Positive and Negative Selection on Noncoding DNA Close to Protein-Coding Genes in Wild House Mice

Athanasios Kousathanas,*1 Fiona Oliver,1 Daniel L. Halligan,1 and Peter D. Keightley1

1Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, Edinburgh, United Kingdom

*Corresponding author: E-mail: a.kousathanas@sms.ed.ac.uk.

Associate editor: Michael Nachman

Abstract

During the past two decades, evidence has accumulated of adaptive evolution within protein-coding genes in a variety of species. However, with the exception of Drosophila and humans, little is known about the extent of adaptive evolution in noncoding DNA. Here, we study regions upstream and downstream of protein-coding genes in the house mouse Mus musculus castaneus, a species that has a much larger effective population size ($N_e$) than humans. We analyze polymorphism data for 78 genes from 15 wild-caught M. m. castaneus individuals and divergence to a closely related species, Mus famulus. We find high levels of nucleotide diversity and moderate levels of selective constraint in upstream and downstream regions compared with nonsynonymous sites of protein-coding genes. From the polymorphism data, we estimate the distribution of fitness effects (DFE) of new mutations and infer that most new mutations in upstream and downstream regions behave as effectively neutral and that only a small fraction is strongly negatively selected. We also estimate the fraction of substitutions that have been driven to fixation by positive selection ($\alpha$) and the ratio of adaptive to neutral divergence ($\omega_a$). We find that $\alpha$ for upstream and downstream regions ($\sim 10\%$) is much lower than $\alpha$ for nonsynonymous sites ($\sim 50\%$). However, $\omega_a$ estimates are very similar for nonsynonymous sites ($\sim 10\%$) and upstream and downstream regions ($\sim 5\%$). We conclude that negative selection operating in upstream and downstream regions of M. m. castaneus is weak and that the low values of $\alpha$ for upstream and downstream regions relative to nonsynonymous sites are most likely due to the presence of a higher proportion of neutrally evolving sites and not due to lower absolute rates of adaptive substitution.

Key words: adaptive evolution, effective population size, wild mice, selective constraint, upstream and downstream regions.

Introduction

In recent years, the search for evidence of adaptive evolution at the molecular level has been at the forefront of genetics research. A principal motivation has been to identify regions of the genome that have experienced adaptive evolution because this might provide clues to their functional importance and may be informative about the features that make each species unique. There have been a wealth of studies focusing on protein-coding genes. Studies in Drosophila, employing variants of the McDonald–Kreitman (MK) test (McDonald and Kreitman 1991), suggest a high proportion of adaptive amino acid substitutions ($\alpha$) (50% or more; Smith and Eyre-Walker 2002; Sawyer et al. 2003; Bierne and Eyre-Walker 2004; Welch 2006; Shapiro et al. 2007; Eyre-Walker and Keightley 2009), whereas in humans similar studies have produced low estimates of $\alpha$ (0–20%; The Chimpanzee Sequencing and Analysis Consortium 2005; Zhang and Li 2005; Boyko et al. 2008; Eyre-Walker and Keightley 2009). These contrasting results between Drosophila and humans have been interpreted to be a consequence of different effective population sizes ($N_e$), that is, the small $N_e$ of the hominid lineage could have resulted in a reduced efficacy of natural selection. Other evidence points to a positive relationship between $\alpha$ and recent $N_e$. For example, $\alpha$ for protein-coding genes has been estimated to be 50% or more in enteric bacteria, which have a large $N_e$ (Charlesworth and Eyre-Walker 2006), close to zero in Arabidopsis (A. lyrata and A. thaliana), which has small $N_e$ (Fexe et al. 2008), and about 40% in Capsella grandiflora, a species closely related to A. thaliana with a larger $N_e$ (Slotte et al. 2010). A recent study of the house mouse Mus musculus castaneus, which has $N_e$ comparable with Drosophila, also produced a high estimate of $\alpha$ for protein-coding genes ($\sim 50\%$; Halligan et al. 2010), again suggesting that $N_e$ is a strong determinant of the efficacy of positive selection on amino acid-changing mutations across taxa.

Estimates of the frequency of adaptive nucleotide substitution in noncoding DNA are currently restricted to Drosophila and humans. In Drosophila, estimates of $\alpha$ for 5′ and 3′ untranslated regions (UTRs) are nearly as high as for protein-coding genes (i.e., 50% or more; Kohn et al. 2004; Andolfatto 2005; Haddrill et al. 2008) and for introns and intergenic regions are relatively low ($\sim 10–20\%$; Andolfatto 2005; Haddrill et al. 2008). In humans, estimates of $\alpha$ for noncoding regions upstream and downstream of protein-coding genes are close to zero (Keightley et al. 2005; Eyre-Walker and Keightley 2009).

