Sokal and Rohlf 1995) to test for differences across the entire allele frequency distribution. Similar results were obtained using a $\chi^2$ test comparing the DAF distribution for SNPs within and outside CNSs, using 10% as a frequency cutoff to separate rare from common SNPs as in Drake et al. (2006).

### Estimating the Effects of Natural Selection

We estimated the selection coefficients operating on CNSs using 2 independent methods (Piganeau and Eyre-Walker 2003; Kryukov et al. 2005) and the proportion of constrained sites and sites undergoing positive selection in CNSs and non-CNSs using the methods of Halligan and Keightley (2006) and Smith and Eyre-Walker (2002), respectively. For the method of Kryukov et al. (2005), possible distributions of selection coefficients ($s$) were modeled by a 5- and 10-column histogram containing bins representing a fraction of sites under a given selection coefficient and where all bins together represent all sites within CNSs. For the 5-bin histograms, columns corresponded to $s$ equal to $-10^{-7}$, $-10^{-6}$, $-10^{-5}$, $-10^{-4}$, and $-10^{-3}$, and for the 10-bin histograms, columns corresponded to $s$ equal to $-10^{-7.5}$, $-10^{-7}$, $-10^{-6.5}$, $-10^{-6}$, $-10^{-5.5}$, $-10^{-5}$, $-10^{-4.5}$, $-10^{-4}$, $-10^{-3.5}$, and $-10^{-3}$. We applied the same theoretical measures as Kryukov et al. (2005), with the exception that we used 10% as a frequency cutoff to separate rare from common SNPs for the theoretical value of the fraction of rare alleles. Furthermore, we did not use any downweighting coefficient for this value when calculating the measure of dissimilarity of the theoretical values to the observed ones because we used high-quality resequencing data.

### Results

#### Data Sets and Definition of CNSs

We compiled 3 population genetic data sets of noncoding regions on the X chromosome in *D. melanogaster* (Glinka et al. 2003; Orenco and Aguade 2004; Ometto, Glinka et al. 2005) using a modified version of the PDA pipeline (Casillas and Barbadilla 2006). AFR is an African sample from Lake Kariba, Zimbabwe; EUR1 is a European sample from Sant Sadurní d’Anoia, Catalonia, Spain. Each data set consists of ~12 independently sampled alleles from ~100–250 noncoding regions each of length ~500 bp and includes both intronic and intergenic regions. Our automated pipeline employing the MUSCLE multiple alignment tool (Edgar 2004) can obtain almost the same results on a locus-by-locus basis as previously reported estimates of nucleotide diversity and divergence based on manually curated alignments, indicating that our alignments are of high quality (Supplementary File 1, Supplementary Material online). Alignments of the
noncoding regions analyzed in this study can be found at http://www.bioinf.manchester.ac.uk/bergman.

After filtering exonic, low-complexity, and poor quality regions at the ends of alignments, we analyzed a total of 249 loci for the AFR sample (160 intronic and 89 intergenic, spanning a total of >124 Kb), 256 loci for the EUR1 sample (166 intronic and 90 intergenic, spanning a total of >134 Kb), and 101 loci for the EUR2 sample (26 intronic and 75 intergenic, spanning a total of 83.5 Kb) (Table 1). Using an aligned reference sequence from the D. melanogaster Release 4 genome assembly (http://www.fruitfly.org/annot/release4.html), we partitioned noncoding DNA into CNS and non-CNS regions using the UCSC dm2 phastConsElements9way track (Hinrichs et al. 2006), which identifies the most conserved regions between GC content and rate of evolution previously observed (Haddrill et al. 2005). We confirmed that intergenic regions are more GC rich than non-CNS regions (Table 1). We also confirmed that intergenic regions are more GC rich overall than intronic regions (Haddrill et al. 2005), potentially because intergenic regions contain a higher proportion of GC-rich CNSs relative to introns. Elevated GC content in slowly evolving CNSs may explain the negative correlation between GC content and rate of evolution previously observed in Drosophila intronic regions (Haddrill et al. 2005).

Sequence Variation in CNS and Non-CNS Regions

We computed standard measures of nucleotide variation within D. melanogaster as well as divergence between D. melanogaster and D. simulans as shown in table 2. Table 2 also summarizes the numbers and density of SNPs, SNFs, IPs, and IFs in each population by genomic compartment (intronic vs. intergenic) and category of sites (CNS vs. non-CNS). Detailed information for each locus is provided in Supplementary File 2 (Supplementary Material online). Preliminary analysis revealed that the density of SNPs is ∼2-fold higher in alignment columns that are deleted in the outgroup species in all 3 data sets (χ² test, P < 2 × 10⁻⁸; Supplementary File 4 (Supplementary Material online), Control Test A, see also [Ometto et al. 2006]). Thus, for the purposes of this study, we excluded all SNPs in IFs to ensure that we are studying the orthologous set of

