Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome

Article - Discoveries

Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome

Tom R. Booker* and Peter D. Keightley

Institute of Evolutionary Biology, University of Edinburgh, Edinburgh, EH9 3FL, United Kingdom

*Current address: *Department of Forest and Conservation Sciences, University of British Columbia, Vancouver, Canada*

Corresponding Author: Tom R. Booker

Email: *booker@zoology.ubc.ca*

Abstract

A major goal of population genetics has been to determine the extent by which selection at linked sites influences patterns of neutral nucleotide diversity in the genome. Multiple lines of evidence suggest that diversity is influenced by both positive and negative selection. For example, in many species there are troughs in diversity surrounding functional genomic elements, consistent with the action of either background selection (BGS) or selective sweeps. In this study, we investigated the causes of the diversity troughs that are observed in the wild house mouse genome. Using the unfolded site frequency spectrum (uSFS), we estimated the strength and frequencies of deleterious and advantageous mutations occurring in different functional elements in the genome. We then used these estimates to parameterize forward-in-time simulations of chromosomes, using realistic distributions of functional elements and recombination rate variation in order to determine if selection at linked sites can explain the observed patterns of nucleotide diversity. The simulations suggest that BGS alone cannot explain the dips in diversity around either exons or conserved non-coding elements (CNEs). A combination of BGS and selective sweeps produces deeper dips in diversity than BGS alone, but the inferred parameters of selection cannot fully explain the patterns observed in the genome. Our results provide evidence of sweeps shaping patterns of nucleotide diversity across the mouse genome, and also suggest that infrequent, strongly advantageous mutations play an important role in this. The limitations of using the uSFS for inferring the frequency and effects of advantageous mutations are discussed.

Introduction

Starting with the discovery of a positive correlation between nucleotide polymorphism and the recombination rate in *Drosophila* in the late 1980s and early 1990s (Aguade, et al. 1989; Begun and Aquadro 1992), it has become clear that natural selection affects genetic diversity across the genomes of many species (Cutter and Payseur 2013; Corbett-Detig, et al. 2015). More recently, models incorporating selection at sites linked to those under observation have been shown to explain a large amount of the variation in diversity across the genome (McVicker, et al. 2009; Charlesworth 2012; Comeron 2014; Elyashiv, et al. 2016). However, a persistent challenge has been to tease apart the contributions of positive and negative selection to the observed patterns.

Because the fates of linked alleles are non-independent, selection acting at one site may have consequences for variation and evolution at another. In broad terms, there are two models describing the effects of directional selection on neutral genetic diversity at linked sites, selective sweeps (SSWs)

and background selection (BGS). SSWs occur when positively selected alleles spread through a population, dragging with them the haplotype on which they arose (Maynard Smith and Haigh 1974; Barton 2000). There are a number of different types of SSW (reviewed in Booker, et al. (2017a)), but in the present study, when not made explicit, we use the term selective sweep to refer to the effects of a single *de novo* advantageous mutation being driven to fixation by selection. BGS, on the other hand, occurs because the removal of deleterious mutations results in a loss of genetic diversity at linked neutral sites (Charlesworth, et al. 1993; Charlesworth 2013). The magnitudes of the effects of SSWs and BGS depend on the strength of selection, the rate of recombination and the mutation rate (Hudson and Kaplan 1995; Nordborg, et al. 1996; Barton 2000). SSWs and BGS have qualitatively similar effects on genetic diversity, however, and many polymorphism summary statistics have little power to distinguish between them (Stephan 2010; Charlesworth 2013).

Several studies have attempted to differentiate between BGS and SSWs. For example, Sattath, et al. (2011) examined patterns of nucleotide diversity around recent nucleotide substitutions in *Drosophila simulans*. Averaging across the entire genome, they observed a trough in diversity around nonsynonymous substitutions, whereas diversity was relatively constant around synonymous ones. This difference is expected under a model of recurrent SSWs, but not under BGS. Their results provide evidence that SSWs have been frequent in *D. simulans* since the species shared a common ancestor with *Drosophila melanogaster* (the outgroup used in that study). Similar results have been reported for *Capsella grandiflora* (Williamson, et al. 2014). In humans (Hernandez, et al. 2011), house mice (Halligan, et al. 2013) and maize (Beissinger, et al. 2016), however, there is very little difference between the patterns of diversity around putatively neutral and potentially adaptive substitutions. These results have been interpreted as evidence that hard SSWs are infrequent in those species. However, Enard, et al. (2014) argued that the proportion of neutral amino acid substitutions in regions of the genome with low functional constraint (and thus weak BGS effects) will be higher than the proportion occurring in regions with high functional constraint (and thus stronger BGS effects), so the Sattath test will be difficult to interpret in species with genomes exhibiting highly variable levels of functional constraint, such as humans and mice (but see (Beissinger, et al. 2016)). Indeed, Enard, et al. (2014) found evidence that adaptive substitutions are fairly frequent in both protein-coding and non-coding portions of the human genome, suggesting that SSWs are common. Furthermore, Nam, et al. (2017) analysed reductions in nucleotide diversity in genomic regions close to genes in the great apes and concluded that strong SSWs, rather than BGS, are required to explain the observed patterns. In that study, although the authors examined a wide range of selection parameters, they did not attempt to identify parameters of selection specific to any ape species.

There are a number of methods for estimating the frequency and strength of advantageous mutations using models of selection at linked sites (Booker, et al. 2017a). Recently, Elyashiv, et al. (2016) produced a map of the expected nucleotide diversity in *D. melanogaster* by fitting a model incorporating both BGS and hard SSWs to the genome-wide patterns of genetic diversity and the divergence between *D. melanogaster* and *D. simulans*. They concluded that sweeps are required to explain much of the genome-wide variation in diversity. Their analysis conditioned the effects of sweeps on the locations of recent substitutions, which is reasonable in *D. melanogaster* where there is a reduction in mean diversity around nonsynonymous substitutions, but not around synonymous ones (Elyashiv, et al. 2016). As described above, this is not the case for wild mice. Indeed, even randomly selected synonymous and nonsynonymous sites in *Mus musculus*, regardless of whether they have experienced a recent substitution, exhibit almost identical reductions in diversity in surrounding regions (Halligan, et al. 2013). Conditioning a sweep model on the locations of recent substitutions in mice may, therefore, produce spurious parameter estimates. Furthermore, the selection parameters estimated by Elyashiv, et al. (2016) were inferred solely from variation in nucleotide diversity. There is information in the distribution of allele frequencies, the site frequency spectrum (SFS), that can be used to estimate the distribution of fitness effects (DFE) for both deleterious and advantageous mutations (Keightley and Eyre-Walker 2007; Boyko, et al. 2008; Schneider, et al. 2011; Tataru, et al. 2017). In the present study, we estimate the DFE using such methods, and then use our estimates to parameterise simulations modelling BGS and SSWs.

In this study, we attempt to understand the influence of natural selection on variation at linked sites in the house mouse, *Mus musculus*. Specifically, we analyse *M. m. castaneus,* a sub-species which has been estimated to have a long-term effective population size (N_e) of around 500,000 (Baines and Harr 2007; Halligan, et al. 2010), making it a powerful system in which to study molecular evolution in mammals. Both protein-coding genes and phylogenetically conserved non-coding elements (CNEs, which have roles in the regulation of gene expression (Lowe, et al. 2011)) exhibit signatures of natural selection in *M. m. castaneus* (Halligan, et al. 2013). In particular, Halligan, et al. (2013) showed that there are substantial reductions in diversity surrounding protein-coding exons and CNEs, consistent with selection reducing diversity at linked sites. The trough in diversity surrounding exons was found to be ~10x wider than the trough surrounding CNEs, suggesting that selection is typically stronger on protein sequences than regulatory sequences. These results, therefore, suggest that selection at linked sites affects nucleotide diversity across large portions of the genome. However, our understanding of the forces that have shaped patterns of diversity is incomplete.

We analyse data on wild-caught *M. m. castaneus* individuals to obtain estimates of the distribution of fitness effects (DFE) for several classes of functional elements in the mouse genome

and then use these to parameterise forward-in-time simulations. We analyse several aspects of our simulation data: 1) the patterns of genetic diversity and the distribution of allele frequencies around both protein-coding exons and conserved non-coding elements; 2) the rates of substitution in different functional elements; and 3) the patterns of diversity around nonsynonymous and synonymous substitutions.

Results

We investigated the causes of variation in genetic diversity around functional elements of house mice by analysing the genomes of 10 wild-caught individuals sequenced to high coverage (Halligan, et al. 2013). We compared nucleotide polymorphism and between-species divergence in three classes of functional sites (0-fold nonsynonymous sites, UTRs and CNEs) with polymorphism and divergence at linked, putatively neutral sequences (4-fold synonymous sites and CNE-flanks). The three classes of functional sites had lower levels of within-species polymorphism and between-species divergence than their neutral comparators (Table 1). This is expected if natural selection keeps deleterious alleles at low frequencies, preventing them from reaching fixation. Tajima's *D* is more negative for 0-fold sites, UTRs and CNEs than for their neutral comparators (Table 1), further indicating the action of purifying selection. It is notable that the two neutral site types exhibited negative Tajima's *D*, indicating that rare variants are more frequent than expected in a Wright-Fisher population (Table 1). This is consistent either with a recent population expansion or the widespread effects of selection on linked sites, both of which may be relevant for this population (Halligan, et al. 2013; Booker, et al. 2017b).

Inferring the unfolded site frequency spectrum

The distribution of derived allele frequencies in a class of sites (the unfolded site frequency spectrum - uSFS) potentially contains information on the frequency and strength of selected mutations. We estimated the uSFSs for 0-fold sites, UTRs and CNEs using a probabilistic method incorporating information from two outgroups (Keightley, et al. 2016). This method attempts to correct for biases inherent in parsimony methods.

A population's demographic history is expected to affect the shape of the SFS. DFE-alpha attempts to correct this by fitting a population size change model to the neutral site class, and, conditional on the estimated demographic parameters, estimates the DFE for linked, selected sites. In the case of 4-fold sites and CNE flanks, a 3-epoch model provided the best fit to the data, based on likelihood ratio tests (Table S1) The trajectories of the inferred population size changes were similar in each case, i.e. a population bottleneck followed by an expansion (Table S2). However, the magnitude

of the changes and the duration of each epoch differed somewhat (Table S2). A possible explanation is that the demographic parameter estimates are affected by selection at linked sites, which differs between site classes (Messer and Petrov 2013; Ewing and Jensen 2016; Schrider, et al. 2016).

To investigate if the inferred parameters of the demographic model are sensitive to the number of sampled individuals, we fitted the 3-epoch model to the 4-fold site data, down-sampled to either 5 or 8 individuals. We found that the magnitude of the population expansion inferred increased with sample size (Table S3). This is presumably caused by an excess of rare variants (for example singletons), as expected under population expansion. Increasing the number of sampled individuals will lead to an increasingly precise estimate of the proportion of rare variants in the population.

We found that the 4-fold site and CNE-flank uSFSs exhibited an excess of high frequency derived alleles relative to expectations under the best-fitting neutral demographic models (Figure S1). For example, χ^2-statistics for the difference between the observed and fitted number of sites for the last uSFS element (i.e. 19 derived alleles) were 245.9 and 505.6 for 4-fold sites and CNE-flanks, respectively. It is reasonable to assume that the differences between fitted and observed values are caused by processes that similarly affect the linked selected site class. We therefore corrected the 0-fold, UTR and CNE uSFSs by subtracting the proportional deviations between fitted and observed values for neutral site uSFSs prior to estimating selection parameters (see Supplementary Methods). Applying this correction (hereafter referred to as the demographic correction) appreciably reduced the proportion of high frequency derived variants (Figure 1).

Estimating the frequencies and strengths of deleterious and advantageous mutations

We estimated the DFE for harmful mutations (dDFE) and the rate and strength of advantageous mutations based on the uSFSs for the three different classes of functional sites using DFE-alpha under two different models (Table 2). The first, as described by Schneider, et al. (2011), used the full uSFS, including sites fixed for the derived allele (hereafter Model A). The second (hereafter Model B), incorporated an additional parameter that absorbs the contribution of sites fixed for the derived allele (see Supplementary Methods). This was motivated by the possibility that between-species divergence may be decoupled from within-species polymorphism (e.g. due to changing selection regimes), and this could lead to spurious estimates of selection parameters (Eyre-Walker and Keightley 2009; Tataru, et al. 2017). Since Model A is nested within Model B, the two can be compared using likelihood ratio tests. In the remainder of the study, results obtained under Model A are shown in parallel with results obtained under Model B.

We performed a comparison of different DFE models, including discrete distributions with one, two or three mutational effect classes and the gamma distribution including or not including advantageous mutations. For each class of functional sites, DFE models with several classes of deleterious mutational effects and a single class of advantageous effects gave the best fit (Table S4). For each class of functional sites, only a single class of advantageous mutations was supported, since additional classes of advantageous mutations did not significantly increase likelihoods (Table S5), presumably reflecting a lack of power. These best-fitting models were identified under both Model A or Model B. Parameter estimates pertaining to the dDFE were also similar between Models A and B (Table 2).