In this study, we investigate positive and negative selection operating on noncoding regions upstream and downstream of the protein-coding genes in a sample of the house mouse M. m. castaneus that were previously studied by Halligan et al. (2010). We study regions upstream and
downstream of protein-coding genes that are known to be enriched for regulatory elements (Xie et al. 2005; Veyrieras et al. 2008) and are implicated in the control of transcription and translation (Gray and Wickens 1998; Shabalina and Spiridonov 2004). Previous studies in murids have shown that ∼30% of sites in 5′- and 3′-UTRs and ∼10% of sites that are within the first 3–5 kb upstream and downstream of the transcription start and stop codon, respectively, are subject to negative selection (Keightley et al. 2005; Gaffney and Keightley 2006). Here, we perform a more thorough investigation of negative selection operating in upstream and downstream regions by estimating the full distribution of fitness effects (DFE) of new mutations. We then proceed to investigate positive selection by estimating α using a method that attempts to account for the presence of slightly deleterious mutations: The DFE is used to predict the expected divergence between two species caused by the fixation of neutral and slightly deleterious mutations, and this is compared with the observed divergence (Eyre-Walker and Keightley 2009). The difference between the observed and expected divergence is used to estimate the amount of adaptive divergence and α. We also estimate ωa, the rate of adaptive divergence relative to neutral divergence, which allows us to better compare rates of adaptive evolution between species, by controlling for the effects of Na on the numbers of effectively neutral substitutions (Gossmann et al. 2010).

Materials and Methods

Sampling of Mice

We analyzed 15 M. m. castaneus individuals sampled from four regions south of the Himalayas in the Himachal Pradesh state of India. We also generated sequence data from a Mus famulus individual originating from India that was previously obtained from the Montpellier wild mice genetic repository (http://www.isem.cnrs.fr/spip.php?article4777). A more detailed description of the sampling of the mice can be found in Halligan et al. (2010).

Choice of Genes

We analyzed 78 autosomal genes from M. m. castaneus whose human orthologs have also been sequenced as part of the Environmental Genome Project (EGP) (Livingston et al. 2004). The EGP data set is enriched for genes that are involved in pathways for DNA repair, cell cycle control, drug metabolism, and apoptosis and therefore non random (Livingston et al. 2004). The genes were chosen if there were African human polymorphism data available; this enabled us to more directly compare the results in humans with mice.

As part of this study, we successfully sequenced upstream and downstream regions for 49 and 51 genes, respectively, in 15 M. m. castaneus individuals and one M. famulus individual. We designed primers to amplify the upstream region of each gene, which lies approximately up to 500 bp upstream of the first codon of the first exon as annotated in the reference mouse genome. Similarly, to amplify the downstream region of each gene, we designed primers that captured the region that lies approximately up to 500 bp downstream of the stop codon of the last exon in the reference mouse genome. We chose to sequence ∼500 bp upstream and downstream of protein-coding genes because evidence from studies of selective constraint, regulatory motifs, and expression-quantitative trait loci suggest that there is high density of functional elements in these regions (Xie et al. 2005; Gaffney and Keightley 2006; Veyrieras et al. 2008). Additionally, the interpretation of sequences further upstream and downstream from these regions was made more difficult by frequent indel variation, making calling of single nucleotide polymorphisms (SNPs) problematic. In supplementary table S1, Supplementary Material online, we give details of the genes analyzed in this study and the upstream and downstream regions that we successfully sequenced. We chose not to restrict our analyses to those genes for which upstream, downstream, exonic, and intronic sequence data were available because the smaller samples for intronic (65), upstream (49), and downstream (51) site classes are unbiased in relation to the larger data set for exonic sequence (78). The set of genes for which we have upstream and downstream sequence data overlap in almost all cases. Additionally, in analyses where a putatively neutral (i.e., synonymous or intronic sites) and a selected class (i.e., nonsynonymous, upstream, or downstream) were required, we only analyzed genes for which both the neutral and the selected class were sequenced. We have also updated the data set of Halligan et al. (2010) with new exonic and intronic sequence of the 78 genes. We reanalyzed the exonic and intronic data sets because we had ∼20% new exonic sequence data and ∼60% new intronic sequence data. Supplementary table S2, Supplementary Material online, shows in detail the differences between the Halligan et al. (2010) data set and the updated data set used in this study.

Sequencing

GoTaq DNA polymerase (Promega) was used in touchdown-style polymerase chain reactions (PCRs): an initial denaturation step of 95 °C for 15 min, followed by 28 cycles of 95 °C for 30 s, 62 °C for 45 s (reducing by 0.5 °C every cycle), 72 °C for 2 min, then 12 cycles of 95 °C for 30 s, 52 °C for 45 s, and 72 °C for 2 min, with a final extension step at 72 °C for 10 min. ExoSAP-IT (USB) was used for the purification of PCR products. If we obtained nonspecific PCR products, we designed new primers to try to increase the specificity. Sequencing was done using Big Dye Terminator Sequencing Kits (Applied Biosystems, Foster City, CA) on an ABI Prism 3730 DNA Analyzer, and both forward and reverse sequences were generated. CodonCode Aligner version 2.0.6 was used to analyze and detect variants (http://www.codoncode.com/aligner/). We used Phred computer program as employed in CodonCode Aligner to assess sequence quality. Sequences had an average Phred score of >60. All sequence traces were manually checked. Sites with a Phred score <30, which could be low quality sequence or heterozygotes, were manually checked but were not automatically excluded in order to avoid excluding
heterozygotes. Where sequence was found to be too low quality, due to multiple indels or repetitive regions, new amplicons were produced on either side of the difficult to sequence area. Such difficult to sequence areas were replaced by 'N's before further analysis. Finally, we generated alignments between the 15 M. m. castaneus, the M. famulus individual, and the M. m. musculus reference sequence using CodonCode Aligner, and all alignments were checked by eye before further analysis.

