Chapter 4: Estimating the parameters of SSWs from patterns of genetic diversity in the house mouse genome

Tom R. Booker[1,*]

[1]Institute of Evolutionary Biology, University of Edinburgh, Edinburgh

July 16, 2018

Abstract

Introduction

Genetically linked sites do not evolve independently, so selection acting at one site may influence the fate of another. The consequences of selection at linked sites are intrinsically linked to the frequency and strength of selected mutations as well as, crucially, the rate of recombination (Maynard Smith and Haigh, 1974; Hudson and Kaplan, 1995; Braverman et al., 1995; Nordborg et al., 1996). Two main modes of selection at linked sites have been identified; selective sweeps (SSWs) caused by the spread of advantageous mutations and background selection (BGS) caused by the removal of deleterious variants. The two processes are related and can both potentially explain the positive correlations between nucleotide diversity and recombination rate reported in many species (Cutter and Payseur, 2013). However, the

proportion of nonsynonymous substitutions attributable to adaptive evolution (α) is typically high (50%) (Galtier 2016; but see Booker et al. 2017a for caveats), suggesting that SSWs may play a substantial role in shaping nucleotide diversity across the genomes of many species.

SSWs have been subject to rigorous population genetic research (Maynard Smith and Haigh, 1974; Coop and Ralph, 2012; Hermisson and Pennings, 2005; Barton, 2000). The classic footprint of a selective sweep is a trough in nucleotide diversity at neutral sites surrounding substitutions. The reductions in nucleotide diversity caused by SSWs are related to the strength of selection acting on advantageous mutations as well as the frequency with which they arise. Taking advantage of this, Wiehe and Stephan (1993) used a model of SSWs to estimate the frequency and strength of advantageous mutations in *Drosophila melanogaster* by fitting the positive correlation between recombination rate and nucleotide diversity. At the time of their analysis, the theory of BGS was in its infancy and models combining the effects of BGS and sweeps had not been developed. However, the effects of BGS are expected to be ubiquitous across the genome (Comeron, 2014; Elyashiv et al., 2016; McVicker et al., 2009), and studies, conceptually similar to Wiehe and Stephan's (1993), have shown that controlling for BGS is highly important when parametrizing sweep models from patterns of nucleotide diversity (Campos et al., 2017; Elyashiv et al., 2016).

Both SSWs and BGS act to reduce nucleotide diversity, so it has proven difficult to distinguish their effects using population genetic data (Stephan, 2010). A number of different approaches have been taken to tease apart the effects of the two processes. For instance, Sattath et al. (2011) showed that, on average, there is a trough in diversity around recent nonsynonymous protein-coding substitutions in *Drosophila melanogaster* but not around synonymous ones. This pattern is strongly suggestive of SSWs, so Sattath et al. (2011) fitted a sweep model to the trough they observed and estimated that strongly advantageous mutations ($2N_e s \approx 5{,}000$) occur in the fruitfly's genome. In the house mouse, there is also a trough in diversity around recent nonsynonymous substitutions, but an almost identical trough is observed

around synonymous substitutions, furthermore a similar trough is observed around even randomly selected synonymous and nonsynonymous sites in the genome (Halligan et al., 2013). This all, perhaps, suggests that the reductions in diversity caused by selection at linked sites extend beyond the average distance separating nonsynonymous substitutions, so that the methods employed by Sattath et al. (2011) are not effective in mice (Halligan et al., 2013). However, values of $\alpha \geq 0.19$ have been reported for multiple classes of functional elements (Halligan et al., 2013) and BGS alone cannot fully explain observed patterns of diversity (Halligan et al. 2013, Chapter 3), suggesting that SSWs do contribute to the observed patterns in mice.

In Chapter 3, we sought to tease apart the contribution of BGS and SSWs to patterns of diversity in mice. We estimated distributions of fitness effects (DFEs) for both harmful and advantageous mutations occurring in multiple classes of functional sites, analysing the distribution of derived allele frequencies (referred to as the unfolded site frequency spectrum, hereafter uSFS). The methods that we used, and related approaches, rely on the assumption that selected mutations segregate in populations of interest, such that they affect the shape of the uSFS. Using simulations, we found that neither BGS nor SSWs, given the selection parameters we estimated from the uSFS, could explain troughs in diversity observed around protein-coding exons and conserved non-coding elements (CNEs). A possible explanation for this finding is that advantageous mutations which have large effects on fitness, and which cause the greatest reduction in neutral diversity, may not be detectable by analysis of the uSFS.

In this study, we use a model of SSWs to estimate the strength and frequency of advantageous mutations that occur within protein-coding exons and regulatory elements. Using simulations, we show that the selection parameters that explain the troughs in diversity are out of the range detectable by analysis of the uSFS. We find that the strength of selection acting on protein-coding exons is far greater than that acting in regulatory elements. Finally, using a simple model of the fitness change brought about by adaptive evolution, we show that, despite adaptation occurring more frequently in regulatory regions,

adaptation in protein-coding regions may contribute more to phenotypic evolution in mice.