### Table 1
Summary of Data Sets and Conserved Noncoding Sequences

<table>
<thead>
<tr>
<th></th>
<th>AFR</th>
<th>EUR1</th>
<th>EUR2</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. loci total</td>
<td>160</td>
<td>166</td>
<td>26</td>
</tr>
<tr>
<td>No. loci non-CNS</td>
<td>160</td>
<td>166</td>
<td>26</td>
</tr>
<tr>
<td>No. loci CNS</td>
<td>143</td>
<td>152</td>
<td>26</td>
</tr>
<tr>
<td>Average no. chromosome total</td>
<td>11.79</td>
<td>11.92</td>
<td>12.58</td>
</tr>
<tr>
<td>Average no. chromosome non-CNS</td>
<td>11.78</td>
<td>11.91</td>
<td>12.58</td>
</tr>
<tr>
<td>Average no. chromosome CNS</td>
<td>11.80</td>
<td>11.92</td>
<td>12.58</td>
</tr>
<tr>
<td>No. aligned sites⁰ total</td>
<td>75,652</td>
<td>82,463</td>
<td>21,481</td>
</tr>
<tr>
<td>No. aligned sites⁰ non-CNS</td>
<td>60,039</td>
<td>65,257</td>
<td>14,779</td>
</tr>
<tr>
<td>No. aligned sites⁰ CNS</td>
<td>15,613</td>
<td>17,206</td>
<td>6702</td>
</tr>
<tr>
<td>No. ungapped sites⁵ total</td>
<td>67,139</td>
<td>74,514</td>
<td>19,557</td>
</tr>
<tr>
<td>No. ungapped sites⁵ non-CNS</td>
<td>52,068</td>
<td>57,858</td>
<td>13,065</td>
</tr>
<tr>
<td>No. ungapped sites⁵ CNS</td>
<td>15,071</td>
<td>16,656</td>
<td>6492</td>
</tr>
<tr>
<td>Average GC content⁶ total</td>
<td>38.33</td>
<td>38.24</td>
<td>40.69</td>
</tr>
<tr>
<td>Average GC content⁶ non-CNS</td>
<td>37.57</td>
<td>37.48</td>
<td>39.48</td>
</tr>
<tr>
<td>Average GC content⁶ CNS</td>
<td>40.98</td>
<td>40.87</td>
<td>40.43</td>
</tr>
<tr>
<td>Average no. of CNS/locus</td>
<td>2.73</td>
<td>2.89</td>
<td>4.65</td>
</tr>
<tr>
<td>Average CNS length (bp)</td>
<td>34.2</td>
<td>34.4</td>
<td>55.4</td>
</tr>
<tr>
<td>Average % CNS⁷</td>
<td>20.6</td>
<td>20.9</td>
<td>31.2</td>
</tr>
<tr>
<td>Average % CNS, ungapped⁸</td>
<td>22.4</td>
<td>22.4</td>
<td>33.2</td>
</tr>
</tbody>
</table>

Note.—Values for all 3 populations are given for introns (IT), intergenic (IG), and all noncoding (ALL) regions.

⁰ Total number of aligned sites (gapped + ungapped).

⁵ Total number of ungapped sites (ungapped in polymorphism and divergence).

⁶ Averages weighted by the number of chromosomes sampled and the number of analyzed sites (ungapped in polymorphism and divergence).

⁷ Calculated using the total number of aligned sites (gapped + ungapped).

⁸ Calculated using the total number of ungapped sites (ungapped in polymorphism and divergence).
nucleotides when estimating polymorphism and divergence in point mutations. Accordingly, densities of point mutations (either SNPs or SNFs) were calculated as the number of point mutations per unaligned site, whereas densities of indels (either IPs or IFs) were calculated as the number of indels per aligned site (Table 2).

As expected under both the mutational cold spot and functional constraint hypotheses, we found reduced polymorphism ($\pi$) and divergence ($K$) (Nei 1987) in CNS compared with non-CNS regions in all 3 data sets in both intronic and intergenic regions (Table 2). Densities of both point and indel variation were reduced in CNS compared with non-CNS regions, and this reduction was higher for point mutations relative to indels and stronger in intergenic regions relative to introns (Table 2). These results alone cannot differentiate between the mutational cold spot and functional constraint hypotheses because both predict a reduction in polymorphism and divergence in CNSs. However, evidence for differential selective constraints operating on CNS and non-CNS regions can be found in the facts that CNSs exhibit a more negative Tajima’s $D$ (Tajima 1989) (Table 2) and a stronger reduction of divergence relative to polymorphism (~60% in SNPs vs. ~80% in SNFs; ~50% in IPs vs. ~70% in IFs) (Table 2). Although the magnitude of the reduction of variation in CNSs relative to non-CNS regions is dependent on our use of the phastCons blocks to define CNSs, alternative definitions of CNS provide similar results (see Materials and Methods for details). We note that the systematic reduction in variation specifically observed in CNSs cannot be explained by sequencing error in the population samples because the same sequencing strategies were applied to both CNS and non-CNS regions.