In this study, we estimated selection parameters based on the uSFS, whereas earlier studies on mice used the distribution of minor allele frequencies, i.e. the 'folded' SFS (Halligan, et al. 2010; Halligan, et al. 2011; Kousathanas, et al. 2011; Halligan, et al. 2013; Kousathanas, et al. 2014). A possible consequence of using the folded SFS is that advantageous mutations segregating at intermediate to high frequencies are allocated to the mildly deleterious class. In the case of 0-fold sites, for example, the best-fitting DFE did not include mutations with scaled effects in the range of $1 < |2N_e s| < 200$ (Table 2). This contrasts with previous studies using the folded SFS, which identified an appreciable proportion of mutations in the $1 < |2N_e s| < 200$ range (Halligan, et al. 2013; Kousathanas and Keightley 2013). This difference may influence the reductions in diversity caused by background selection, so we performed simulations incorporating either the gamma dDFEs inferred from analysis of the folded SFS by Halligan, et al. (2013) or the discrete dDFEs inferred in the present study.

For all classes of functional sites, we inferred that moderately positively selected mutations are fairly frequent under both Models A and B (Table 2). In the case of 0-fold sites, for example, the frequency of advantageous mutations was 0.3% (under Model A). Across the three classes of sites, the scaled selection strengths of advantageous mutations were fairly similar (Table 2), i.e. $2N_e s \sim 16$, implying that s is on the order of 10^{-5} (assuming N_e = 500,000; (Geraldes, et al. 2011)). However, we found that estimates of the frequency of advantageous mutations (p_a) obtained under Model B for 0-fold sites and UTRs were ~3 times higher than those obtained under Model A. In the cases of both 0-fold sites and UTRs, Model B fitted significantly better than Model A, as judged by likelihood ratio tests (0-fold sites, $\chi^2_{1\ \text{d.f.}} = 4.2$; $p = 0.04$; UTRs, $\chi^2_{1\ \text{d.f.}} = 9.9$; $p = 0.002$). Interestingly, in the case of CNEs, Models A and B did not differ significantly in fit ($\chi^2_{1\ \text{d.f.}} = 0.26$; $p = 0.60$) and estimates of the advantageous mutation parameters were similar (Table 2).

Forward-in-time population genetic simulations

We conducted forward-in-time simulations to examine whether estimates of the DFE obtained by analysis of the uSFS predict patterns of diversity observed around functional elements. In our simulations, we used estimates of selection parameters obtained by DFE-alpha for 0-fold sites, UTRs and CNEs assuming either Model A (i.e. from the full uSFS) or Model B (i.e. by absorbing the contribution of sites fixed for the derived allele with an additional parameter). The selection parameter estimates obtained under Models A and B resulted in major differences in the patterns of diversity around functional elements.

i) Patterns of nucleotide diversity around functional elements in simulated populations

Using the selection parameter estimates obtained from DFE-alpha (Table 2), we performed simulations incorporating deleterious mutations, advantageous mutations or both advantageous and deleterious. Our analysis involved computing diversity in windows surrounding functional elements and comparing the diversity patterns with those seen in *M. m. castaneus.* In order to aid visual comparisons, we divided nucleotide diversity (π) at all positions by the mean π at physical distances greater than 75Kbp and 4Kbp away from exons and CNEs, respectively. When comparisons were made on the scale of genetic distance, we divided π by its mean at distances greater than $4N_e r$ = 1,500 for protein-coding exons and $4N_e r$ = 200 for CNEs. These distances were chosen because they are the values beyond which π remains approximately constant.

In our simulations, we scaled recombination, mutation and selection parameters by N in a linear fashion. However, linear scaling can become problematic when selection coefficients are strong (Uricchio and Hernandez 2014). To test whether linear scaling was appropriate for the parameters we estimated, we simulated populations with N = 100, 500, 750, 1,000 and 2,000. We found that patterns of genetic diversity converged in populations with N = 750, 1,000 and 2,000 (Figure S2). The following simulations results were obtained assuming N = 1,000.

Simulations incorporating only deleterious mutations predicted a chromosome-wide reduction in genetic diversity. Around exons and CNEs, diversity plateaued at ~94% of the neutral expectation (Figure S3-S4). Simulations involving only BGS did not fully predict the observed troughs in diversity around functional elements. Specifically, the predicted troughs in diversity around both protein-coding exons and CNEs, were not as wide nor as deep as those observed in the real data (Figures 2-3; S3-S4). Similar predictions were obtained for Models A or B (Figures 2-3; S3-S4) and for the gamma

dDFEs inferred by Halligan, et al. (2013) (Figure S5). Our simulations incorporating deleterious mutations suggest, then, that while BGS affects overall genetic diversity across much of the genome, positive selection presumably also makes a substantial contribution to the dips in diversity around functional elements.

In our simulations of exons and surrounding regions, recurrent SSWs produced troughs in diversity, but they were both narrower and shallower than those observed in the house mouse. However, the results are sensitive to the model used to estimate selection parameters (Figure 2; Table 3). Assuming the selection parameters estimated under Model A (i.e. analysing the full uSFS) we found that advantageous mutations produced a small dip in diversity around exons, which was both shallower and narrower than the one generated by deleterious mutations alone (Figure 2; Table 3). In contrast, the advantageous mutation parameters estimated under Model B (i.e. where sites fixed for the derived allele do not influence selection parameters) resulted in a marked trough in diversity around exons in simulations (Figure 2; Table 3). In simulations incorporating both advantageous and deleterious mutations, the troughs in diversity around exons were not as large as those observed in *M. m. castaneus* (Figure 2; Table 3), and even at very large distances from exons, diversity was around 90% of neutral expectation (Figure S3). Assuming Model B selection parameters resulted in a trough in diversity that was both deeper and wider than the one generated when assuming Model A parameters (Figure 2). The differences between Model A simulations and Model B simulations presumably arise because under Model B the frequency of advantageous nonsynonymous mutations was ~3 times higher than under Model A (Table 2). When analysis windows are binned based on genetic distance rather than physical distance, the differences between observed and simulated diversity patterns are even more striking (Figure 3 and Figure S4).

We also carried out simulations focussing on CNEs and found that the combined effects of BGS and recurrent SSWs, as generated by our estimates of selection parameters, resulted in diversity troughs that were slightly shallower than observed (Figure 2; Table 3). Selection parameters obtained under Models A and B produced similar results. The troughs in diversity around CNEs in simulations incorporating only advantageous mutations were slightly shallower than the ones generated by deleterious mutations alone (Figure S4). The troughs in diversity around CNEs in our simulations were also slightly shallower than those observed (Figure 2). This could be because we failed to detect infrequent, strongly selected advantageous mutations in CNEs or we underestimated the true frequency of advantageous mutations occurring in those elements. When plotted on the scale of genetic distance, the differences between our simulated data and the *M. m. castaneus* data become strikingly apparent (Figure 3).

ii) The site frequency spectrum around functional elements

SSWs and BGS are known to affect the shape of the SFS for linked neutral sites (Braverman, et al. 1995; Charlesworth, et al. 1995; Kim 2006). SSWs and BGS generate troughs in diversity at linked sites (Figures 2-3), but nucleotide diversity on its own does not contain information about the shape of the SFS. Tajima's *D* is a useful statistic for this purpose, because it is reduced when there is an excess of rare polymorphisms relative to the neutral expectation and increased when intermediate frequency variants are more common (Tajima 1989). We therefore compared Tajima's *D* in the regions surrounding functional elements in simulations with values observed in the real data. It is notable that average Tajima's *D* is far lower in *M. m. castaneus* than in our simulations (Figure 4). This likely reflects a genome-wide process, such as population size change, that we have not modelled.

If we assume selection parameters obtained under Model A, Tajima's *D* around protein-coding exons is relatively invariant, and matches the pattern observed in the real data fairly well (Figure 4). However, under Model B, the simulations exhibit a marked dip in Tajima's *D*, which is not observed in the real data (Figure 4).

In the case of CNEs, we observed a trough in Tajima's *D* in the real data (Figure 4), and simulations predict similar troughs under Models A and B (Figure 4). However, the trough in Tajima's *D* may be caused by the presence of functionally constrained sequences in the immediate flanks of CNEs (*See Methods*), making a comparison between the simulations and the observed data problematic.

iii) Rates of substitution in functional elements

Incorporating information from sites fixed for the derived allele when estimating the DFE (as in Model A) or disregarding this information (as in Model B) had a striking effect on estimates of the frequency and effects of advantageous mutations (Table 2). In the case of 0-fold sites, for example, p_a was ~3x higher under Model B than Model A (Table 2). We therefore investigated the extent by which such differences affect the divergence at selected sites under the two models. Nucleotide divergence at putatively neutral sites between the mouse and the rat is approximately 15%, so we simulated an expected neutral divergence of 7.5% for one lineage.

We compared the ratio of nucleotide divergence at selected sites to the divergence at neutral sites (d_{sel}/d_{neut}) between the simulated and observed data. In simulations that assumed the selection

parameters obtained under Model A, d_{sel}/d_{neut} values were similar to those observed in *M. m. castaneus* for all classes of selected sites (Table 4). Under Model B, however, the simulations predicted substantially more substitutions at nonsynonymous sites and UTRs than were seen in the real data (Table 4). This suggests that, under Model B, the frequency of advantageous mutations for 0-fold sites and UTRs may be overestimated.

iv) Examining the uSFS and estimating the DFE from simulated data

BGS and SSWs both perturb allele frequencies at linked neutral sites, distorting site frequency spectra, which can lead to the inference of spurious demographic histories (Messer and Petrov 2013; Ewing and Jensen 2016; Schrider, et al. 2016). By fitting a model incorporating three epochs of population size to the putatively neutral site data, we inferred that *M. m. castaneus* has experienced a population bottleneck followed by an expansion (Table S2). To investigate the possibility that the inferred demographic histories could be an artefact of selection at linked sites, we fitted demographic models to the uSFS obtained from simulated synonymous sites. Visual comparison of the uSFS from simulated populations with the uSFS obtained from *M. m. castaneus* reveals that our simulations do not fully capture the excess of high frequency variants observed in the mouse population (Figure S6). However, for simulations assuming the selection parameters obtained under either Model A or B, a 3-epoch population size model gave the best fit to the data. The estimated demographic histories were somewhat different between simulations assuming Model A or Model B, but in each case a population bottleneck followed by an expansion was inferred (Table S6). This is an interesting observation: our simulations assumed a constant population size, but selection at linked sites appears to distort the neutral uSFS, such that a demographic history similar to the one inferred from the real data is estimated (Table S6).

Our simulations also indicate that selection parameters are difficult to accurately infer using the uSFS alone. In the case of Model A simulations, the selection strength and frequency of deleterious mutations was accurately estimated, as was the combined frequency of all effectively neutral mutations (Table S6). However, in Model A simulations, DFE-alpha did not accurately estimate the strength and frequency of advantageous mutations. Estimates of selection parameters in Model B simulations were similar to the input parameters, but a notable exception was that the frequency of advantageous mutations (p_a) was overestimated (Table S6). These results suggest that some features of the uSFS inferred for *M. m. castaneus* have been captured by our simulated data. However**,** the demographic correction which we applied to the uSFS before estimating selection parameters (see Supplementary Methods), had a substantially greater impact on the *M. m. castaneus* data than for the simulated data, particularly in the case of high frequency derived alleles (Figure S6).

A possible explanation is that strong SSWs, which cause a greater increase in the proportions of high frequency derived alleles for linked sites (Kim 2006), have left a signal in the neutral site uSFS, but these cannot be accurately inferred from the selected site uSFS itself. If this were the case, then there may be information in the uSFS for linked neutrally evolving sites that could be used when estimating selection parameters. This would require the expected uSFS arising under the joint effects of SSWs and BGS.

v) Patterns of diversity around sites that have recently experienced a substitution

In general, it has been difficult to discriminate between BGS and SSWs, because their effects on genetic diversity and the site frequency spectrum are qualitatively similar. It has been suggested that the two processes can be teased apart by taking advantage of the fact that hard SSWs should be centred on a nucleotide substitution, whereas this is not the case for BGS. Comparing the average genetic diversity in regions surrounding recent putatively selected and putatively neutral substitutions (e.g. 0-fold and 4-fold sites, respectively) may therefore reveal the action of SSWs (Hernandez, et al. 2011; Sattath, et al. 2011). Halligan, et al. (2013) performed such an analysis in *M. m. castaneus* using the closely related *M. famulus* as an outgroup, and found that the profiles of neutral diversity around 0-fold and 4-fold substitutions were virtually identical. Similar findings have been reported in other species (Hernandez, et al. 2011; Beissinger, et al. 2016). One interpretation of these results is that hard SSWs are rare. To investigate this, we measured the average neutral diversity around nonsynonymous and synonymous substitutions in simulations for the case of frequent hard SSWs.