Dealing with Heterozygous Indels

Heterozygous indels can make SNP calls from sequence traces problematic, and we took several steps to ensure that SNP calls were accurate. First, we obtained sequences from both strands for every amplicon. This allowed us to accurately call SNPs up to a heterozygous indel without having to interpret the sequence traces after the occurrence of heterozygous indel. Second, for many of the genomic regions that we sequenced, we designed multiple overlapping amplicons to cover the region to ensure that SNP calls on either side of heterozygous indels were accurate. Third, we used CodonCode Aligner to call SNPs. This software has specialized algorithms that are able to call SNPs, following heterozygous indels, by interpreting the out-of-sync signals from both strands in the sequence traces. Fourth, we also manually checked all sequence traces and SNP calls by eye, including manually disentangling out-of-sync signals in sequence traces where heterozygous indels were encountered. Fifth, we also very carefully checked SNPs at sites where the genotype frequencies of the individuals were not at Hardy–Weinberg equilibrium (because this signal could be indicative of heterozygotes being miscalled as homozygotes).

Sequence Processing

We obtained orthologous Rattus norvegicus sequences for each amplicon using a reciprocal best-hits Blast approach. To do this, we Blasted the mouse reference sequence (mm9) for each amplicon, plus 200 bp of flanking DNA, against two different assemblies (labeled “standard” and “alternative”) of the rat genome and searched for a reciprocal best hit. If we failed to find a reciprocal best hit for the standard assembly, we searched the alternative assembly. Both assemblies were downloaded from the University of California–Santa Cruz (UCSC) genome browser; the standard was produced by the Baylor College of Medicine Human Genome Sequencing Center (BCM-HGSC) as part of the Rat Genome Sequencing Consortium, and the alternative was produced by Celera Genomics. If we failed to identify an ortholog via the reciprocal best-hits approach, we checked the relevant section in the “multiz30way” whole-genome sequence alignments of 30 vertebrates (http://genome.ucsc.edu/). We considered sequences to be orthologous if the sequence of interest was located entirely within a single unbroken alignment for mouse and rat. We realigned all alignments obtained by either method using MAVVID (Bray and Pachter 2004) and then subsequently checked them all by eye. Any obviously misaligned sections identified when checking by eye were masked from any further analysis. Using this procedure, putatively orthologous rat sequences were obtained for at least part of every mouse amplicon.

We constructed alignments for each amplicon between the mouse reference (mm9) sequence, the sequences from all M. m. castaneus individuals, M. famulus, and rat. We annotated sites according to the mouse reference genome into the following categories: 5′, 3′, intron or coding. Within the coding category, sites were categorized as first, second, or third positions as well as the level of degeneracy in the genetic code (zero-fold-, 2-fold-, or 4-fold degenerate). We excluded potential splice sites of introns (defined as the first 6 bp or last 16 bp of an intron) from any analysis. We also categorized sites on the basis of their CpG-prone status (defined as being preceded by a C or followed by a G in any species).

Summary Statistics

We assume that segregating polymorphisms are biallelic. If there were more than two alleles segregating at a site, we only consider the two most frequent alleles. We calculated two statistics for nucleotide diversity, \( \pi \) and Watterson’s \( \theta_w \).

Let the site frequency spectrum (SFS) of a class of sites be the vector \( \nu_i \) containing \( i \) \( (0 \leq i < n) \) segregating alleles in a sample of \( n \) alleles from the population. Then, \( \pi \) and \( \theta_w \) are calculated as follows (Watterson 1975; Tajima 1983):

\[
\pi = 2 \sum_{i=1}^{n-1} \frac{i(n - i)\nu_i}{n(n - 1)}, \quad \theta_w = \frac{\sum_{i=1}^{n-1} \nu_i}{\sum_{i=1}^{n-1} \frac{1}{i}}.
\]

For our data set, given that we sequenced both chromosomes of each of the 15 M. m. castaneus individuals, the sampled number of alleles (\( n \)) was 30 if the sequencing was successful for every individual. However, due to sequencing failures, our data set did not contain 30 sequenced alleles for each site, so we calculated composite estimates of \( \pi \) and \( \theta_w \). We calculated \( \pi \) and \( \theta_w \) for sites that had the same number of alleles sequenced (categories of coverage) and then summed the estimates across categories of coverage as per Halligan et al. (2010).

For a population at Fisher–Wright equilibrium, and assuming no selection, \( \pi \) and \( \theta_w \) estimates are expected to be equal to one another. They are expected to differ, however, if there is a skew in the SFS toward low- or high-frequency alleles. The level of skew can be quantified by the Tajima’s \( D \) statistic (Tajima 1989). However, to calculate \( D \), there needs to be an equal number of alleles sequenced at each site. We therefore rejected any sites where we had fewer than 20 alleles sequenced. We then sampled without replacement 20 alleles from each of the remaining sites such that the number of alleles sampled at each site was constant.