### CNSs Are Selectively Constrained

To test whether the ratios of polymorphism (SNPs) to divergence (SNFs) differ significantly in CNS and non-CNS regions, we applied a modified MK test to point mutations in CNS and non-CNS regions. Separate tests were performed for all 3 data sets and for intronic and intergenic regions (Fig. 1). We present results here for all fragments
available on the X-chromosome, which did not differ from analyses using only noncoding sequences from regions of high recombination as defined in Andolfatto (2005) (results not shown). For point mutations (Fig. 1A–C), we observed very strong deviations from the null hypothesis ($\chi^2$ tests; all $P < 10^{-5}$) with a significant excess of polymorphisms relative to fixed differences in CNSs in all cases but on one with the least amount of data (intronic regions in the EUR2 sample, which exhibit the same trend). One can summarize an entire MK table as a ratio of ratios termed the neutrality index (Rand and Kann 1996) which is defined as $NI = (SNP_{CNS}/SNP_{non-CNS})/(SNF_{CNS}/SNF_{non-CNS})$. Assuming the ratio of polymorphism to divergence in non-CNS regions is closer to that expected under neutrality than in CNS regions, the $NI$ attempts to quantify whether the levels of divergence in CNSs are too low ($NI < 1$) or too high ($NI > 1$) relative to levels of polymorphism. In all data sets, we observe $NI > 1$ overall and in both intronic and intergenic regions, although $NI$ for intergenic regions is greater than $NI$ for intronic regions, indicating that point mutations in intergenic CNSs are less likely to fix relative to those in intronic CNS. These results reject the mutational cold spot model to explain the mode of CNS evolution in intronic and intergenic regions and support the interpretation that a significant proportion of new point mutations in CNSs are deleterious and do not contribute to divergence between species. It is important to emphasize that under the null hypothesis, the definition of CNS and non-CNS regions does not affect the predictions of equal ratios of polymorphism and divergence in both categories of sites. However, because CNSs are defined a priori by low divergence, our test is biased against detecting excess divergence, and thus, this analysis does not eliminate the action of positive selection on CNSs.

Evidence for functional constraint acting on CNSs can also be found in differences between the DAF spectra for CNS and non-CNS regions (Keightley, Kryukov et al. 2005; Drake et al. 2006). We observed a highly significant excess of low-frequency derived alleles for SNPs within CNSs relative to non-CNS regions in all 3 populations (Fig. 2A–C). For example, in the AFR population, 63% of SNPs inside CNSs are singletons, compared with only 46% of SNPs outside CNSs ($\chi^2$ test $= 47.5$, df $= 1$, $P = 5.45 \times 10^{-12}$). When data are partitioned into intronic and intergenic regions, all tests show a significant excess of low-frequency SNPs in CNSs relative to non-CNSs for all data sets, except for intronic regions in EUR2, which has the least amount of data (Supplementary File 3, Supplementary Material online, panels A–F). The excess of rare alleles in CNSs was more prominent for intergenic than intronic SNPs, consistent with the results of the MK test which indicate that fewer SNPs in intergenic CNSs go to fixation.
Similar results were obtained by testing for differences in the unpolarized minor allele frequency spectrum between CNS and non-CNS regions (results not shown), indicating that differences in the polarized DAF spectrum between these classes of sites are not strongly influenced by misinference of ancestral states (Hernandez et al. 2007). Overall, we find a strong signal that SNPs in CNS regions are specifically maintained at low frequency relative to SNPs in non-CNS regions, a finding which is inconsistent with the mutational cold spot hypothesis but is compatible with the presence of deleterious SNPs segregating at low frequency in functionally constrained CNSs.

One difficulty that arises from using non-CNS regions to detect purifying selection on CNSs is the fact that base composition differences in these 2 categories of sites (Table 1) (Drake et al. 2006). For example, recent changes in mutation biases could mimic the signature of selective differences between CNS and non-CNS regions and affect the tests of neutrality used here (Eyre-Walker 1997). In particular, a recent increase in the rate of G:C → A:T mutation has recently been suggested to explain nonequilibrium patterns of base composition evolution in D. melanogaster introns and intergenic regions (Kern and Begun 2005; Ometto et al. 2006) and may potentially cause an excess of SNPs in CNSs relative to non-CNS regions that would be restricted to low population frequency. Under this model of a recent increase in the rate of G:C → A:T mutation, differences in base composition between CNS and non-CNS regions cannot explain the reduction in polymorphism and divergence in CNSs because CNSs are more GC rich than non-CNS regions and G:C → A:T mutations occur at a higher rate than A:T → G:C mutations. Nevertheless, to control for any potential effects of biased mutation patterns that result from differences in base composition between CNS and non-CNS regions, we performed DAF tests for G:C → A:T and A:T → G:C mutations separately. We found an excess of rare alleles in CNSs when both G:C → A:T mutations and A:T → G:C mutations were considered separately in all samples except for A:T → G:C mutations in the EUR2 sample, which had the least amount of data but still showed the same trend (Supplementary File 4, Supplementary Material online, Control Test B). Thus, we can detect evidence for purifying selection on CNSs even when potential changes in base composition are controlled for, indicating that a recent increase in G:C → A:T mutation rate is unlikely to confound the conclusion that CNSs are selectively constrained. Moreover, if our interpretation that CNSs are constrained is correct, we suggest that the excess of low-frequency G:C → A:T mutations and other aspects of non-equilibrium base composition evolution in Drosophila may in fact be a consequence of the preservation of functional GC nucleotides in noncoding DNA by purifying selection, rather than evidence for a change in mutation rate or biased segregation as suggested previously (Kern and Begun 2005; Galtier et al. 2006; Ometto et al. 2006).