In our simulations, we measured diversity around substitutions occurring on a time-scale equivalent to the divergence time between *M. m. castaneus* and *M. famulus*. The average diversities around nonsynonymous and synonymous substitutions in the simulated data were very similar, regardless of whether simulations assumed Model A or Model B selection parameters (Figure 5). However, the troughs in diversity around substitutions were deeper in the simulations assuming Model B (Figure 5), reflecting the higher frequency of advantageous mutations (Table 2). In the immediate vicinity of nonsynonymous substitutions, diversity was lower than the corresponding value for synonymous substitutions (Figure 5). However, the differences are slight, so it would be difficult to draw firm conclusions about the action of either SSWs or BGS. Taken together, these results suggest that analysing patterns of diversity around recent substitutions does not provide enough information to convincingly discriminate between SSWs and BGS in *M. m. castaneus*, even when hard sweeps are fairly frequent. Further analysis is required to assess whether this is also the case for other organisms.

Discussion

There are a number of observations suggesting that natural selection is pervasive in the murid genome. First, there is a positive correlation between synonymous site diversity and the rate of recombination (Booker, et al. 2017b). Secondly, there is reduced diversity on the X-chromosome compared to the autosomes, which cannot readily be explained by neutral or demographic processes (Baines and Harr 2007). Thirdly, there are troughs in genetic diversity surrounding functional elements, such as protein-coding exons and CNEs, which are consistent with the action of BGS and/or SSWs (Halligan, et al. 2013). In this paper, we analysed the sequences of 10 *M. m. castaneus* individuals sampled from the ancestral range of the species (Halligan, et al. 2013). We estimated the DFEs for several classes of functional sites (0-fold nonsynonymous sites, UTRs and CNEs), and used these estimates to parameterise forward-in-time simulations. We investigated whether the simulations predict the observed troughs in diversity around functional elements along with the between-species divergence observed between mice and rats.

Estimating selection parameters based on the uSFS

Relative to putatively neutral comparators, 0-fold sites, UTRs and CNEs all exhibited reduced nucleotide diversity, reduced nucleotide divergence and an excess of low frequency variants (Table 1; Figure 1), consistent with the action of natural selection (Halligan, et al. 2010; Halligan, et al. 2013). The estimates of the DFEs included substantial proportions of strongly deleterious mutations (Table 2). In addition, the best-fitting models also included a single class of advantageous mutations. Although additional classes were not statistically supported, in reality, there is almost certainly a distribution of advantageous selection coefficients (Bank, et al. 2014; McDonald, et al. 2016). A visual examination of the fitted and observed uSFSs, however, shows that the estimated DFEs fitted the data very well (Figure S7), suggesting that there is limited information in the uSFS to estimate a range of positive selection coefficients.

When estimating the DFE for a particular class of sites, we analysed either the full uSFS including sites fixed for the derived allele (Model A) or we ignored sites fixed for the derived allele (i.e. Model B). Recently, Tataru, et al. (2017) showed that selection parameters can be accurately estimated from the uSFS in simulations that ignored between-species divergence, if the freuqnecy of advantageous mutations (p_a) is sufficiently high. In our analysis of 0-fold sites and UTRs, Model B gave a significantly better fit and higher estimates p_a than Model A (Table 2). For CNEs, however, Models A and B did not significantly differ in fit, and the selection parameter estimates were very similar (Table 2). The goodness-of-fit and parameter estimates obtained under Models A and B may

differ if the processes that generated between species-divergence are decoupled from the processes that produce within species diversity. There are several factors that could potentially cause such a decoupling. 1) Past demographic processes may have distorted the uSFS in ways not captured by the corrections we applied; 2) there may be error in assigning alleles as ancestral or derived; 3) the nature of the DFE may have changed since the accumulation of between-species differences began; and 4) there could be rare, strongly advantageous mutations that contribute to divergence, but contribute negligibly to polymorphism. It is difficult to know which of these factors affected the outcome of our analyses. However, we found that Model B gave a better fit to the uSFS than Model A for 0-fold sites and UTRs, but not CNEs. There is an *a priori* expectation that the strength of selection on protein-coding sequences is greater than regulatory sequences (Halligan, et al. 2013), so we think the latter factor is likely to have been important.

Patterns of diversity and Tajima's D around functional elements

We performed simulations incorporating our estimates of deleterious and advantageous mutation parameters to dissect the contribution of BGS and selective sweeps to diversity dips around functional elements. Our simulations suggest that BGS and SSWs both produce genome-wide reductions in neutral diversity (Figures S3-4), but neither process on its own fully explains the troughs in diversity around protein-coding exons and CNEs, regardless of which model (A or B) is used to estimate selection parameters (Figures 2-3). Around protein-coding exons, the combined effects of advantageous and deleterious mutations generated a shallower trough in diversity than the one observed (Figure 2). These patterns are qualitatively similar when analysing physical or genetic distances, but the differences between observed and simulated patterns are more apparent when analysing genetic distances (Figure 3). A possible explanation for this is that rare, strongly selected advantageous mutations are undetectable by analyses based on the uSFS (discussed below). In contrast, the combined effects of BGS and SSWs predicted troughs in diversity surrounding CNEs that closely match those observed, when measured on the physical scale (Figure 2). Troughs of diversity around CNEs on the scale of genetic distance are not nearly so similar in simulated populations to those observed in *M. m. castaneus* as they are on the physical scale. Specifically, the observed trough in diversity around CNEs on the genetic distance scale is both deeper and wider than in simulated populations. Analysing patterns of diversity on the physical scale is analogous to assuming that there is a uniform recombination rate across the genome. Our results highlight the importance of incorporating recombination rate variation when performing such analyses, particularly in species that exhibit highly variable recombination rates.

There is an excess of rare variants in *M. m. castaneus* relative to the neutral expectation, as indicated by a strongly negative Tajima's *Ds* for putatively neutral sites (Table 1) and for regions surrounding exons and CNEs (Figure 4). Our simulations incorporating both advantageous and deleterious mutations also exhibited negative Tajima's *D*, but not nearly so negative as in the real data (Figure 4). This difference between the observed data and the simulations indicates that there may be processes generating an excess of rare variants, such as a recent population expansion, which were not incorporated in the simulations.

Rates of nucleotide substitutions in simulations

Our simulations suggest that the frequency of advantageous mutations (p_a) estimated for 0-fold sites and UTRs under Model B may be unrealistically high. This is because several aspects of the results were incompatible with the observed data. Firstly, we found that the substitution rates for simulated nonsynonymous and UTR sites were higher than those observed between mouse and rat (Table 4). Secondly, we observed a pronounced dip in Tajima's *D* around simulated exons, which is not present in the real data (Figure 4), suggesting that under Model B, either the strength or frequency of positive selection at 0-fold sites is overestimated.

Do our results provide evidence for strongly selected advantageous mutations?

Estimation of the strength and frequency of advantageous mutations based on the uSFS relies on the presence of positively selected variants segregating within the population (Boyko, et al. 2008; Schneider, et al. 2011; Tataru, et al. 2017). The frequency of advantageous mutations may impose a limit on the parameters of positive selection that can be accurately estimated. Indeed, Tataru, et al. (2017) recently showed that p_a may be overestimated when analysing the uSFS, if the true value of p_a is low.

If advantageous mutations are infrequent, those with larger effects on fitness will be less likely to observed segregating than those with milder effects, as strongly selected mutations have shorter sojourn times than weakly selected ones (Fisher 1930; Kimura and Ohta 1969). This could explain why the selection parameters we estimated fail to predict the troughs in diversity observed in the real data (Figure 2). Furthermore, the fact that Model B gave a better fit than Model A for 0-fold sites and UTRs suggests that polymorphism and divergence have become decoupled for those sites. This is also consistent with the presence of infrequent, strongly selected mutations that become fixed rapidly and are thus not commonly observed as polymorphisms.

Relevant to this point, an interesting comparison can be made between two recent studies to estimate the frequency and strength of positive selection using the same *D. melanogaster* dataset. The first, by Keightley, et al. (2016), used the uSFS analysis methods of Schneider, et al. (2011) (i.e. Model A in the present study), and estimated the frequency of advantageous mutations (p_a) = 4.5 x 10^{-3} and the scaled strength of selection ($2N_e s_a$) = 23.0 for 0-fold nonsynonymous sites. In the second study, Campos, et al. (2017) estimated p_a = 2.2 x 10^{-4} and $2N_e s_a$ = 241, based on the correlation between synonymous site diversity and nonsynonymous site divergence. Although the individual parameter estimates differ substantially, the compound parameter $2N_e s_a p_a$ (which approximates the rate of SSWs) was similar between the studies (0.055 and 0.104 for Campos, et al. (2017) and Keightley, et al. (2016) respectively). It is expected that synonymous site diversity is reduced by SSWs, so the method used by Campos, et al. (2017) may be sensitive to the presence of strongly selected mutations, whereas the Keightley, et al. (2016) approach may have been more sensitive to weakly selected ones. It seems plausible then, that the two studies capture different aspects of the DFE for advantageous mutations (a similar argument was made by Sella, et al. (2009)). Supporting this view, Elyashiv, et al. (2016) recently estimated the DFE in *D. melanogaster*, incorporating both strongly and weakly selected advantageous mutations, by fitting a model incorporating BGS and SSWs to genome-wide variation in genetic diversity. They inferred that weakly selected mutations are far more frequent than strongly selected ones, but that both contribute to variation in genetic diversity across the *D. melanogaster* genome. In the present study, we used similar methods as Keightley, et al. (2016) to estimate the frequency and strength of advantageous mutations, so the estimated parameters of positive selection may represent only weakly selected mutations. Indeed, patterns of diversity at microsatellite loci suggest that there are strongly selected, infrequent sweeps in multiple European *M. musculus* populations (Teschke, et al. 2008), so infrequent strong sweeps may be a general feature of mouse evolution.

We tested the hypothesis that undetected, strongly selected mutations are chiefly responsible for the reductions in diversity around functional elements by performing additional simulations. In this exercise, we assumed far stronger selection on advantageous mutations for protein-coding regions than we estimated for 0-fold sites (Table S7). When modelling strong selection, we reduced p_a such that rate of sweeps (proportional to the product $2N_e s_a p_a$) was either that estimated from the uSFS under Model A, or double that value (Table S7). As can be seen in Figure 6, increasing the strength of selection acting on advantageous mutations, while simultaneously decreasing the p_a parameter resulted in troughs in diversity around protein-coding exons that were both deeper and wider than those observed in *M. m. castaneus*. By chance, we identified a set of parameters ($2N_e s$ = 400; p_a = 0002) that can provide a relatively close correspondence between simulated and observed patterns of

diversity. However, it must be stressed that these parameters were chosen arbitrarily and that there is no statistical support for them.

Although strongly selected mutations can generate a similar trough in average diversity around protein-coding exons as observed in the real data, they do not produce the apparent genome-wide reduction in Tajima's *D* observed in *M. m. castaneus* (Figure S8). Indeed, in all simulations modelling strongly advantageous mutations, we observed a trough in Tajima's *D* that plateaued at values close to 0 in regions surrounding protein-coding exons (Figure S8). In this exercise, we manipulated the selection parameters for 0-fold sites only, but it is possible that there are strongly advantageous mutations in all of the site classes. A combination of DFE parameters for the different functional elements in mice could therefore explain the reductions in diversity and the genome-wide negative Tajima's *D.* On the other hand, it could also be that recent demographic processes have swamped the signal of linked positive selection in the site frequency spectrum. Our results highlight the need for methods that can simultaneously estimate selection parameters for multiple functional elements and demographic history.

Understanding the contributions of regulatory and protein change to phenotypic evolution has been an enduring goal in evolutionary biology (King and Wilson 1975; Carroll 2005; Franchini and Pollard 2017). If selection is strong relative to drift (i.e. $2N_e s_a > 1$) then the rate of change of fitness from the fixation of advantageous mutations is expected to be proportional to the square of the selection coefficient (Falconer and Mackay 1996). In this study, we inferred that the strength of selection acting on new advantageous mutations in CNEs and 0-fold sites are roughly equivalent, but that advantageous mutations occur more frequently in CNEs (Table 2). Given that there are more CNE nucleotides in the genome than there are 0-fold sites (Table 1), this could imply that adaptation at regulatory sites causes the greatest fitness change in mice. However, our analyses suggest that both protein-coding genes and CNEs may experience strongly selected advantageous mutations, which are undetectable by analysis of the uSFS. If this were the case, protein-coding mutations could make a larger contribution to fitness change than mutations in regulatory sites.