Nucleotide divergence (\( d \)) between M. m. castaneus and both M. famulus and rat was calculated using the Kimura 2-parameter correction (Kimura 1980). We had multiple sequences for M. m. castaneus, so we computed an average
divergence. We calculated evolutionary constraint \( C_d \) by comparing substitution rates at a putatively neutral and a selected site class. We use substitution rates at neutral sites to estimate expected numbers of substitutions at selected sites. Expected parameter numbers of substitutions are compared, and constraint is calculated as \( C_d = 1 - O_d/E_d \) (Eyre-Walker and Keightley 1999; Keightley and Gaffney 2003). We used synonymous sites as the neutral class.

**Nonsynonymous and Synonymous Sites**

We treated nonsynonymous and synonymous sites as in Li (1993) and Pamilo and Bianchi (1993): All zero-fold degenerate sites were treated as nonsynonymous and 4-fold degenerate sites as synonymous. At 2-fold degenerate sites, transitions were considered synonymous and transversions nonsynonymous. Unmutated 2-fold degenerate sites were divided into nonsynonymous and synonymous by considering the ratio of transitional and transversional changes \((t/t_v)\) as calculated at 4-fold degenerate sites across all genes in a comparison of \( M. m. castaneus \) and \( M. famulus \).

**DFE of New Mutations**

We employed a maximum likelihood (ML) approach described by Keightley and Eyre-Walker (2007) to infer the DFE of new mutations at nonsynonymous sites of coding regions and in upstream and downstream regions as implemented in the program DFE-alpha (available online at http://homepages.ed.ac.uk/eang33/). The method assumes two classes of sites, one neutral and one selected, and contrasts SFSs of the two classes. Fitness effects of new mutations \((s)\) are assumed to be zero in the neutral class and unconditionally deleterious in the selected class and are sampled from a gamma distribution with parameters \(a\) (scale) and \(b\) (shape). The method also incorporates a simple demographic model: The population at an initial size \(N_1\) experiences a step change to \(N_2\) for \(t\) generations in the past.

We use a constant \(N_2\) of 100 so that the ratio \(N_2/N_1\), that is, the change in population size, is actually estimated. An additional parameter, \(f_0\), estimates the proportion of unmutated sites. The parameter space of \(N_2/N_1\), \(t/N_2\), \(f_0\), \(a\), and \(b\) is searched to find the values that maximize the likelihood of observing the neutral and selected SFSs. In order to account for variation in the number of alleles at each site, we generated SFSs for sites that had the same number of alleles sampled in both neutral and selected classes. We summed the log-likelihoods of each SFS to produce the overall log-likelihood as per Halligan et al. (2010). We also explored the likelihood surface for the demographic and selection parameters to check for any irregularities such as local maxima (see supplementary figs. S1 and S2, Supplementary Material online). The likelihood surfaces were plotted using R (Ihaka and Gentleman 1996). We interpolated from the estimated parameters of the gamma distribution, the percentages of mutations that fall within four \(N_s\) ranges: 0–1, 1–10, 10–100, and >100.

**Estimating Evolutionary Constraint by Using Polymorphism Data**

Evolutionary constraint, calculated using divergences \((C_d)\), will be biased downward if some fraction of the observed divergence at the focal class is adaptive. We obtained a second estimate of evolutionary constraint, which is not subject to such biases, by using information from polymorphism data only \((C_p)\). We first estimate the average fixation probability of a new, negatively selected mutation \((\bar{u})\) at the focal class by integrating over the DFE as in Eyre-Walker and Keightley (2009):

\[
\bar{u} = \int_0^\infty 2Nu(N, s)f(s|a, b) \, ds,
\]

where \(u(N, s)\) is the fixation probability of a new negatively selected mutation \((N\) is assumed equal to \(N_a\)).

\(C_p\) can then be calculated as follows:

\[
C_p = 1 - \bar{u}.
\]

**Adaptive Evolution**

To estimate the proportion of adaptive substitutions \((z)\), approaches based on the MK test are frequently used (Eyre-Walker 2006). However, these approaches do not take into account slightly deleterious mutations, which contribute proportionally more to polymorphism than divergence and therefore can lead to underestimates of \(z\). They also ignore demographic history, which can be problematic, because a population size change in the past could produce evolutionary signatures similar to selection. A recent extension of the MK test (Eyre-Walker and Keightley 2009) attempts to take into account both slightly deleterious mutations and population demography.

The method assumes a neutral and a selected class of sites. The DFE in the selected class is first estimated, simultaneously accounting for demography as in Keightley and Eyre-Walker (2007). The nucleotide divergence of the neutral class \((d_N)\) is assumed to be proportional to the mutation rate and divergence due to deleterious mutations in the selected class is the product of the mutation rate and the average fixation probability of a new deleterious mutation \((\bar{u})\). We can estimate the expected divergence \((d_{est})\) in the selected class due to neutral and deleterious mutations as follows:

\[
d_{est} = d_N \bar{u}.
\]

The difference between the observed \((d_X)\) and estimated divergence \((d_{est})\) estimates the amount of adaptive divergence \((d_{adaptive} = d_X - d_{est})\) in the selected class \((X)\). If we scale \(d_{adaptive}\) by \(d_X\) we obtain \(z\), the fraction of adaptive substitutions in the selected class:

\[
z = \frac{d_{adaptive}}{d_X}.
\]

However, as noted by Gossmann et al. (2010), caution should be exercised when comparing estimates of \(z\) from
Table 1. Details of Genes Sequenced and Percentages of Sites Sequenced for All 30 Alleles and for at Least 20 Alleles.