We also investigated 2 possible alternatives for the striking differences we observed between CNS and non-CNS regions that might result from alignment error. The first possibility is that indels may create low-quality regions of multiple alignments, causing SNPs and SNFs to accumulate in the vicinity of gaps and that the differences we observe in point mutations may be a byproduct of differences in indel rates between CNS and non-CNS regions. To control for this possibility, we repeated our analyses excluding CNS and non-CNS regions that contain indels (Supplementary File 4, Supplementary Material online, Control Test C). All MK tests remained highly significant when all regions with either IPs or IFs were excluded. Likewise, all DAF tests remained highly significant when regions with IPs were excluded. We also explored a related source indel-associated alignment error that might result from 2 indels of exactly the same length in essentially the same position of a single sequence, one an insertion and one a deletion. Two such indels may collapse in the alignment and thus result in a run of consecutive substitutions. Consecutive SNPs or SNFs may also occur through complex mutational events that replace more than a single nucleotide (Averof et al. 2000; Haag-Liautard et al. 2007). We repeated our analyses excluding consecutive SNPs or SNFs (all SNPs followed or preceded by another SNP in the alignment or all SNFs followed or preceded by another SNF; Supplementary File 4, Supplementary Material online, Control Test D). All tests remained highly significant, showing no evidence of this type of alignment error. These results demonstrate that indel-associated misalignment cannot explain the differences in point mutations between CNS and non-CNS regions.

Indels themselves, in contrast to point mutations, show less striking differences in the ratios of polymorphism to divergence in CNSs relative to non-CNS regions (Fig. 1D–F). In contrast to previous analysis of indels in an MK framework that contrast SNPs in silent sites with either insertions or deletions in introns (Presgraves 2006), here we directly contrast levels of polymorphism and divergence for all indels in CNSs to all indels in non-CNS regions. The NI values we observe for indels are still above 1, consistent with purifying selection on indels in CNS regions but are always lower than the corresponding values for SNPs. MK tests are only significant for the combined intronic plus intergenic regions in the 2 larger data sets (AFR and EUR1) and intronic regions in EUR1. Likewise, we observed no strong differences within species in the DAF spectra between indels in CNS and non-CNS regions, with nonsignificant DAF tests for all populations overall (Fig. 2D–F) and for intronic and intergenic regions separately (Supplementary File 3, Supplementary Material online, panels G–L). We note that potential differences in CNS and non-CNS regions are not diluted or obscured by a high rate of indels due to simple slippage because low-complexity repeat regions were filtered from the data.

Because there are 8-fold fewer IPs than SNPs and 5-fold fewer IFs than SNFs in our data set, the lack of strong differences in CNS and non-CNS regions for indels may simply result from low power to reject the null hypothesis. We tested if differences as strong as those observed for point mutations could also be observed in the smaller sample of indels by rescaling the point mutation data (which show significant results) to the sample sizes observed for indels and repeated the MK and DAF tests (Supplementary File 4, Supplementary Material online, Control Test E). Specifically, numbers of SNPs and SNFs were reduced to the numbers of observed IPs and IFs, while maintaining the observed ratio of SNPs to SNFs in contingency tables of
Fig. 2.—DAF distributions for polymorphic point mutations and indels in the AFR, EUR1, and EUR2 data sets. Dark gray and light gray bars represent data for all SNPs/IPs within CNSs and non-CNS regions, respectively. Reported P values are for 2-sample Komogorov–Smirnov tests.
the MK tests or the frequency bins of the DAF distributions. MK tests and DAF tests using rescaled point mutation data sets for combined intronic and intergenic were still highly significant in all 3 populations ($P < 0.01$). Assuming the same degree of purifying selection is acting on point mutations and indels, it is unlikely that low power is the main cause of the lack of significance for indels in MK and DAF tests between CNS and non-CNS regions. Given the fact that we find strong evidence that CNSs are constrained for point substitutions, we do not interpret these results as support for an indel cold spot hypothesis to explain the mode of CNS evolution. Rather, we interpret the lack of strong differences in indel evolution between CNS and non-CNS regions as evidence for spatial constraints acting on both CNS and non-CNS sequences (see Discussion).