Limitations of the study

There is a growing body of evidence suggesting that hard sweeps may not be the primary mode of adaptation in both *D. melanogaster* and humans. Firstly, soft sweeps, where multiple haplotypes reach fixation due to the presence of multiple *de novo* mutations or selection acted on standing variation, may be common. Garud, et al. (2015) developed a suite of haplotype-based statistics that can discriminate between soft and hard SSWs. The application of these statistics to

North American and Zambian populations of *D. melanogaster* suggested that soft sweeps are the dominant mode of adaptation in that species, at least in recent evolutionary time (Garud, et al. 2015; Garud and Petrov 2016). Furthermore, Schrider and Kern (2017) recently reported that signatures of soft sweeps are more frequent than those of hard sweeps in humans. However, their method did not explicitly include the effects of partial sweeps and/or BGS. Under a model of stabilising selection acting on a polygenic trait, if the environment changes, adaptation to a new optimum may cause small shifts in allele frequency at numerous loci without necessarily resulting in fixations (Barton and Keightley 2002; Pritchard, et al. 2010). Genome-wide association study hits in humans exhibit evidence that such partial SSWs may be common (Field, et al. 2016). These results all suggest that the landscape of adaptation may be more complex than the model of directional selection acting on a *de novo* mutation assumed in this study. For example, our simulations did not incorporate changing environments or stabilising selection, so we were unable to model adaptive scenarios other than hard sweeps.

Further work should aim to understand the probabilities of the different types of sweeps. Different functional elements have different DFEs for harmful mutations. In particular, regulatory elements seem to experience more mildly selected deleterious mutations than coding sequences (Halligan, et al. 2013; Williamson, et al. 2014) (Table 2). It has been argued that such differences in constraint between coding and non-coding elements may be due to a lower pleiotropic burden on regulatory sequences (Carroll 2005). Differences in the DFE among different genomic elements is expected to affect genetic diversity within these elements. This, in turn, may affect the types of sweeps that occur, since the relative probabilities of a hard versus soft sweep depend on the level of standing genetic variation (reviewed in (Hermisson and Pennings 2017)).

In our simulations, we treated N_e as constant through time, but this is an oversimplification. We analysed two different classes of putatively neutral sites, and inferred there has been a population size bottleneck followed by an expansion (Table S2). In our simulations, however, we showed that the inferred demographic history may largely be an artefact of selection at linked sites (Table S6). There is a strongly negative Tajima's *D* in genomic regions far from functional elements, which is not explained by selection (or at least the selection parameters we inferred) (Figure 4). This reduction is presumably caused by a demographic history or strong selection that was not included in our simulations. Better estimates of the demographic history of *M. m. castaneus* may be obtained, for example, from regions of the genome experiencing high recombination rates, located far from functional elements. Finally, the size of mouse populations may oscillate seasonally (Pennycuik, et al. 1986) and if this were the case, so would the effective selection strength of new mutations (and thus the probabilities of SSWs) (Otto and Whitlock 1997).

In house mice, crossing over events predominantly occur in narrow windows of the genome termed recombination hotspots (Brick, et al. 2012). The locations of recombination hotspots have evolved rapidly between and within *M. musculus* sub-species (Smagulova, et al. 2016), but at broad-scales recombination rates are relatively conserved (Booker, et al. 2017b). Assuming a single suite of recombination hotspots in simulations may produce misleading results if hotspot locations evolve faster than the rate of neutral coalescence. While we included fine-scale variation in recombination rates in our simulations, we used a recombination map was inferred at a broader scale than the scale of hotspots (Booker, et al. 2017b). However, hotspots are an important feature of the recombination landscape in mice and thus potentially influence the patterns of diversity around functional elements, but the appropriate way to model them is unclear.

Conclusions

Using simulations, we have shown that estimates of the DFE obtained by analysis of the uSFS cannot fully explain the patterns of diversity around both CNEs and protein-coding exons. We argue that, while frequent mutations with moderate advantageous effects occur in different functional elements in the mouse genome (Table 2), strongly advantageous mutations that are undetectable by analysis of the uSFS generate the bulk of the reductions in diversity. Estimates of the strength and rate of advantageous mutations could be obtained by directly fitting a sweep model to the troughs in diversity around functional elements. We have shown that BGS makes a substantial contribution to these troughs, and applying models that incorporate both BGS and sweeps (Kim and Stephan 2000; Elyashiv, et al. 2016; Campos, et al. 2017) might allow us to make more robust estimates of selection parameters.

Materials and Methods

Samples and polymorphism data

We analysed the genome sequences of 10 wild-caught *M. m. castaneus* individuals sequenced by Halligan *et al.* (Halligan, et al. 2013). The individuals were sampled from an area that is thought to include the ancestral range of the species (Baines and Harr 2007). A population structure analysis suggested that the individuals chosen for sequencing came from a single randomly mating population (Halligan, et al. 2010). Sampled individuals were sequenced to an average depth of ~30x using Illumina technology. Reads were mapped to version mm9 of the mouse genome and variants called as described in Halligan, et al. (2013). Only single nucleotide polymorphisms were considered, and insertion/deletion polymorphisms were excluded from downstream analyses. We used the

genome sequences of *Mus famulus* and *Rattus norvegicus* as outgroups in this study. For *M. famulus*, a single individual was sequenced to high coverage and mapped to the mm9 genome (Halligan, et al. 2013). For *R. norvegicus*, we used the whole genome alignment of the mouse (mm9) and rat (rn4) reference genomes from UCSC.

For the DFE-alpha analysis (*see below*), the underlying model assumes a single, constant mutation rate. Hypermutable CpG sites strongly violate this assumption, so CpG-prone sites were excluded as a conservative way to remove CpG sites from our analyses. A site was labelled as CpG-prone if it is preceded by a C or followed by a G in the 5' to 3' direction in either *M. m. castaneus*, *M. famulus* or *R. norvegicus*. Additionally, sites that failed a Hardy-Weinberg equilibrium test ($p < 0.002$) were excluded from further analysis, because they may represent alignment errors.

Functional elements in the murid genome

In this study, we considered three different classes of functional elements in the genome: the exons and untranslated regions (UTRs) of protein-coding genes and conserved non-coding elements (CNEs).

Coordinates for canonical splice-forms of protein-coding gene orthologs between *Mus musculus* and *Rattus norvegicus* were obtained from version 67 of the Ensembl database. We used these to identify untranslated regions (UTRs) as well as 4-fold and 0-fold degenerate sites in the coding regions. We made no distinction between 3' and 5' UTRs in the analysis. Genes containing alignment gaps affecting >80% of sites in either outgroup and genes containing overlapping reading frames were excluded. This left a total of 18,171 autosomal protein-coding genes.

The locations of conserved non-coding elements (CNEs) in the house mouse genome were identified as described by Halligan, et al. (2013).

Estimating the parameters of the distribution of fitness effects (DFE) for a particular class of sites using DFE-alpha (*see below*) requires neutrally evolving sequences for comparison. When analysing 0-fold degenerate sites and UTRs, we used 4-fold degenerate sites as the comparator. For CNEs, we used non-conserved sequence in the flanks of CNEs. Halligan *et al.* (Halligan, et al. 2013) found that, compared to the genome-wide average, nucleotide divergence between mouse and rat in the ~500bp on either side of CNEs is ~20% lower than that of intergenic DNA distant from CNEs, suggesting functional constraint in these regions. For the purpose of obtaining a quasi-neutrally evolving reference class of sequence and to avoid these potentially functional sequences, we

therefore used sequence flanking the edges of each CNE, offset by 500bps. For each CNE, the total amount of flanking sequence used in the analysis was equal to the length of the focal CNE, split evenly between the upstream and downstream regions. CNE-flanking sequences overlapping with another annotated feature (i.e. exon, UTR or CNE) or the flanking sequence of another CNE were excluded.

The site frequency spectrum around functional elements

For distances of up to 100Kbp on either side of exons and 5Kbp on either side of CNEs, the non-CpG-prone sites in non-overlapping windows of 1Kbp and 100bp, respectively, were extracted. For each analysis window, we calculated the genetic distance to the focal element in terms of the population-scaled recombination rate ($\rho = 4N_e r$) using the *M. m. castaneus* recombination map we constructed in an earlier study (Booker, et al. 2017b). Sites within analysis windows that overlapped with any of the annotated features described above, or that contained missing data in *M. m. castaneus* or either outgroup were excluded. The data for analysis windows were collated based on either physical or genetic distances distance from the nearest CNE or exon, from which we calculated nucleotide diversity and Tajima's *D*.

Overview of DFE-alpha analysis

The distribution of allele frequencies in a sample, referred to as the site frequency spectrum (SFS), provides information on evolutionary processes. Under neutrality the SFS reflects past demographic processes, such as population expansions and bottlenecks, and potentially the effects of selection at linked sites. The allele frequency distribution will also be distorted if focal sites are subject to functional constraints. The SFS therefore contains information on the strengths and frequencies of mutations with different selective effects, known as the distribution of fitness effects (hereafter the DFE). Note that balancing selection may maintain alleles at intermediate frequencies (Charlesworth 2006), but we assume that the contribution of this form of selection to overall genomic diversity is negligible.

DFE-alpha estimates selection parameters using information contained in the SFS by a two-step procedure (Keightley and Eyre-Walker 2007). First, a demographic model is fitted to data for a class of putatively neutral sites. Conditional on the demographic parameter estimates, the DFE is then estimated for the selected sites. In the absence of knowledge of ancestral or derived alleles, the 'folded' SFS can be used to estimate the demographic model and the DFE for harmful variants (hereafter referred to as the dDFE) (Keightley and Eyre-Walker 2007). If information from one or more

outgroup species is available, and the ancestral state for a segregating site can be inferred, one can construct the 'unfolded' SFS (uSFS). In the presence of positive selection, such that advantageous alleles segregate at an appreciable frequency, the parameters of the distribution of fitness effects for advantageous mutations can be estimated from the uSFS (Schneider, et al. 2011; Keightley, et al. 2016; Tataru, et al. 2017). In this study, we estimate the proportion of new mutations occurring at a site that are advantageous (p_a) and the strength of selection acting on them ($N_e s_a$).

Inference of the uSFS and the DFE

We inferred the distributions of derived allele frequencies in our sample for 0-fold and 4-fold sites, UTRs, CNEs and CNE-flanks using *M. famulus* and *R. norvegicus* as outgroups, using the two-outgroup method implemented in ml-est-sfs v1.1 (Keightley, et al. 2016). This method employs a two-step procedure conceived to address the biases inherent in parsimony methods. The first step estimates the rate parameters for the tree under the Jukes-Cantor model by maximum likelihood assuming a single mutation rate. Conditional on the rate parameters, the individual elements of the uSFS are then estimated.

DFE-alpha fits discrete population size models, allowing up to two changes in population size through time. For each class of putatively neutral sites, one-, two- and three-epoch models were fitted by maximum likelihood and the models with the best fit (as judged by likelihood ratio tests) were used in further analyses. When fitting the three-epoch model, we ran DFE-alpha (v2.16) 10 times with a range of different search algorithm starting values, in order to check convergence. We explored the effect of fitting the demographic model to a smaller number of individuals by down-sampling the 4-fold sites dataset to 5 and 8 individuals, with respect to frequency.

In the cases of 4-fold sites and CNE-flanks, the inferred uSFSs exhibited a higher proportion of high frequency derived alleles than expected under the best-fitting demographic model (Figure S1) (hereafter referred to as an uptick). Such an increase is not possible under the single population, single locus demographic models assumed. There are several possible explanations for the uptick: 1) mis-inference of the uSFS due to an inadequacy of the model assumed in ml-est-sfs; 2) failure to capture the demographic history of *M. m. castaneus* by the models implemented in DFE-alpha; 3) sequencing errors in *M. m. castaneus* or either outgroup generating spurious signals of divergence; 4) SSWs, since they can drag linked alleles to high frequencies (Braverman, et al. 1995; Kim 2006); 5) cryptic population sub-division in our sample of mouse individuals; and 6) positive selection, acting on the putatively neutral sites themselves. We think this latter explanation is unlikely, however, since there is little evidence for selection on synonymous codon usage in *Mus musculus* (dos Reis and

Wernisch 2009). With the exception of direct selection affecting the putatively neutral class of sites, the above sources of bias should also affect the selected class of sites (Eyre-Walker, et al. 2006; Glemin, et al. 2015; Keightley, et al. 2016). We therefore corrected the selected sites uSFS prior to inferring selection parameters by subtracting the proportional deviation between the neutral uSFS expected under the best-fitting demographic model and the observed neutral uSFS (following Keightley *et al.* (Keightley, et al. 2016); see Supplementary Methods).

Simultaneous inference of the DFE for harmful mutations (dDFE) and adaptive mutation parameters was performed using DFE-alpha (v.2.16) (Schneider, et al. 2011). A gamma distribution has previously been used to model the dDFE, since it can take a variety of shapes and has only two parameters (Eyre-Walker and Keightley 2007). However, more parameter-rich discrete point mass distributions provide a better fit to nonsynonymous polymorphism site data in wild house mice (Kousathanas and Keightley 2013). We therefore compared the fit of one, two and three discrete class dDFEs and the gamma distribution, and also included one or more classes of advantageous mutations. Nested DFE models were compared using likelihood ratio tests, and non-nested models were compared using Akaike's Information Criteria (AIC). Goodness of fit was also assessed by comparing observed and expected uSFSs using the χ^2-statistic, but the numbers of sites in the i^{th} and $n\text{-}i^{th}$ classes are non-independent, so formal hypothesis tests were not performed.