<table>
<thead>
<tr>
<th>Site Class</th>
<th>Number of Genes</th>
<th>Number of Sites</th>
<th>Mean Number Sites Per Gene (SD)</th>
<th>30 Alleles</th>
<th>&gt;20 Alleles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsynonymous</td>
<td>78</td>
<td>34,532</td>
<td>443 (160)</td>
<td>60</td>
<td>95</td>
</tr>
<tr>
<td>Synonymous</td>
<td>78</td>
<td>13,056</td>
<td>167 (63)</td>
<td>60</td>
<td>94</td>
</tr>
<tr>
<td>Intron</td>
<td>65</td>
<td>43,672</td>
<td>672 (413)</td>
<td>45</td>
<td>88</td>
</tr>
<tr>
<td>Upstream</td>
<td>49</td>
<td>25,303</td>
<td>516 (132)</td>
<td>50</td>
<td>93</td>
</tr>
<tr>
<td>Downstream</td>
<td>51</td>
<td>26,622</td>
<td>522 (182)</td>
<td>57</td>
<td>91</td>
</tr>
</tbody>
</table>

different species. Differences in the estimates of \( \alpha \) could reflect differences in the contribution of slightly deleterious mutations to \( d_s \) rather than different rates of adaptive substitution. We can control for differences in the frequency of effectively neutral mutations in the selected class by computing \( \omega_{sr} \), the ratio of \( d_{\text{adaptive}} \) to \( d_s \):

\[
\omega_{sr} = \frac{d_{\text{adaptive}}}{d_s}
\]  

We also estimated \( \alpha \) using a simple but frequently used method (Fay and Wu 2001):

\[
\alpha_{\text{FWW}} = 1 - \frac{D_s P_X}{D_X P_s},
\]  

where \( D_X \) and \( D_s \) are counts of divergent sites between \( M. \text{m castaneus} \) and an outgroup species for selected and neutral site classes, respectively, and \( P_X \) and \( P_s \) counts of polymorphic sites for selected and neutral site classes, respectively.

Confidence intervals and standard error for all parameters were obtained by bootstrapping 1,000 times by gene. \( P \) values, computed for comparisons between site classes or with zero, were obtained by two-tailed bootstrap tests.

Results

Data
Our data set consists of sequences from 78 autosomal genes from a sample of 15 wild, unrelated \( M. \text{m castaneus} \) individuals sampled from NW India. Part of the coding region of these genes and partial introns were sequenced in a previous study (Halligan et al. 2010). In this study, we focus on regions directly upstream and downstream of the coding region of these genes. We successfully amplified and sequenced \( \sim 500 \) bp upstream and downstream from the start and stop codons for a subset of 49 and 51 of the 78 genes, respectively (table 1). We also sequenced the orthologous genes in a \( M. \text{famulus} \) individual, which we used together with the rat as an outgroup to estimate divergence, constraint, \( \alpha \) and \( \omega_{sr} \).

Summary Statistics
Nucleotide diversity, Tajima’s \( D \), divergence to \( M. \text{famulus} \) and rat and evolutionary constraint estimates are shown in table 2. The upstream and downstream site classes show intermediate levels of nucleotide diversity (\( \pi = 0.56 \) % in both cases) and divergence to \( M. \text{famulus} \) (\( d = 2.49 \% \) for upstream and 2.38% for downstream) compared with nonsynonymous sites (\( \pi = 0.15 \% \) and \( d = 0.82 \% \)) but are closer to the synonymous site estimates (\( \pi = 0.75 \% \) and \( d = 3.27 \% \)). Divergence to the rat is about five times higher than divergence to \( M. \text{famulus} \) for all site classes.

In contrast to nonsynonymous sites, the upstream, downstream, and intronic site classes do not show discernible differences in the shape of their SFSs compared with synonymous sites (fig. 1). Tajima’s \( D \) estimates, which quantify the skew in the SFS, are significantly lower than zero for all cases examined even for synonymous sites (\( P < 10^{-3} \) in all cases examined). A negative Tajima’s \( D \) indicates an excess of rare variants, which can be caused by negative selection. However, population expansion, a prolonged bottleneck or population subdivision can also produce a similar pattern. Different mutation rates between the compared regions could also alter the SFS. For example, CpG dinucleotides have higher mutation rates in mammals (Arndt et al. 2003) and their frequencies differ between coding and noncoding DNA. However, if CpG-prone sites are excluded, we observe little changes in the SFS in all cases (fig. 1).

We calculate evolutionary constraint \( C_d \) by comparing inter specific divergence between the focal site class and a putatively neutral site class. The estimate for \( C_d \) is moderately high for upstream and downstream site classes (24.1% and 28.4%, respectively, table 2) and significantly different than zero in both cases (\( P = 0.004 \) for upstream; \( P = 0.002 \) for downstream).