**Estimating the Effects of Deleterious Mutations on CNSs**

Deleterious alleles that are not immediately purged by natural selection are expected to be maintained at low frequencies and not go to fixation in natural populations. The results that CNSs show a significant excess of polymorphism to divergence and an enrichment of low-frequency alleles indicate that CNSs do indeed harbor more deleteriousSNPs relative to the non-CNS spacer sequences between them. Can we infer at what frequency deleterious SNPs in CNSs are segregating or the magnitude of fitness effects acting on SNPs in CNS regions? To address these questions, we restricted our analysis to the AFR data set, which is taken from a population of *D. melanogaster* that is assumed to be closest to ancestral conditions (Glinka et al. 2003; Thornton and Andolfatto 2006) and for which we have the largest sample of polymorphism to estimate properties of deleterious alleles in CNSs.

We found evidence that deleterious SNPs were restricted almost exclusively to the singleton class in the AFR data set by an analysis of the effects of removing singletons on the results of the MK test. This procedure has been used in tests for positive selection assuming that if deleterious mutations are restricted to sites with low population frequencies, derived alleles present only once in the sample should preferentially be enriched for deleterious mutations (Fay et al. 2002; Andolfatto 2005). In all cases, NIs remained greater than one after the removal of singleton SNPs but decreased relative to values for the total data set (Supplementary File 4, Supplementary Material online, Control Test F). For introns, intergenic regions and the combined data set, the resulting MK tests yielded nonsignificant differences in the ratio of nonsingleton polymorphism to divergence ratio, suggesting that common SNPs in CNSs in the AFR data set are effectively neutral. This result indicates that virtually all signals of deleterious alleles segregating in CNSs are restricted to rare SNPs. In addition, this result also indicates that there is no evidence for positive selection acting on CNSs, even when the confounding effects of purifying selection are taken into consideration (see Discussion). We note that similar tests on the EUR1 and EUR2 populations yielded NI $> 1$ which remained significant in the absence of singletons, suggesting that purifying selection may be acting even among common SNPs in these data sets (Supplementary File 4, Supplementary Material online, Control Test F). Further evidence that SNPs are restricted to low frequency can be found in the ratio of SNPs in CNSs relative to non-CNSs at different derived allele frequencies. We find that in the AFR population, the SNP$_{CNS}$/SNP$_{non-CNS}$ ratio is significantly heterogeneous across all 10 DAF classes ($\chi^2$ test; $P < 4.17 \times 10^{-5}$), consistent with categorical differences in the complete DAF spectra shown above. However, when low-frequency SNPs are removed, the remaining 9 DAF classes show no significant heterogeneity in their SNP$_{CNS}$/SNP$_{non-CNS}$ ratio ($P > 0.07$). This result indicates that common SNPs in CNS and non-CNS regions have similar DAF spectra and that the majority of deleterious SNPs in CNSs are restricted to a DAF of less than 10% in the population samples.

To quantify the observed differences in selective pressure acting on CNSs relative to non-CNS regions, we estimated the distribution of selection coefficients in these regions using an exhaustive computational search method developed by Kryukov et al. (2005) (see Materials and Methods). Unlike other methods to estimate selection coefficients from population genetic data, no explicit distribution of selection coefficients is assumed by this method. Rather, distributions of selection coefficients ($s$) are modeled by histograms, where bins represent the fraction of sites under a given selection coefficient. All possible distributions are enumerated under a model of weakly deleterious evolution, and the fit of the data is evaluated for each possible distribution. Assuming an effective population size ($N_e$) for *D. melanogaster* of 10$^6$ (Kreitman 1983), our results indicate that best fit of the data is to a distribution where 80–85% of sites in CNSs are subject to weak purifying selection ($s \sim 10^{-5}$) and the remaining 15–20% of CNS sites are effectively neutral ($s \sim 10^{-7}$). Likewise, using the method of Piganeau and Eyre-Walker (2003), which assumes an underlying gamma distribution of selection coefficients, we obtain an average strength of selection on CNSs of $N_e s = 30.7$ (95% confidence interval: 13–117) with a shape parameter of $\beta = 0.31$ (95% confidence interval: 0.22–0.42). These results indicate that purifying selection on *Drosophila* CNSs exceeds the boundaries of nearly neutral evolution. However, the strength of purifying selection for bulk noncoding DNA in *Drosophila* may be on the boundaries of nearly neutral evolution because non-CNS regions are more abundant but less constrained than CNS regions. Thus, the evolution of *Drosophila* noncoding DNA in general may be sensitive to changes in $N_e$, both across time through changes in census population size (Keightley, Kryukov et al. 2005; Keightley, Lercher, and Eyre-Walker 2005) or across the genome such as in regions of reduced recombination (Haddrill et al. 2007). Our results also indicate that purifying selection is stronger (Chen et al. 2007) and affects more sites (Kryukov et al. 2005) in *Drosophila* CNSs than for mammalian CNSs.