We constructed profile likelihoods to obtain confidence intervals. Two unit reductions in *logL*, on either side of the maximum likelihood estimates (MLEs) were taken as approximate 95% confidence limits.

Two methods for inferring the rates and effects of advantageous mutations based on the uSFS

It has been suggested that estimates of the DFE obtained based on the uSFS may be biased if sites fixed for the derived allele are included in calculations (Tataru, et al. 2017). Sites fixed for the derived allele are typically a frequent class in the uSFS, and therefore strongly influence parameter estimates. Bias can arise, for example, if the selection strength has changed since the split with the outgroup, such that the number of sites fixed for the derived allele do not reflect the selection regime that generated current levels of polymorphism. If nucleotide divergence and polymorphism are decoupled in this way, selection parameter estimated from only polymorphism data (and sites fixed for ancestral alleles) may therefore be less biased than those obtained when using the full uSFS. To investigate this possibility, we estimated selection parameters either utilising the full uSFS (we refer to this method as Model A) or by analysing the uSFS while fitting an additional parameter

(Supplementary Methods), such that sites fixed for the derived allele do not contribute to estimates of the selection parameters (we refer to this method as Model B).

Certain alleles present in a sample of individuals drawn from a population may appear to be fixed that are, in fact, polymorphic. Attributing such polymorphisms to between-species divergence may then influence estimates of the DFE by increasing the number of sites fixed for the derived allele (note that this would only affect estimates obtained under Model A). We corrected the effect of polymorphism attributed to divergence using an iterative approach as follows. When fitting selection or demographic models, DFE-alpha produces a vector of expected allele frequencies. Using this vector, we inferred the expected proportion of polymorphic sites that appear to be fixed for the derived allele. This proportion was then subtracted from the fixed derived class and distributed among the polymorphism bins according to the allele frequency vector. We then refitted the model using this corrected uSFS, and this procedure was applied iteratively until convergence (See Supplementary Methods). For each site class, convergence was achieved within five iterations and the selection parameters for each class did not substantially change between iterations.

Forward-in-time simulations modelling background selection and selective sweeps

We performed forward-in-time simulations in SLiM v1.8 (Messer 2013) to assess whether the observed patterns of diversity around functional elements (Halligan, et al. 2013) can be explained by SSWs or BGS caused by mutations originating in the elements themselves. These simulations focussed on either protein-coding exons or CNEs. We also ran SLiM simulations to model the accumulation of between-species divergence under our estimates of the DFE. In all our simulations, we either assumed the estimates of selection parameters obtained from the full uSFS (Model A) or those obtained when sites fixed for the derived allele do not contribute to parameter estimates (Model B).

Models of BGS and recurrent SSWs predict that the magnitudes of their effects are sensitive to the rate of recombination and mutation rate and the strength of selection (Wiehe and Stephan 1993; Nordborg, et al. 1996; Coop and Ralph 2012). To parameterise our simulations, we used estimates of compound parameters scaled by N_e. For example, estimates of selection parameters obtained from DFE-alpha are expressed in terms of $N_e s$ (where s is the difference in fitness between homozygotes for ancestral and derived alleles, assuming semi-dominance). For a population where N_e = 1,000 and s = 0.05, for example, the strength of selection is therefore approximately equivalent to that of a population where N_e = 10,000 and s = 0.005. By scaling parameter values according to the population size of the simulations (N_{sim}), we modelled the much larger *M. m. castaneus* population (N_e

≅ 500,000 (Geraldes, et al. 2011) in a computationally tractable way. However, this linear scaling can be problematic, particularly for strong positive selection (Uricchio and Hernandez 2014). We compared patterns of diversity in simulations with population sizes of N = 100, 500, 750, 1,000 and 2,000 diploid individuals to assess the effect of the linear scaling given the selection, recombination and mutation parameters we assumed.

1. Annotating simulated chromosomes

Functional elements are non-randomly distributed across the house mouse genome. For example, protein-coding exons are clustered into genes and CNEs are often found close to other CNEs (Halligan, et al. 2013). Incorporating this distribution into simulations is important when modelling BGS and recurrent SSWs, because their effects on neutral diversity depend on the density of functional sequence (Nordborg, et al. 1996; Campos, et al. 2017). We incorporated the distribution as follows. For each simulation replicate, we chose a random position on an autosome, which was itself randomly selected (with respect to length). The coordinates of the functional elements (exons, UTRs and CNEs) in the 500Kbp downstream of that position were used to annotate a simulated chromosome of the same length. For simulations focussing on exons (CNEs), we only used chromosomal regions that had at least one exon (CNE).

2. Mutation, recombination and selection in simulations

We used an estimate of the population scaled mutation rate, $\theta=4N_e\mu$, to set the mutation rate (μ) in simulations, such that levels of neutral polymorphism approximately matched those of *M. m. castaneus*. Diversity at putatively neutral sites located close to functional elements (for example, 4-fold synonymous sites) may be affected by BGS and SSWs. To correct for this, we used an estimate of θ = 0.0083, based on the average nucleotide diversity at non-CpG-prone sites at distances >75Kbp from protein-coding exons. This distance was used, because it the approximate distance beyond which nucleotide diversity remains flat. The mutation rate in simulations was thus set to $0.0083/4N_{sim}$.

Variations in the effectiveness of selection at linked sites, due to variation in the rate of recombination across the genome, may not be captured by simulations that assume a single rate of crossing over. Recently, we generated a map of variation in the rate of crossing-over for *M. m. castaneus* using a coalescent approach (Booker, et al. 2017b), quantified in terms of the population scaled recombination rate $\rho=4N_er$. Recombination rate variation in the 500Kbp region used to obtain functional annotation was used to specify the genetic map for individual simulations.

We modelled natural selection at sites within protein-coding exons, UTRs and CNEs in the simulations using the estimates of selection parameters obtained from the DFE-alpha analysis. In the case of protein-coding exons, 25% of sites were set to evolve neutrally (i.e. synonymous sites), and the fitness effects of the remaining 75% were drawn from the DFE inferred for 0-fold sites (hereafter termed nonsynonymous sites in the simulations). For mutations in UTRs and CNEs, 100% were drawn from the DFEs inferred for those elements. Population scaled selection coefficients were divided by $2N_{sim}$ to obtain values of s for use in simulations. All selected mutations were assigned a dominance coefficient of 0.5, as assumed by DFE-alpha.

3. Patterns of diversity around functional elements in simulations

We examined the contributions of BGS and recurrent SSWs to the troughs in diversity observed around protein-coding exons and CNEs using forward-in-time simulations. Focussing on either protein-coding exons or CNEs, we performed three sets of simulations. The first incorporated only harmful mutations (causing BGS), the second only advantageous mutations (causing SSWs), and the third set incorporated both (causing both processes). Thus, under a given set of DFE estimates, we performed six sets of simulations (three sets focussing on exons and three sets focussing on CNEs). For each simulation set, 2,000 SLiM runs were performed, each using a randomly sampled 500Kbp region of the genome. In each SLiM run, populations of N_{sim}= 1,000 diploid individuals were allowed to evolve for 10,000 generations ($10N_{sim}$) in order to approach mutation-selection-drift balance. At this point, 200 randomly chosen haploid chromosomes were sampled from the population and used to construct SFSs.

For each set of simulations, segregating sites in windows surrounding functional elements were analysed in the same way as for the *M. m. castaneus* data (*see above*). The SFSs for all windows at the same distance from an element were collated. Analysis windows around protein-coding exons were oriented with respect to the strand orientation of the actual gene. Neutral sites near the tips of simulated chromosomes only experience selection at linked sites from one direction, so analysis windows located within 60Kbp of either end of a simulated chromosome were discarded. For a given distance to a functional element, we obtained confidence intervals around individual statistics by bootstrapping analysis window 1,000 times.

Mutation rate variation is expected to contribute to variation in nucleotide diversity. Nucleotide divergence between mouse and rat is relatively constant in the intergenic regions surrounding protein-coding exons (Halligan, et al. 2013), suggesting that mutation rate variation is not responsible for the

troughs in diversity around exons. Around CNEs, however, there is a pronounced dip in nucleotide divergence between *M. m. castaneus* and the rat. A likely explanation for this is that alignment-based approaches to identify CNEs fail to identify the edges of some elements, resulting in the inclusion of functionally constrained sequence in the analysis windows close to CNEs. This factor was not incorporated in our simulations, so in order to correct for this constraint, allowing us to compare diversity around CNEs in *M. m. castaneus* with our simulation data, we scaled values as follows. We divided nucleotide diversity by between-species divergence, in this case mouse-rat divergence, giving a statistic (π/d_{rat}) that reflects diversity corrected for mutation rate variation. We then multiplied the π/d_{rat} values by the mean mouse-rat divergence in regions further than 3Kbp from the edges of CNEs to obtain values on the same scale as our simulation data.

When comparing the patterns of diversity around functional elements in our simulations with the observations from *M. m. castaneus*, we used the root mean square (RMS) as a measure of goodness-of-fit.

$$RMS = \sqrt{\frac{1}{n_w}\sum_{i=1}^{n_w}(\pi_{sim}(i) - \pi_{obs}(i))^2},$$

where $\pi_{sim}(i)$ and $\pi_{obs}(i)$ are the diversity values from simulations and *M. m. castaneus*, respectively, in window *i* around a particular class of functional element and n_w is the total number of analysis windows. Approximate confidence intervals for RMS values were obtained using the bootstrap replicates described above.

4. Re-inferring the DFE based on simulated population data

We performed two additional sets of simulations to model the accumulation of between-species nucleotide divergence under the DFE estimates obtained by analysis of the full uSFS (i.e. Model A) and those obtained when sites fixed for the derived allele did not contribute to selection parameters (i.e. Model B). These simulations were the same as those described above, except that we ran them for additional generations to approximate the mouse-rat divergence. We ran 4,000 replicates of these simulations. Using polymorphic sites and sites fixed for the derived allele, we constructed the uSFS for each class of functional sites.

In order to model the mouse-rat divergence, we required a time frame to approximate the neutral divergence between those two species. Neutral divergence between *M. m. castaneus* and *R. norvegicus* (K_{rat}) is ~15% at non-CpG-prone sites far from protein-coding exons. Under neutrality, divergence is expected to be equal to $2T\mu$, where T is the time in generations since the two-species

shared a common ancestor and μ is the mutation rate per base pair per generation. In the simulations, the mutation rate was 2.075×10^{-6} bp^{-1} (recall that we scaled mutation rates using an estimate of $4N_e\mu$) and since $K_{rat} = 0.15$, $T = 36{,}145$ generations. We thus ran simulations incorporating both deleterious and advantageous mutations, focussing on exons, for 46,145 generations, discarding the first 10,000 as burn-in. At the final generation, we constructed the uSFS for synonymous and nonsynonymous sites from 20 randomly sampled haploid chromosomes. To obtain a proxy for mouse-rat divergence, we counted all substitutions that occurred after the $10N_{sim}$ burn-in phase plus any derived alleles present in all 20 haploid chromosomes.

Using the uSFSs for synonymous and nonsynonymous sites obtained from the simulations, we estimated selection parameters using the methods described above. We first fitted one-, two- and three- epoch demographic models to simulated synonymous site data. For the simulations assuming Model A or Model B, we found that the three-epoch demographic model gave the best fit to the simulated synonymous site uSFS in both cases. Using the expected uSFS under the three-epoch model, we performed the demographic correction (Supplementary Methods) before estimating selection parameters. When estimating selection parameters based on simulation data, we used the same methods as used for the analysis of the *M. m. castaneus* data, i.e. the DFE for Model A simulations was estimated using Model A *etc.*

5. Patterns of diversity around recent nonsynonymous and synonymous substitutions

Comparisons of the average level of nucleotide diversity around recent synonymous and nonsynonymous substitutions have been used to test for positive selection (Hernandez, et al. 2011; Sattath, et al. 2011; Halligan, et al. 2013; Williamson, et al. 2014; Beissinger, et al. 2016). In *M. m. castaneus* there is essentially no difference in diversity around recent substitutions at 0-fold and 4-fold sites (Halligan, et al. 2013). This could reflect a paucity of SSWs, or alternatively, this particular test may be unable to discriminate between BGS and SSWs in mice. Using our simulation data, in which SSWs are relatively frequent, we tested whether patterns of diversity around selected and neutral substitutions reveals the action of positive selection. In their study, Halligan *et al.* (Halligan, et al. 2013) used *M. famulus* as an outgroup to locate recent substitutions, because it is much more closely related to *M. musculus* than the rat. We obtained the locations of nucleotide substitutions in our simulations as follows. Neutral divergence between *M. m. castaneus* and *M. famulus* (K_{fam}) is 3.4%. In the simulations, given that the mutation rate was 2.075×10^{-6}, 8,193 generations are sufficient to approximate the *M. m. castaneus* lineage since its split with *M. famulus* K_{fam}. Thus, all substitutions that occurred in 8,193 generations were analysed. Neutral diversity around synonymous and nonsynonymous substitutions in non-overlapping windows of 1,000bp up to 100Kbp from substituted

sites were then extracted from the simulations. Sites in analysis windows that overlapped with functional elements were excluded. If two substitutions of the same type were located less than 100Kbp apart, analysis windows extended only to the midpoint of the two sites.