DFE of New Mutations
We inferred the DFE of new mutations along with demographic parameters using a ML-based approach (Keightley and Eyre-Walker 2007). We first tested whether a model that incorporates demographic change plus selection (Model 3) fits the data significantly better than a model that assumes only demographic change (Model 1). The likelihood ratios for this comparison were highly significant in all cases examined (\(-\Delta\log L\) values are reported in table 3; \( P < 10^{-2} \) in all cases, with d.f. = 2). We also examined a model where we fitted only selection under constant population size (Model 2). We found that the fit of Model 2 to the data was significantly poorer than Model 3 in all cases examined (\(-\Delta\log L\) values are reported in table 3; \( P < 10^{-9} \) in all cases, with d.f. = 2).
Using the inferred parameters of the DFE (see supplementary tables S3 and S4, Supplementary Material online, for estimates of the demographic parameters, the mean selective effect ($N_e s$), and the shape parameter of the fitted gamma distribution), we can estimate the proportion of mutations falling into each of four categories of selective effects ($N_e s$): 0–1, 1–10, 10–100, and >100 (table 4). For upstream and downstream sites, most new mutations fall into the effectively neutral category ($N_e s$, 0–1) (69.8% and 67.5%, respectively), which is in sharp contrast and significantly different ($P < 0.05$ for both upstream and downstream in all comparisons) from the estimate for nonsynonymous sites (15.4%). Although most new mutations in upstream and downstream sequences are effectively neutral, there is a substantial fraction (21.7% and 19.6%, respectively, not significantly different from zero) of strongly selected mutations ($N_e s > 10$) in these regions.

We also calculate evolutionary constraint $C_p$, a statistic that summarizes the DFE, and is the average probability of a new deleterious mutation to be lost. $C_p$ is moderately high for upstream (30.7%) and downstream (33.9%) site classes and is significantly different from zero only for the downstream site class ($P < 0.05$).

**Adaptive Evolution**

We then estimated the fraction of substitutions driven to fixation by positive selection ($\alpha$) using an extension of the MK test (Eyre-Walker and Keightley 2009). This method uses neutral divergence between *M. m. castaneus* and an outgroup (either *M. famulus* or rat) along with the DFE, inferred from polymorphism data of *M. m. castaneus* to infer the expected divergence between *M. m. castaneus* and the outgroup. The difference between the observed and the expected divergence estimates the adaptive divergence between *M. m. castaneus* and the outgroup. $\alpha$ is then calculated by scaling the adaptive divergence by the observed divergence (see detailed description in the Materials and Methods). Estimates of $\alpha$ for the nonsynonymous, upstream, and downstream site classes for *M. m. castaneus* are presented in table 5. We report moderately low estimates of $\alpha$ for upstream and downstream site classes (11.8% and 9.3%, respectively), which are not significantly different from zero when using *M. famulus* as the outgroup and synonymous sites as the neutral reference. By combining upstream and downstream sequences, we obtain a similar point estimate for $\alpha$ (9.3%) and a narrower confidence interval (9.2–22.3), which includes zero but excludes high estimates of $\alpha$ and is not significantly different from the estimate for nonsynonymous sites ($P = 0.078$). The point estimates are very similar when using the rat as the outgroup or intronic sites are used as the neutral reference, but the confidence intervals are narrower for the latter case because more data are included. The estimate for $\alpha$ for combined upstream and downstream sequences is significantly different from nonsynonymous sites when using intronic sites as the neutral reference ($P = 0.014$ when

**Table 2.** Estimates of Percentage Diversity ($\% \pi_1$, $\% \theta_{uw}$) Summed over All Sites for *M. m. castaneus*, Tajima’s $D$, Percentage Divergence ($\% d$) to *Mus famulus*, and the Rat and Evolutionary Constraint ($C_p$) Calculated using Synonymous Sites as the Neutral Class.

<table>
<thead>
<tr>
<th>Site Class</th>
<th>$% \pi$ (SE)</th>
<th>$% \theta_{uw}$ (SE)</th>
<th>Tajima’s $D$ (SE)</th>
<th>$% d$ (<em>M. famulus</em>) (SE)</th>
<th>$% d$ (rat) (SE)</th>
<th>$C_p$ (<em>M. famulus</em>) (SE)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nonsynonymous</td>
<td>0.15 (0.02)</td>
<td>0.22 (0.02)</td>
<td>-0.93 (0.17)</td>
<td>0.82 (0.10)</td>
<td>3.69 (0.40)</td>
<td>74.5 (3.3)</td>
</tr>
<tr>
<td>Synonymous</td>
<td>0.75 (0.06)</td>
<td>0.93 (0.06)</td>
<td>-0.53 (0.12)</td>
<td>3.27 (0.21)</td>
<td>18.11 (0.56)</td>
<td>—</td>
</tr>
<tr>
<td>Intron</td>
<td>0.66 (0.04)</td>
<td>0.83 (0.04)</td>
<td>-0.75 (0.09)</td>
<td>2.90 (0.14)</td>
<td>15.61 (0.42)</td>
<td>12.1 (7.2)</td>
</tr>
<tr>
<td>Upstream</td>
<td>0.56 (0.05)</td>
<td>0.71 (0.06)</td>
<td>-0.78 (0.14)</td>
<td>2.49 (0.21)</td>
<td>12.20 (0.65)</td>
<td>24.1 (7.1)</td>
</tr>
<tr>
<td>Downstream</td>
<td>0.56 (0.06)</td>
<td>0.69 (0.06)</td>
<td>-0.59 (0.16)</td>
<td>2.38 (0.21)</td>
<td>11.78 (0.78)</td>
<td>28.4 (7.7)</td>
</tr>
</tbody>
</table>

**Fig. 1** Plots of the SFSs for nonsynonymous, upstream, downstream, intron, and synonymous site classes for all sites and for non-CpG-prone sites only. Numbers of polymorphic sites are given in parentheses and refer to the “All Sites” plot.
Table 3. Likelihood-Ratio Tests Contrasting Models Fitted to the SFS Data when Estimating the DFE. Models Fitting Only Demography (M1) or Only Selection (M2) Are Contrasted with a Model That Fits Both Demography and Selection (M3) to the Data.