**Discussion**

The major conclusion of this work is that highly CNSs in *Drosophila* are maintained by purifying selection and are not simply regions of the genome with extremely low
mutation rate as predicted by the mutation cold spot hypothesis. In addition, we find that the strength of purifying selection acting to maintain CNSs is moderately strong, with most nucleotides in CNSs being preserved by selection coefficients 10- to 100-fold greater than the reciprocal of the effective population size. The conclusion that Drosophila CNSs are maintained by purifying selection supports previous analyses that have made this assumption based on reduced rates of molecular evolution (Bergman and Kreitman 2001; Siepel et al. 2005; Halligan and Keightley 2006). Specifically, our results support the UCSC phastCons highly conserved track (Siepel et al. 2005) as being able to identify selectively constrained regions of the D. melanogaster genome. These findings in Drosophila closely parallel those recently found for mammalian CNSs and predicted micro-RNA binding sites using population genetic data from human SNP studies (Keightley, Kryukov et al. 2005; Kryukov et al. 2005; Chen and Rajewsky 2006; Drake et al. 2006; Chen et al. 2007). Thus, purifying selection may be a general force acting to maintain highly CNSs in metazoan genomes. As no population genomic evidence (or molecular mechanism) has yet been put forth to support the mutation cold spot hypothesis, similar results in disparate organisms such as flies and mammals (together with the nonrandom spacing of CNSs in flies and worms [Bergman et al. 2002; Webb et al. 2002]) argue against the general likelihood that CNSs will be shown to be mutational cold spots in any organism. Further studies in disparate taxa including plants and other metazoans will be necessary to confirm the generality of this conclusion.

Selective constraint on CNSs is consistent with the large body of evidence from experimental studies that highly CNSs in Drosophila are often associated with regulatory function, such as cis-regulatory elements or noncoding RNAs (Bergman et al. 2002; Enright et al. 2003; Lai et al. 2003; Costas et al. 2004). Furthermore, several facts reported over the last decade collectively point to widespread selective constraint operating on Drosophila noncoding DNA. First, unconstrained noncoding DNA is quickly deleted from the Drosophila genome (Petrov et al. 1996; Petrov and Hartl 1998). This process is predicted to purge the fly genome of “junk” DNA, making nonfunctional sequence-like pseudogenes rare (Petrov et al. 1996; Harrison et al. 2003) and enriching noncoding DNA that remains in the genome for functional elements. Second, genes with complex transcriptional regulation have longer flanking intergenic regions (Nelson et al. 2004), suggesting that the mere presence of noncoding DNA in Drosophila may imply function. Third, for both intronic and intergenic DNA, the rate of molecular evolution between closely related Drosophila species decreases with increasing noncoding sequence length (Haddrill et al. 2005; Halligan and Keightley 2006), consistent with the interpretation that long noncoding regions may have increased functional constraints. Fifth, long introns have a higher proportion of CNS sequences and genes with CNSs in their introns have more complex regulation (Petit et al. 2007). Finally, adaptive substitutions may be commonplace in both intronic and intergenic regions (Andolfatto 2005), which can only occur if the density of functional nucleotide sites in noncoding DNA is high.

Given that constraints on noncoding sequences are widespread in Drosophila, and the possibility that adaptive substitution occurs in noncoding DNA, it is worth considering whether flanking non-CNSs are appropriate control sequences to detect selection on CNSs. Halligan and Keightley (2006) report a method to measure constraints on regions of genomic DNA as the reduction in the rate of substitution relative to that expected based on putatively unconstrained sequences, such as 4-fold degenerate silent sites. Despite widespread evidence for weak selection on silent sites in Drosophila (Shield et al. 1988; Akashi 1995), we applied this method to evaluate if stronger primary sequence constraints act on non-CNS regions relative to 4-fold silent sites, and, if so, what effect this may have on our conclusion that CNSs are selectively constrained and not mutational cold spots. As shown in table 3, we find that non-CNS regions exhibit an ~20% reduced rate of sequence evolution relative to 4-fold silent sites, indicating that primary sequence constraints act on non-CNS regions. Levels of constraint on CNSs are estimated to be ~85% by the Halligan and Keightley (2006) method (consistent with results above using the Kryukov et al. (2005) method), and thus, we infer that selective constraints operating on non-CNS regions affect ~4 times fewer sites than in CNS regions. Constraints on non-CNS regions are perhaps not surprising because even the most rigorous definition of CNSs (Siepel et al. 2005) is unlikely to capture all functionally constrained noncoding DNA, especially those which arise through lineage-specific gain-of-function events. Nevertheless, as constraints on non-CNS regions would only tend to obscure differences between CNS and non-CNS categories by making their patterns of evolution more similar, our conclusion that CNSs are selectively constrained is conservative with respect to the null hypothesis that they are mutational cold spots. However, by using non-CNS regions as putatively unconstrained control sequences, the proportion of sites under constraint in CNSs and the magnitude of their selective effects are likely to be underestimated in our analysis.