Except where noted, all analyses were conducted using custom Python and Perl scripts. Scripts used in this study are deposited on GitHub (https://github.com/TBooker)

Acknowledgements

We owe thanks to Brian Charlesworth, Ben Jackson and two anonymous reviewers for useful comments on the manuscript and to Deborah Charlesworth, Dan Halligan, Rory Craig and the evolutionary genetics lab group at the University of Edinburgh for helpful discussions throughout the course of the project. Tom Booker is supported by a BBSRC EASTBIO Studentship. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 694212).

References

Aguade M, Miyashita N, Langley CH. 1989. Reduced variation in the yello-achaete-schute region in natural populations of *Drosophila melanogaster*. Genetics 122:607-615.
Baines JF, Harr B. 2007. Reduced X-Linked Diversity in Derived Populations of House Mice. Genetics 175:1911-1921.
Bank C, Hietpas RT, Wong A, Bolon DN, Jensen JD. 2014. A bayesian MCMC approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. Genetics 196:841-852.
Barton NH. 2000. Genetic hitchhiking. Philos Trans R Soc Lond B Biol Sci 355:1553-1562.
Barton NH, Keightley PD. 2002. Understanding quantitative genetic variation. Nat Rev Genet 3:11-21.
Begun DJ, Aquadro CF. 1992. Levels of naturally occuring DNA polymorphism correlate with recombination rate in *Drosophila melanogaster*. Nature 356.
Beissinger TM, Wang L, Crosby K, Durvasula A, Hufford MB, Ross-Ibarra J. 2016. Recent demography drives changes in linked selection across the maize genome. Nature Plants 2:16084.
Booker TR, Jackson BC, Keightley PD. 2017a. Detecting positive selection in the genome. BMC Biol 15:98.
Booker TR, Ness RW, Keightley PD. 2017b. The Recombination Landscape in Wild House Mice Inferred Using Population Genomic Data. Genetics 207:297-309.
Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS Genet 4:e1000083.
Braverman JM, Hudson RR, Kaplan NL, Langley CH, Stephan W. 1995. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. Genetics 140:783-796.
Brick K, Smagulova F, Khil P, Camerini-Otero RD, Petukhova GV. 2012. Genetic recombination is directed away from functional genomic elements in mice. Nature 485:642-645.
Campos JL, Zhao L, Charlesworth B. 2017. Estimating the parameters of background selection and selective sweeps in *Drosophila* in the presence of gene conversion. Proc Natl Acad Sci Early Online.

Carroll SB. 2005. Evolution at two levels: on genes and form. PLoS Biol 3:e245.
Charlesworth B. 2013. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. J Hered 104:161-171.
Charlesworth B. 2012. The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the *Drosophila* X chromosome. Genetics 191:233-246.
Charlesworth B, Morgan MT, Charlesworth D. 1993. The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289-1303.
Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. PLoS Genet 2:e64.
Charlesworth D, Charlesworth B, Morgan MT. 1995. The Pattern of Neutral Molecular Variation Under the Background Selection Model. Genetics 141:1619-1632.
Comeron J. 2014. Background selection as a baseline for nucleotide variation across the *Drosophila* genome. PLoS Genet 10.
Coop G, Ralph P. 2012. Patterns of neutral diversity under general models of selective sweeps. Genetics 192:205-224.
Corbett-Detig RB, Hartl DL, Sackton TB. 2015. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. PLoS Biol 13:e1002112.
Cutter AD, Payseur BA. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. Nat Rev Genet 14:262-274.
dos Reis M, Wernisch L. 2009. Estimating translational selection in eukaryotic genomes. Mol Biol Evol 26:451-461.
Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, Coop G, Sella G. 2016. A genomic map of the effects of linked selection in *Drosophila*. PLoS Genet 12:e1006130.
Enard D, Messer PW, Petrov DA. 2014. Genome-wide signals of positive selection in human evolution. Genome Res 24:885-895.
Ewing GB, Jensen JD. 2016. The consequences of not accounting for background selection in demographic inference. Mol Ecol 25:135-141.
Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. Nat Rev Genet 8:610-618.
Eyre-Walker A, Keightley PD. 2009. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Mol Biol Evol 26:2097-2108.
Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. Genetics 173:891-900.
Falconer DS, Mackay FC. 1996. Introduction to Quantitative Genetics. Harlow: Longman.
Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, Yengo L, Rocheleau G, Froguel P, McCarthy MI, et al. 2016. Detection of human adaptation during the past 2000 years. Science 354:760-764.
Fisher RA. 1930. The distribution of gene ratios for rare mutations. Proc R Soc of Edinb 50:205-220.
Franchini LF, Pollard KS. 2017. Human evolution: the non-coding revolution. BMC Biology 15.
Garud NR, Messer PW, Buzbas EO, Petrov DA. 2015. Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. PLoS Genet 11:e1005004.
Garud NR, Petrov DA. 2016. Elevated linkage disequilibrium and signatures of soft sweeps are common in *Drosophila melanogaster*. Genetics 203:863-880.
Geraldes A, Basset P, Smith KL, Nachman MW. 2011. Higher differentiation among subspecies of the house mouse (Mus musculus) in genomic regions with low recombination. Mol Ecol 20:4722-4736.
Glemin S, Arndt PF, Messer PW, Petrov D, Galtier N, Duret L. 2015. Quantification of GC-biased gene conversion in the human genome. Genome Res 25:1215-1228.
Halligan DL, Kousathanas A, Ness RW, Harr B, Eory L, Keane TM, Adams DJ, Keightley PD. 2013. Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. PLoS Genet 9:e1003995.
Halligan DL, Oliver F, Eyre-Walker A, Harr B, Keightley PD. 2010. Evidence for pervasive adaptive protein evolution in wild mice. PLoS Genet 6:e1000825.
Halligan DL, Oliver F, Guthrie J, Stemshorn KC, Harr B, Keightley PD. 2011. Positive and negative selection in murine ultraconserved noncoding elements. Mol Biol Evol 28:2651-2660.

Hermisson J, Pennings PS. 2017. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods in Ecology and Evolution 8:700-716.
Hernandez RD, Kelly JL, Elyashiv E, Melton SC, Auton A, McVean G, Project G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. Science 331:920-924.
Hudson RR, Kaplan NL. 1995. Deleterious background selection with recombination. Genetics 141:1605-1617.
Keightley PD, Campos JL, Booker TR, Charlesworth B. 2016. Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of *Drosophila melanogaster*. Genetics 203:975-984.
Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. Genetics 177:2251-2261.
Kim Y. 2006. Allele frequency distribution under recurrent selective sweeps. Genetics 172:1967-1978.
Kim Y, Stephan W. 2000. Joint Efects of Genetic Hitchhiking and Background Selection on Neutral Variation. Genetics 155:1415-1427.
Kimura M, Ohta T. 1969. The average number of generations until fixation of a mutant gene in a finite population. Genetics 61:763-771.
King M-C, Wilson AC. 1975. Evolution at Two Levels in Humans and Chimpanzees. Science 188:107-116.
Kousathanas A, Halligan DL, Keightley PD. 2014. Faster-X adaptive protein evolution in house mice. Genetics 196:1131-1143.
Kousathanas A, Keightley PD. 2013. A comparison of models to infer the distribution of fitness effects of new mutations. Genetics 193:1197-1208.
Kousathanas A, Oliver F, Halligan DL, Keightley PD. 2011. Positive and negative selection on noncoding DNA close to protein-coding genes in wild house mice. Mol Biol Evol 28:1183-1191.
Lowe CB, Kellis M, Siepel A, Raney BJ, Clamp M, Salama SR, Kingsley DM, Lindblad-Toh K, Haussler D. 2011. Three Periods of Regularory Innovation During Vertebrate Evolution. Science 333:pp. 1019-1024.
Maynard Smith J, Haigh J. 1974. The hitch-hiking effect of a favourable gene. Genetical Research 23:23-25.
McDonald MJ, Rice DP, Desai MM. 2016. Sex speeds adaptation by altering the dynamics of molecular evolution. Nature 531:233-236.
McVicker G, Gordon D, Davis C, Green P. 2009. Widespread genomic signatures of natural selection in hominid evolution. PLoS Genet 5:e1000471.
Messer PW. 2013. SLiM: simulating evolution with selection and linkage. Genetics 194:1037-1039.
Messer PW, Petrov DA. 2013. Frequent adaptation and the McDonald-Kreitman test. Proc Natl Acad Sci 110:8615-8620.
Nam K, Munch K, Mailund T, Nater A, Greminger MP, Krutzen M, Marques-Bonet T, Schierup MH. 2017. Evidence that the rate of strong selective sweeps increases with population size in the great apes. Proc Natl Acad Sci U S A 114:1613-1618.
Nordborg M, Charlesworth B, Charlesworth D. 1996. The effect of recombination on background selection. Genetical Research 67:159-174.
Otto SP, Whitlock MC. 1997. The Probability of Fixation in Populations of Changing Size. Genetics 146.
Pennycuik PR, Johnston PG, Westwood NH, Reisner AH. 1986. Variation in number in a house mouse population housed in a large outdoor enclosure. Journal of Animal Ecology 55:371-391.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20:R208-215.
Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. 2011. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in *Drosophila simulans*. PLoS Genet 7:e1001302.

Schneider A, Charlesworth B, Eyre-Walker A, Keightley PD. 2011. A method for inferring the rate of occurrence and fitness effects of advantageous mutations. Genetics 189:1427-1437.
Schrider DR, Kern AD. 2017. Soft sweeps are the dominant mode of adaptation in the human genome. Mol Biol Evol.
Schrider RD, Shanku GA, Kern DA. 2016. Effects of linked selective sweeps on demographic inference and model selection. Genetics.
Sella G, Petrov DA, Przeworski M, Andolfatto P. 2009. Pervasive natural selection in the *Drosophila g*enome? PLoS Genet 19.
Smagulova F, Brick K, Yongmei P, Camerini-Otero RD, Petukhova GV. 2016. The Evolutionary Turnover of Recombination Hotspots Contributes to Speciation in Mice. Genes & Development 30:277-280.
Stephan W. 2010. Genetic hitchhiking versus background selection: the controversy and its implications. Philos Trans R Soc Lond B Biol Sci 365:1245-1253.
Tajima F. 1989. Statistical Method for Testing the Neutral Mutation Hypothesis by Polymorphism. Genetics 123:585-595.
Tataru P, Mollion M, Glemin S, Bataillon T. 2017. Inference of Distribution of Fitness Effects and Proportion of Adaptive Substitutions from Polymorphism Data. Genetics 207:1103-1119.
Teschke M, Mukabayire O, Wiehe T, Tautz D. 2008. Identification of Selective Sweeps in Closely Related Populations of the House Mouse Based on Microsatellite Scans. Genetics 180:1537-1545.
Uricchio LH, Hernandez RD. 2014. Robust Forward Simulations of Recurrent Hitchhiking. Genetics 197:221-236.
Wiehe T, Stephan W. 1993. Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. Mol Biol Evol 10:842-854.
Williamson RJ, Josephs EB, Platts AE, Hazzouri KM, Haudry A, Blanchette M, Wright SI. 2014. Evidence for widespread positive and negative selection in coding and conserved noncoding regions of *Capsella grandiflora*. PLoS Genet 10:e1004622.

Figure Legends

Figure 1 The uSFS for three classes of functional sites (yellow and blue bars) compared to a putatively neutral comparator (grey bars). The neutral comparator for 0-fold sites and UTRs was 4-fold degenerate synonymous sites in both cases. For CNEs, the neutral comparator was CNE-flanking sequence. The expected uSFS under a demographic model fitted to a neutral comparator was used to correct the uSFS for the corresponding selected sites (see *Methods*).

Figure 2 Estimates of scaled diversity (π/π_{Ref}) around protein-coding exons and CNEs (black lines) in *M. m. castaneus* compared to results from simulations (colored ribbons). The panels show diversity observed in simulated populations assuming DFE estimates obtained by analysis of the full uSFS (Model A) or when sites fixed for the derived allele do not influence selection parameters (Model B). Colored ribbons represent 95% confidence intervals obtained from 1,000 bootstrap samples.

Figure 3 Estimates of scaled diversity (π/π_{Ref}) plotted against genetic distance from exons and conserved non-coding elements (CNEs) in *M. m. castaneus* compared to results from simulations (colored ribbons). The panels show diversity observed in simulated populations assuming DFE estimates obtained by analysis of the full uSFS (Model A) or when sites fixed for the derived allele do not influence selection parameters (Model B). Nucleotide diversity (π) is scaled by the mean diversity at population-scaled genetic distances ($4N_e r$) more than 1,500 from exons and 200 from CNEs (π_{Ref}). Colored ribbons represent 95% confidence intervals obtained from 1,000 bootstrap samples. Diversity downstream of functional elements is shown.