<table>
<thead>
<tr>
<th>Site Class</th>
<th>Selected</th>
<th>M1 versus M3</th>
<th>M2 versus M3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>312.4</td>
<td>47.2</td>
<td></td>
</tr>
<tr>
<td>Upstream</td>
<td>12.4</td>
<td>50.5</td>
<td>43.7</td>
</tr>
<tr>
<td>Downstream</td>
<td>13.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Synonymous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>412.6</td>
<td>120.3</td>
<td></td>
</tr>
<tr>
<td>Upstream</td>
<td>10.6</td>
<td>116.8</td>
<td></td>
</tr>
<tr>
<td>Downstream</td>
<td>12.1</td>
<td>126.0</td>
<td></td>
</tr>
</tbody>
</table>

Note: M1, demography, no selection; M2, no demography, selection; M3, demography and selection.

using *M. famulus* as the outgroup and $P = 0.018$ when using the rat as the outgroup.

The different estimates of $\alpha$ between nonsynonymous, upstream, and downstream site classes might be due to differences in the rate of slightly deleterious rather than adaptive substitution (Gossmann et al. 2010). In order to account for any differences in the slightly deleterious substitution rate between the selected site classes, we computed the ratio of adaptive divergence to neutral divergence ($\omega_a$). The resulting $\omega_a$ estimates are very similar and not significantly different for the nonsynonymous, upstream, and downstream site classes (11.9%, 9%, 6.6%, respectively, table 5). These results indicate that the lower estimates of $\alpha$ in the upstream and downstream site classes compared with nonsynonymous sites could be due to a higher proportion of the upstream and downstream sites evolving close to neutrally rather than a lower absolute rate of adaptive substitution. The confidence intervals for $\omega_a$ are very wide when examining upstream and downstream site classes individually but narrow down when we combine data from upstream and downstream site classes (the upper boundary for $\omega_a$ is never higher than 30.6%). Similarly, with estimates for $\alpha$, when using intronic sites as the neutral reference and the rat as the outgroup, we get narrower confidence intervals because more data are included.

Finally, we use a simple frequently used approach to estimate $\alpha$ (Fay et al. 2001) in order to be able to make comparisons with studies that have not employed the Eyre-Walker and Keightley (2009) methodology. We control for slightly deleterious mutations by excluding low-frequency polymorphisms (<10%) as suggested by (Fay et al. 2001). By using this method, we obtained zero or negative estimates of $\alpha$ for upstream and downstream site classes (supplementary table S5, Supplementary Material online) that roughly agree with the estimates we obtained using the Eyre-Walker and Keightley (2009) methodology.

Discussion

In this study, we present results suggesting that sites upstream and downstream of protein-coding regions in *M. castaneus* are, on average, under weak positive and negative selection. Several lines of evidence support this conclusion. Nucleotide diversity in *M. castaneus* and divergence to *M. famulus* or rat in upstream and downstream regions are much higher than for nonsynonymous sites and slightly but significantly lower than synonymous sites. Evolutionary constraint is also significantly lower in upstream and downstream regions than for nonsynonymous sites. Tajima’s D estimates are not significantly different between site classes, except for the synonymous and nonsynonymous sites comparison, which suggests either that all site classes investigated are under negative selection or that a population expansion or bottleneck has occurred in the past in *M. castaneus*. Indeed, if we fit a simple demographic model of a step change in population size, we find evidence for population expansion of *M. castaneus*, which might explain the negative Tajima’s D at synonymous sites. A population expansion might also explain the negative Tajima’s D in upstream and downstream regions. However, a model of a demographic change plus negative selection fits the data significantly better than a model of demographic change with no selection or a model with selection only in all cases examined. Therefore, we obtained statistically significant evidence for both a population expansion in *M. castaneus* in the past and negative selection acting on upstream and downstream regions. The DFE inferred for upstream and downstream regions implies that most new mutations have $N_s$ values in range of 0–1, but a small fraction is strongly negatively selected. At nonsynonymous sites the pattern is reversed because most new mutations are strongly deleterious. This result further supports the conclusion that upstream and downstream regions are, on average, under weak selective constraint compared with nonsynonymous sites.

Table 4. Estimates of Percentages of Mutations in Four $N_s$ Ranges and Evolutionary Constraint Estimated from Polymorphism ($C_P$).