Conversely, if adaptive substitutions preferentially occur in non-CNS regions, it may be possible that the rate of substitution in non-CNS regions is elevated relative to the unconstrained neutral substitution rate, which could cause us spuriously to reject the mutational cold spot hypothesis. To evaluate if any signature of adaptive substitution is detected in our data set, we conducted MK tests on CNS and non-CNS regions as selected classes of sites using 4-fold silent sites as putatively unconstrained controls. For these analyses, we reprocessed the sequence data reported in Andolfatto (2005) using PDA to extract SNPs and SNFs for 4-fold degenerate silent sites only. As is observed using sites from linked non-CNS regions above (Fig. 1), the NI for CNSs remains significantly greater than one when using partially linked 4-fold silent sites as controls (Table 3). As before, when we removed singletons to reduce the confounding effects of deleterious mutations present in low-frequency alleles, we found no departure from neutral expectations between CNS and 4-fold silent sites (Table 3). Additionally, SNPs in CNSs are more skewed to lower frequencies than the SNPs in 4-fold silent sites (Supplementary File 5, Supplementary Material online) as has been
shown recently for bulk noncoding DNA in *D. melanogaster* (Andolfatto 2005; Mustonen and Lassig 2007) and *Drosophila miranda* (Bachtrog and Andolfatto 2006). These results confirm our main claims that CNSs are selectively constrained with deleterious SNPs restricted to low frequencies and clearly demonstrate that the conclusion that CNSs are functionally constrained does not depend on our use of linked non-CNS regions as controls.

Intriguingly, we find evidence for positive selection in non-CNS regions when 4-fold silent sites are used as unconstrained controls, both when all sites are used or when singletons are removed (Table 3). The same trends are observed when CNS and non-CNS regions are pooled into bulk noncoding DNA, as in Andolfatto (2005). Thus, we confirm the results of Andolfatto (2005) for the putative signature of adaptive substitution on noncoding DNA when 4-fold silent sites are used as controls, even despite differences in methods of estimating polymorphism and divergence. Somewhat counterintuitively, perhaps, the 4-fold silent sites are used as controls, even despite differences in methods of estimating polymorphism and divergence. Somewhat counterintuitively, perhaps, the signature of adaptive substitution in *Drosophila* noncoding DNA does not appear to occur in CNSs, which might be expected to contain the functional elements that are targets for positive selection. Conversely, the signal of excess substitution in bulk noncoding DNA relative to silent sites appears to be restricted to the non-CNS regions that are divergent among other insect species. Using the method of Smith and Eyre-Walker (2002), we estimate that ~18% of substitutions in non-CNS regions are driven to fixation by positive selection relative to neutral expectations (Table 3), which is compatible with previous estimates for bulk noncoding DNA (Andolfatto 2005). Selective constraints on CNSs coupled with the signature of adaptive substitution in non-CNS regions might be expected under the model of stabilizing selection proposed by Ludwing et al. (2000) for the *even-skipped* stripe 2 enhancer, whereby loss of ancestral transcription factor binding sites in CNS regions may lead to compensatory adaptive fixations of lineage-specific binding sites in non-CNS regions (e.g., *bcd-3*) that restore *cis*-regulatory function.

In summary, using 4-fold silent sites as another class of putatively unconstrained neutrally evolving sequences, we find evidence that both constraint and adaptation influence rates of substitution in non-CNS regions. Assuming additive influences of these 2 opposing forces (Andolfatto 2005; Halligan and Keightley 2006), we find that the estimated proportion of mutations purged by purifying selection in non-CNS regions (C ~18%) is approximately the same as the estimated proportion of substitutions that have been driven to fixation by positive selection (α ~17%). Assuming no adaptive evolution in CNS regions, these results also imply that the proportion of functionally relevant nucleotides in non-CNS regions is FRN = C + (1 − C)α ≈ 33%, or ~2.5-fold less than the proportion of functional sites in CNS regions (85%). Thus, given that the majority of *Drosophila* noncoding DNA is found in non-CNS regions (65–80%, Table 1), the number of functional sites in CNS and non-CNS regions is approximately equivalent. Although less densely packed than in CNS regions, the greater number of functional sites in non-CNS regions coupled with their relaxed selective constraints may explain why these regions appear to be the most likely targets for positive selection in noncoding DNA. Further work will be necessary to determine if (and how) the distribution of positive and negative selection coefficients acting on polymorphisms in non-CNS regions affects their utility in testing the mutational cold spot hypothesis. Likewise, the influence of alternative CNS definitions on quantitative estimates of constraint and adaptation in CNS and non-CNS regions needs to be investigated further. Nevertheless, the direct result for a significant constraint and skew toward rare alleles in CNSs using 4-fold silent sites as controls (see above) indicates that our main claim that CNSs are selectively constrained and not mutational cold spots is unaffected by the potentially confounding effects of either constraint or adaptive substitution in non-CNS regions.