Figure 4 Tajima's D around protein-coding exons and CNEs in *M. m. castaneus* compared to simulated data. The black lines show Tajima's D computed from the *M. m. castaneus* genome sequence data around protein-coding exons or CNEs. The colored ribbons show the 95% bootstrap intervals from simulated data assuming the DFEs estimated under either Model A (i.e. analyzing the full uSFS) or Model B (i.e. fixed derived sites do not contribute to the likelihood for selection parameters).

Figure 5 Nucleotide diversity (π) around substituted sites in *M. m. castaneus* compared to the same pattern obtained from simulation data. Nucleotide diversity in *M. m. castaneus* was scaled by divergence between mouse and rat to correct for variation in local mutation rates. The *M. m. castaneus* data are from Halligan, et al. (2013).

Figure 6 Patterns of nucleotide diversity around protein-coding exons in simulations assuming strongly selected advantageous mutations. Nucleotide diversity in *M. m. castaneus* is shown in black. Nucleotide diversity (π) is scaled by the mean diversity at distances more than 75 Kbp from exons (π_{Ref}). Colored ribbons represent 95% confidence intervals obtained from 1,000 bootstrap samples.

Figure 1

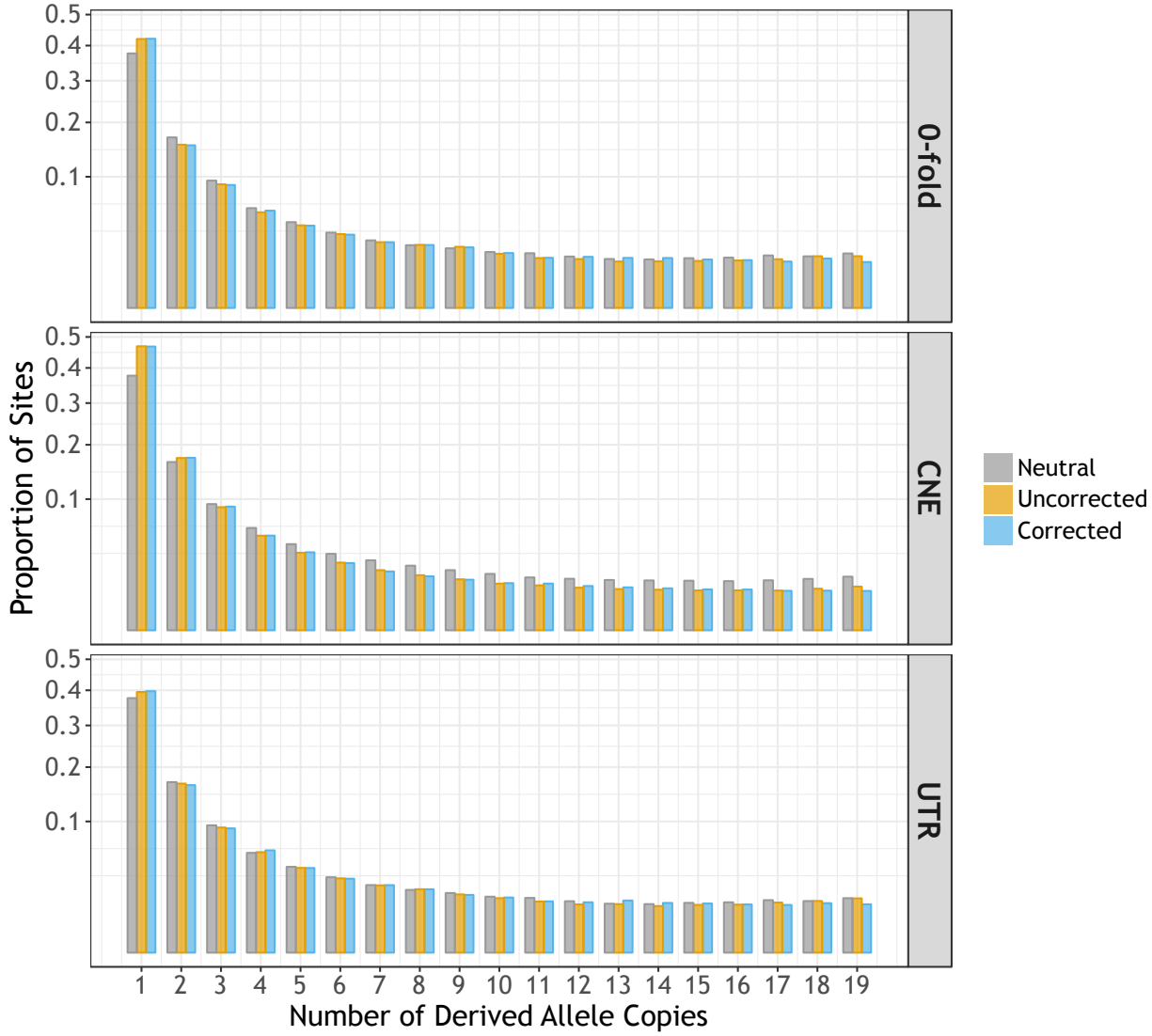


Figure 2

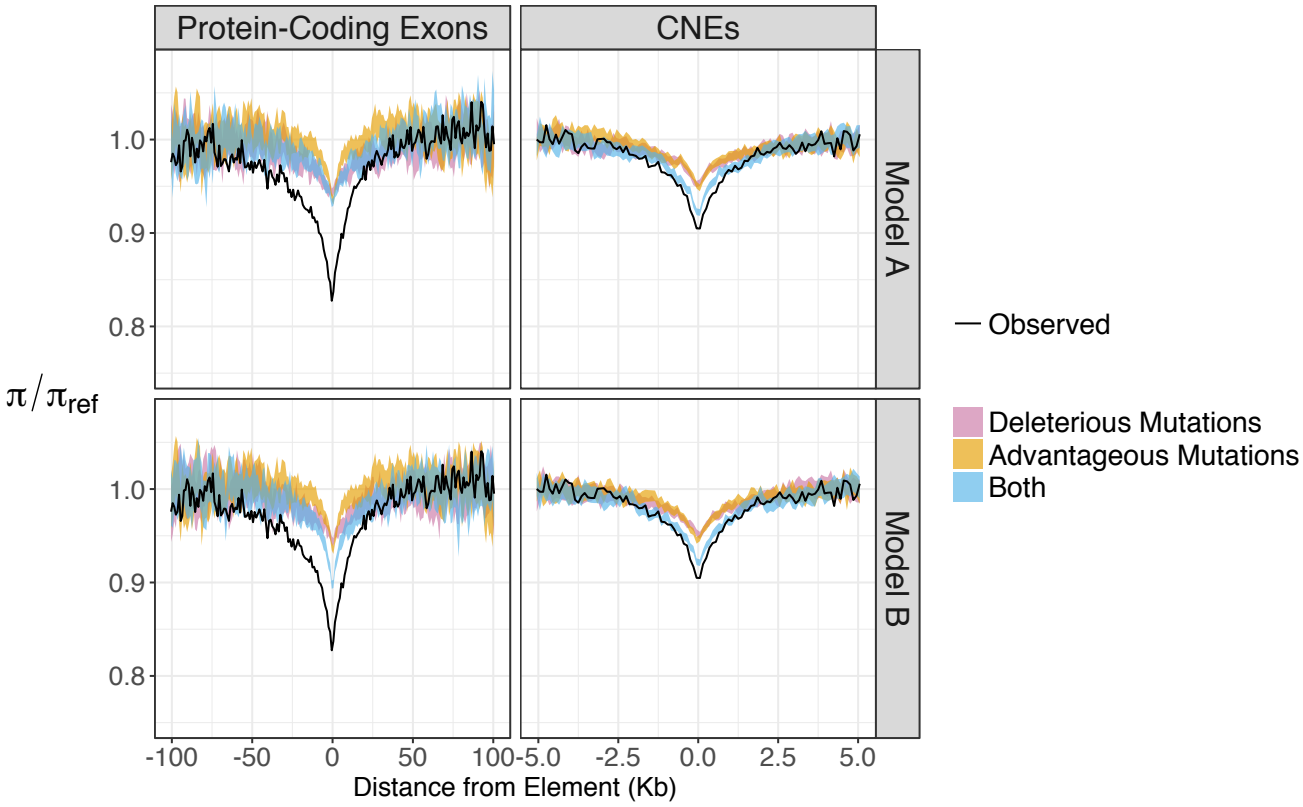


Figure 3

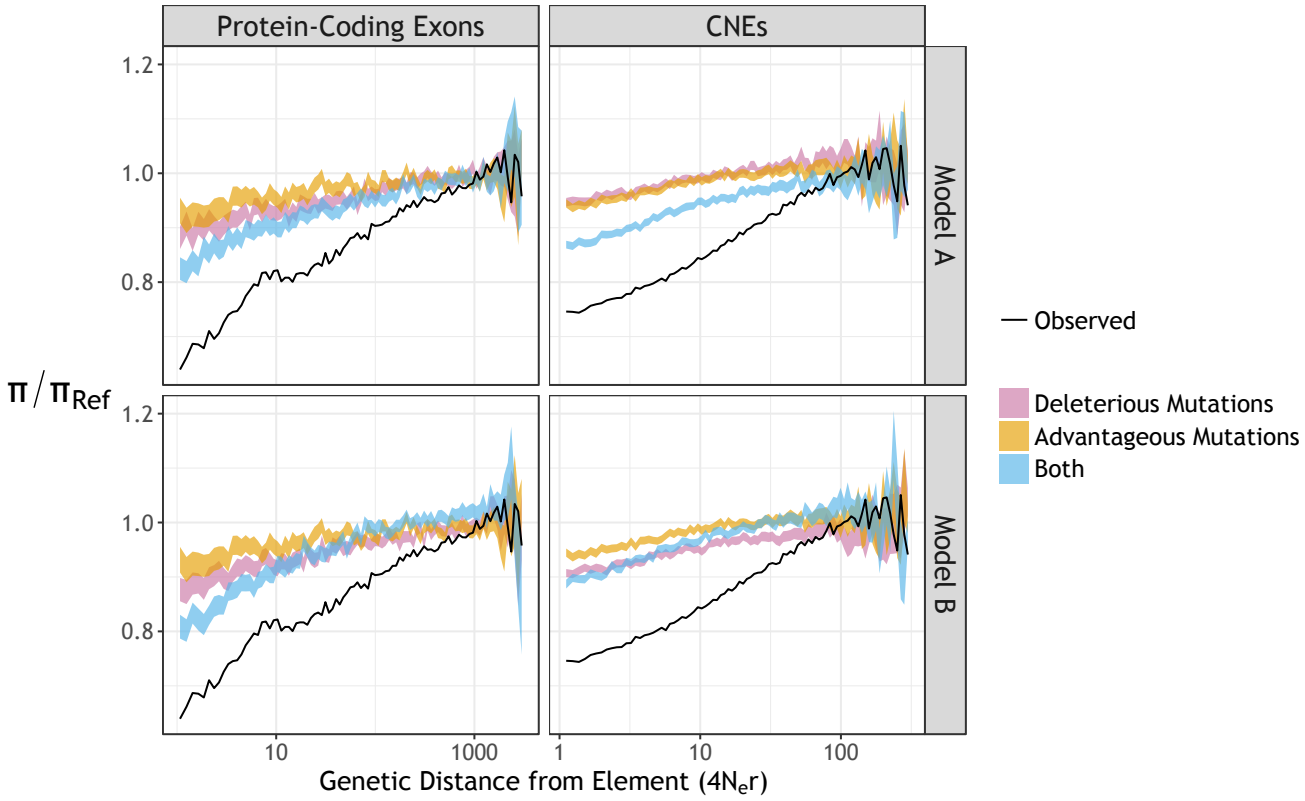


Figure 4

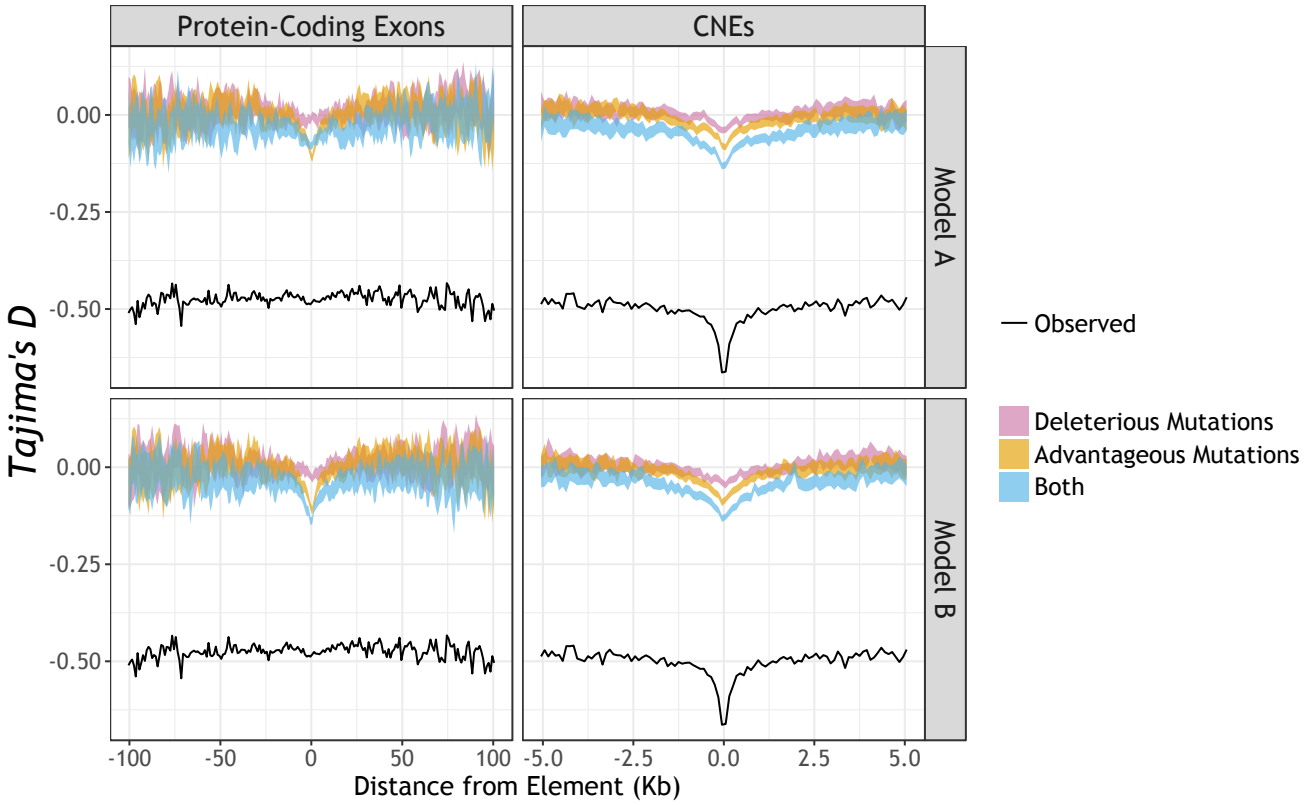


Figure 5

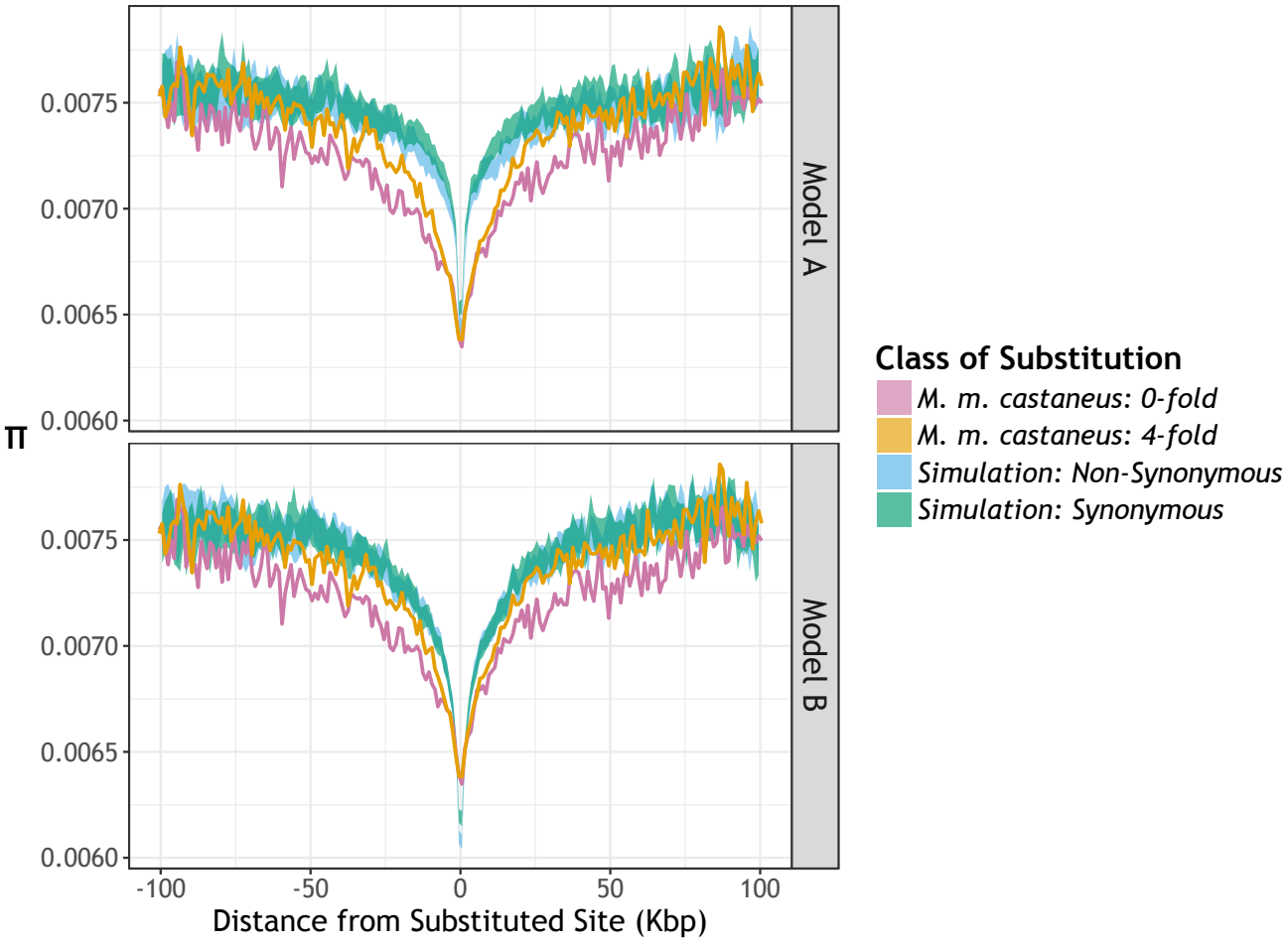
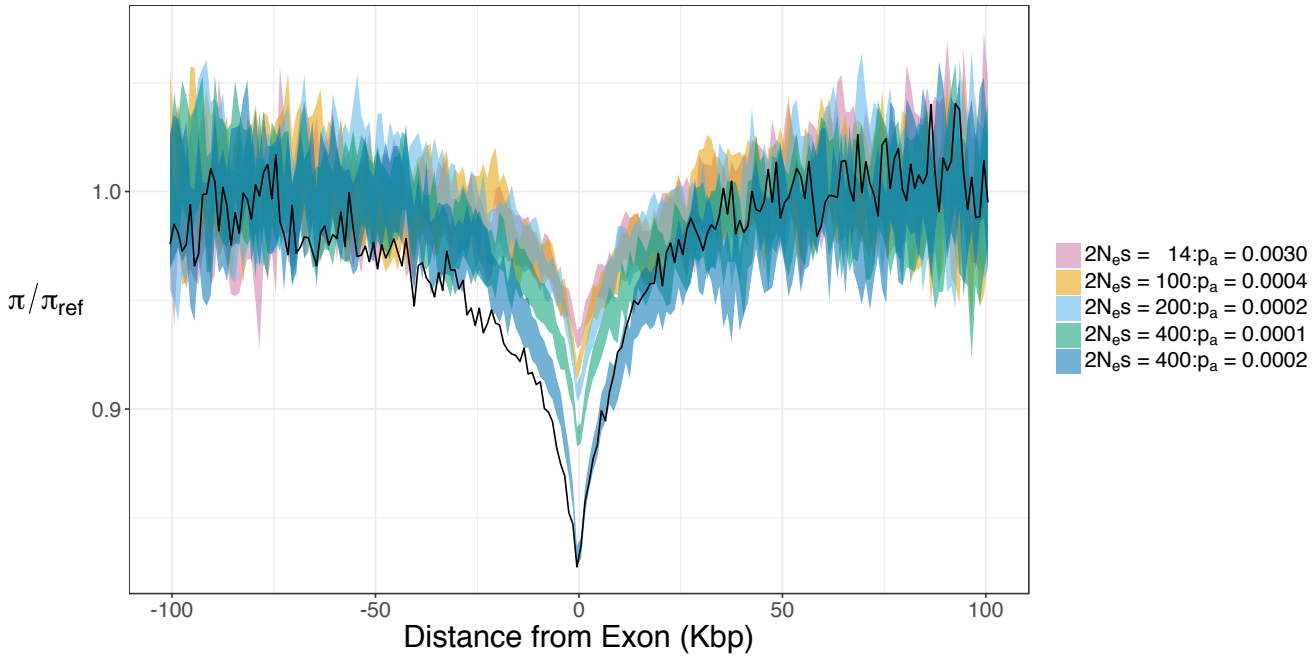


Figure 6



Tables

Table 1. Summary statistics for five classes of sites in *M. m. castaneus.*

	π (%)	Tajima's *D*	d_{fam} (%)	d_{rat} (%)	Sites (Mb)
0-fold	0.134	-0.763	0.239	2.93	10.2
4-fold	0.628	-0.627	1.06	12.7	1.49
CNE	0.274	-1.03	0.418	3.67	24.6
CNE flank	0.670	-0.602	1.03	13.8	17.8
UTR	0.438	-0.702	0.802	10.0	11.3

All values refer to non-CpG prone sites. Nucleotide divergences between *M. m. castaneus* and *M. famulus* (d_{fam}) and between *M. m. castaneus* and *R. norvegicus* (d_{rat}) were estimated by maximum likelihood using the method described in Keightley, et al. (2016).

Table 2. Parameter estimates for the distribution of fitness effects for three classes of sites in *M. m. castaneus* obtained under two models.

Model A: DFE inferred from the full uSFS			
	0-fold	**UTR**	**CNE**
$2N_es_1$	-0.090	-0.194	-0.646
p_1	0.191	0.701	0.352
$2N_es_2$	-208	-78.2	-7.96
p_2	0.806	0.286	0.278
$2N_es_3$	-	-	-155.8
p_3	-	-	0.360