<table>
<thead>
<tr>
<th>Site Class</th>
<th>Selected</th>
<th>Percentage of Mutations in $N_s$ Range [95% CI]</th>
<th>$C_P$ [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neutral</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nonsynonymous</td>
<td>15.4 [9/23.3]</td>
<td>11.4 [3.1/18.6]</td>
<td>19.5 [3.6/42.1]</td>
</tr>
<tr>
<td>Upstream</td>
<td>66.9 [48.1/100]</td>
<td>8.5 [0/39.0]</td>
<td>9.4 [0/20.3]</td>
</tr>
<tr>
<td>Downstream</td>
<td>67.5 [29.1/100]</td>
<td>13.0 [0/69.2]</td>
<td>13.6 [0/23.6]</td>
</tr>
<tr>
<td>Synonymous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upstream + Downstream</td>
<td>70.7 [52.6/92.2]</td>
<td>8.6 [5.9/31.8]</td>
<td>9.5 [0/17.7]</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Site Class</th>
<th>Selected</th>
<th>Percentage of Mutations in $N_s$ Range [95% CI]</th>
<th>$C_P$ [95% CI]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upstream</td>
<td>77.6 [56.6/100]</td>
<td>9.3 [0/35.2]</td>
<td>9.2 [0/15.7]</td>
</tr>
<tr>
<td>Downstream</td>
<td>75.9 [50.8/100]</td>
<td>9.2 [0/30.2]</td>
<td>9.5 [0/19.7]</td>
</tr>
<tr>
<td>Intron</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Upstream + Downstream</td>
<td>77.1 [63.9/100]</td>
<td>9.3 [0/21.3]</td>
<td>9.3 [0/14.5]</td>
</tr>
</tbody>
</table>
Our low point estimates of α for upstream and downstream regions are suggestive of weak positive selection operating in these regions compared with nonsynonymous sites. Although the confidence intervals for the estimates are very wide when we examine upstream and downstream regions individually, the confidence intervals for combined upstream and downstream regions exclude rates of adaptive substitution higher than ~36% on upstream and downstream regions of *M. m. castaneus*.

In sequencing diploid outbred individuals, regions between heterozygous indels can be problematical for SNP calling. Our data set contains <10% of such regions. If such regions are excluded, we obtain almost identical estimates for constraint, the parameters of the DFE and α (results not shown). Another consideration about our data set is that because the EGP sample is not random (Livingston et al. 2004), we might have excluded genes that have high rates of adaptive evolution in upstream and downstream regions. For example, promoter regions of many neural and nutrition-related genes in humans have been found to be subject to positive selection (Haygood et al. 2007). However, a comparison of estimates of α for regions ~500 bp upstream and downstream of the start and stop codon of protein-coding genes in humans, obtained with the methodology employed in the present study and using the EGP and PGA (Program for Genomic Applications; Akay et al. 2004) data sets, showed nonsignificant differences between data sets (Eyre-Walker and Keightley 2009). Additionally, a comparison of estimates of α, obtained with the methodology employed in the present study and using the EGP, PGA, and Boyko et al. (2008) data sets, has shown non significant differences between data sets for nonsynonymous sites in humans (Halligan et al. 2010).

It has been suggested that regulatory noncoding regions might be more important for evolution than protein-coding genes in primates (King and Wilson 1975). However, studies that have used a simple extension of the MK test (Keightley et al. 2005) and the methodology employed in this study (Eyre-Walker and Keightley 2009) have estimated that α in upstream and downstream regions in humans is close to zero. Humans might have low rates of adaptive substitution in upstream and downstream regions because of their historically low *N_e*. However, in the current study, we also obtain low estimates for α for upstream and downstream regions in *M. m. castaneus*, a mammalian species with a *N_e* much larger than humans (Halligan et al. 2010). The low estimate of α in upstream and downstream regions in *M. m. castaneus* may be due to the sparse distribution of regulatory elements in the mammalian genome. Therefore, the upstream and downstream sequences we have focused on could include a substantial amount of neutral sequence along with some functionally relevant elements.

In order to control for differences between site classes in the contribution of slightly deleterious mutations to the observed divergence, we calculated the ratio of adaptive to neutral divergence (ω_α), and we obtained similar estimates for nonsynonymous, upstream, and downstream site classes (~5–10%). Therefore, upstream and downstream regions of protein-coding genes in *M. m. castaneus* appear to have a similar absolute rate of adaptive substitution with nonsynonymous sites. This finding implies that the difference in the estimate of α observed at nonsynonymous sites between humans (~0–20%) and *M. m. castaneus* (~50%) might also be due to differences in the relative proportion of slightly deleterious mutations between the two species.

Finally, if noncoding regulatory elements are distributed over many thousands of base pairs in the mammalian genome, then the net input of adaptive substitutions to regulatory regions of mammals could be higher than protein-coding genes. Eyre-Walker and Keightley’s (2009) study in humans and our study in *M. m. castaneus* only examined ~500 bp upstream and downstream of the start and stop codon, respectively, of a limited collection of protein-coding genes. We suggest that genome-wide studies of putative regulatory noncoding regions are needed in *M. m. castaneus* and humans so that the role of regulatory regions to adaptation can be more confidently ascertained.

**Supplementary Material**

Supplementary figures S1 and S2 and supplementary tables S1–S5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

**Acknowledgments**

We thank Bettina Harr for providing the *M. m. castaneus* samples. We would also like to thank Bettina Harr, Adam Eyre-Walker, Paul Sharp, and three anonymous referees for providing insightful comments on the manuscript. We
acknowledge funding from grants from the Biotechnology and Biological Sciences Research Council (BBSRC) and the Wellcome Trust. A.K. is funded by a BBSRC postgraduate studentship.

References