Selective constraints may also operate on the length of noncoding DNA as well as on primary sequence. This possibility may explain why differences in the ratio of polymorphism to divergence or DAF spectrum between CNS and non-CNS regions are not as strong for indels as they are for point mutations, assuming that spatial constraints act on both CNS and non-CNS regions. Mechanistically, this might be expected to occur if CNSs represent the constraints imposed by transcription factor binding sites that could be disrupted by both point and indel mutations, whereas non-CNS regions that act to position neighboring binding sites would be affected only by indel mutations (Ondek et al. 1988). Spatial constraints have been argued previously in *Drosophila* noncoding DNA based on the non-random distribution of CNSs and the strong correlation in the length between neighboring CNSs across divergent species (Bergman et al. 2002). Ometto, Stephan, and De Lorenzo et al. (2005) have also argued for spatial constraints acting within *Drosophila* noncoding DNA based on the ratio of insertions to deletions and the size distribution of deletions segregating in natural populations. More recently, Lunter et al. (2006) have inferred that spatial constraints act on human noncoding DNA based on the distribution of indel positions in

<table>
<thead>
<tr>
<th>Constraint</th>
<th>α</th>
<th>All polymorphisms</th>
<th>Excluding singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNS vs. 4-fold silent</td>
<td>0.84525</td>
<td>0.18360</td>
<td>0.02623</td>
</tr>
<tr>
<td>Non-CNS vs. 4-fold silent</td>
<td>0.13860</td>
<td>0.17577</td>
<td>0.684</td>
</tr>
<tr>
<td>CNS + non-CNS vs. 4-fold silent</td>
<td>0.36616</td>
<td>0.13443</td>
<td>0.00012</td>
</tr>
</tbody>
</table>

Note.—The reported α is based on all polymorphisms including singletons. CNS/non-CNS ratios for polymorphism and divergence are summarized by the neutrality index as NI = (SNFCNS/SNPnon-CNS)/(SNFCNS/SNFCNS), and P values are computed using a χ² test of independence. Tests are performed both including and excluding singletons.
alignments with other mammalian species, and Sun et al. (2006) have argued for spatial constraints between neighboring vertebrate ultraconserved regions. A lack of strong differences in indel evolution between CNS and non-CNS regions may alternatively arise because of technical reasons such as the fact that phastCons permits indels in CNSs, blurring real differences in the pattern of indel evolution between these categories. Another possible explanation is that CNSs are mutation cold spots for indels, although this seems unlikely if CNSs are selectively constrained for point mutations as we argue here. If selective constraints are indeed operating on indels in both CNS and non-CNS regions, the lack of strong differences in the ratio of polymorphism to divergence or DAF spectrum suggests that the strength of selection against indels in noncoding DNA may be stronger than for point mutations because only relatively weak purifying selection allows an accumulation of low-frequency variants in nature (as is observed for SNPs). Future progress in detecting evidence for spatial constraints and quantifying the mode and strength of selection acting on both indels and point mutations in noncoding DNA will shed light on the functions encoded in this most abundant, yet least explored, territory of metazoan genomes.

Supplementary Material

Supplementary Files 1–5 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Douda Bensasson, Francesco Catania, Manolis Dermitzakis, and Dan Halligan for valuable discussions and Peter Andolfatto, Francesco Catania, Dan Halligan, Martin Kreitman, Gregory Kryukov, Alfredo Ruiz, Josh Shapiro, Kevin Thornton, and 3 anonymous reviewers for comments on the manuscript. We also thank Gregory Kryukov for guidance in implementing his method to estimate selection coefficients, Adam Eyre-Walker for providing programs and assistance to estimate the strength of selection, and Dan Halligan for providing programs and data sets to estimate constraint using 4-fold silent sites. We are grateful to the Washington University Genome Sequencing Center, the Broad Institute, Agencourt Biosciences, and the laboratories of Monserrat Aguade, Peter Andolfatto, David DeLorenzo, and Wolfgang Stephan for producing the sequence data used in this work. S.C. was supported by a Marie Curie fellowship from the European Commission (CHPMT-GH-01-00285-13), by the Ministerio de Ciencia y Tecnología (BES-2003-0416), and grant BFU2006-08640/BMC from the Ministerio de Educación y Ciencia, Spain.

Literature Cited


Casillas S, Barbudilla A. 2006. PDA v.2; improving the exploration and estimation of nucleotide polymorphism in large datasets of heterogeneous DNA. Nucleic Acids Res. 34:W632–W634.


William Martin, Associate Editor

Accepted July 16, 2007