$2N_es_a$	14.54 [9.24 – 23.4]	10.64 [7.82 – 14.1]	18.34 [14.0 – 41.8]
p_a	0.0030 [0.0019 – 0.0048]	0.013 [0.0097 – 0.019]	0.0098 [0.0037 – 0.0099]
Model B: Sites fixed for the derived allele do not contribute to parameter estimates			
	0-fold	**UTR**	**CNE**
$2N_es_1$	-0.342	-0.320	-0.506
p_1	0.184	0.689	0.342
$2N_es_2$	-200	-64.0	-7.68
p_2	0.806	0.281	0.286
$2N_es_3$	-	-	-152.6
p_3	-	-	0.365
$2N_es_a$	16.6 [12.5 – 20.2]	13.9 [11.1 – 17.4]	17.2 [8.74 – 25.2]
p_a	0.010 [0.0030 – 0.0183]	0.0294 [0.0181 – 0.0436]	0.008 [0.0004 – 0.0100]

The first (Model A) estimates of selection parameters based on the full uSFS. Under the second method (Model B), sites fixed for the derived allele were prevented from influencing estimates of selection parameters. The bracketed values are 95% confidence intervals obtained from profile likelihoods. The parameters shown are: p_i = the proportion of mutations falling into the i^{th} deleterious

class; $2N_e s_i$ = the scaled homozygous selection coefficient of the i^{th} deleterious class; p_a = the proportion of advantageous mutations; $2N_e s_a$ = the scaled homozygous selection coefficient of the advantageous mutation class.

Table 3. The root mean square difference between values of π around functional elements predicted in simulations and π observed in *M. m. castaneus*.

		Exons		CNEs	
		Median	95% range	Median	95% range
Model A	*Deleterious Mutations*	0.0327	0.0311 - 0.0343	0.0164	0.0151 - 0.0179
	Advantageous Mutations	0.0422	0.0403 - 0.0442	0.0177	0.0161 - 0.0195
	Both	0.0312	0.0297 - 0.0340	0.0100	0.0088 - 0.0113
Model B	*Deleterious Mutations*	0.0331	0.0314 - 0.0351	0.0157	0.0144 - 0.0171
	Advantageous Mutations	0.0380	0.0355 - 0.0406	0.0162	0.0147 - 0.0179
	Both	0.0274	0.0253 - 0.0294	0.0088	0.0078 - 0.0101

Confidence intervals were obtained from 1,000 bootstrap samples (*see Methods*). Values shown are for the patterns of diversity when measured on the scale of physical distance.

Table 4. Comparison of the accumulation of nucleotide divergence in simulated populations between different functional site types.

		Simulation DFE			
	M. m. castaneus	**Model A**		**Model B**	
Site Class	d_{sel}/d_{neu}	d (%)	d_{sel}/d_{neu}	d (%)	d_{sel}/d_{neu}
0-fold	0.225	1.66	0.221	2.26	0.301
UTR	0.757	5.76	0.767	6.85	0.914
CNE	0.406	3.31	0.440	3.07	0.409

In the cases of 0-fold sites and UTRs, d_{neu} refers to 4-fold sites. For CNEs, d_{neu} refers to CNE flanking sites. In all simulations, d_{neu} was set to 7.5%.