Materials and Methods

Model of SSWs with BGS

Campos et al. (2017) gave expressions for the neutral diversity expected under the combined effects of BGS (BGS) and SSWs (SSWs). They assumed that the effects of BGS and SSWs act independently so that their effects can simply be summed. However, BGS causes a reduction to the effective population size (N_e) at a neutral locus, k by some fraction B_k, so may influence the rate and fixation probability of new advantageous mutations since both are dependant upon N_e. We scale the sweep effect by B_k in a modified version of the model used by Campos et al. (2017),

$$\frac{\pi_k}{\pi_0} \approx \frac{1}{B_k^{-1} + B_k 2N_e P_{sc,k}}. \tag{1}$$

Where π_k is genetic diversity observed at neutral site k and π_0 is diversity expected in the absence of selection at linked sites. $P_{sc,k}$ is the reduction in coalescence times at site k caused by the effects of SSWs,

$$P_{sc,k} \approx V_a \tau \gamma_a^{\frac{-4r_{i,k}}{s}} \tag{2}$$

The term $V_a = 2\mu p_a \gamma_a$ is the rate of sweeps per base pair per generation, where μ is the point mutation rate, p_a is the proportion of new mutations that are advantageous and γ_a is the scaled selection coefficient ($2N_e s_a$) of those mutations (Kimura and Ohta 1971). τ is the number of selected sites in a functional element and the recombination fraction between a functional element (i) and the focal neutral

site is $r_{i,k}$. When assuming that recombination proceeds solely by crossing over $r_{i,k}$ is simply the product of the physical distance ($d_{i,k}$) and the local crossing-over rate (r_c). When incorporating gene conversion, we use Equation 1 from Frisse et al. (2001):

$$r_{i,k} = d_{i,k} r_c + g_c d_g \left(1 - e^{-\frac{d_{i,k}}{d_g}}\right) \tag{3}$$

where g_c is the rate of gene conversion and d_g is the mean gene conversion tract length, assuming that the distribution of tract lengths is exponential. When applying Equation 3 we use $g_c = \kappa r_c$ where κ is the ratio of the non-crossovers to crossovers.

Both theoretical and experimental results suggest that the distribution of fitness effects for advantageous mutations is exponential (Eyre-Walker and Keightley, 2007). We incorporate an exponential distribution of advantageous mutation effects to Equation 3 as follows:

$$P_{sc,k} \approx \int_0^\infty f(\gamma) V_a \tau \gamma_a^{\frac{-4r_{i,k}}{s}} \, d\gamma \tag{4}$$

We estimated γ_a and p_a by fitting Equation 1 to the relationship between nucleotide diversity and genetic distance to functional elements using non-linear least squares with the *lmfit* (0.9.7) package for Python 2.7. When analysing the mouse data, see below, we compared the fit of Equation 1 incorporating either one or two discrete classes of advantageous mutations (Equation 2) or the exponential distribution (Equation 4) using Aikieke's Information Criterion (AIC).

Analysis of Mouse Data

We analysed patterns of genetic diversity present in 10 wild-caught *M. m. castaneus* individuals, first reported by Halligan et al. (2013). Briefly, Halligan et al. (2013) sequenced individual genomes to high coverage ($\approx$ 30x) using Illumina paired-end reads, which were then mapped to the mm9 mouse reference genome using BWA. Variants were called using a Samtools pipeline. Note that we only analyse

SNP data in this study, insertion/deletion variants are not included. For further details of the sequencing and variant calling methods see Halligan et al. (2013). Protein-coding exons present in version 67 of the Ensembl annotation database and the locations of conserved non-coding elements (CNEs) identified by Halligan et al. (2013) using an alignment of placental mammals were used in this study. The mean length of a protein-coding exon is 151bp, of which we assume 75% of sites are subject to selection and the mean length of a conserved non-coding element is 51bp, of which 100% of sites are subject to selection.

From the edges of exons (CNEs), polymorphism data and divergence to the rn4 rat reference genome were extracted for non-CpG sites in windows of 1Kbp (100bp) extending to distances of 100Kbp (5Kbp). Analysis windows were then binned based on genetic distance to the focal element. Genetic distances were calculated using either the LD-based recombination map for *M. m. castaneus* constructed by Booker et al. (2017b) or the pedigree-based genetic map constructed using common lab strains of *M. musculus* by Cox et al. (2009). Because LD-based and pedigree-based recombination maps have different benefits and drawbacks (*see Results*), we performed analyses based on both of these maps in parallel. Genetic distances calculated using the Cox et al. (2009) map were scaled assuming an N_e of 426,200.

Compared to crossing-over rates, gene conversion parameters are very difficult to estimate (Paigen and Petkov, 2010). Empirical estimates of the ratio of crossovers to non-crossovers (a parameter we have termed κ) vary across orders of magnitude in mammals (Paigen and Petkov, 2010). Paigen et al. (2008) measured non-crossover gene conversion rates in three recombination hotspots in mice and estimated a mean gene conversion tract length of 144bp and $\kappa = 0.105$, we refer to this estimate as the low gene conversion rate. Values of κ as high as 12.0 have been reported in humans (Paigen and Petkov 2010), so to explore the effects of high gene conversion rates on the parameters of SSWs inferred from models of selection at linked sites, we also assumed $\kappa = 12.0$, which we refer to as the high gene conversion rate.

In order to disentangle the sweep parameters we obtained, we assume a point mutation rate of 5.4×10^{-9}, which is based on a mutation-accumulation experiment in *M. musculus* (Uchimura et al.,

2015).

Estimates of B

BGS contributes to the troughs in diversity around both protein-coding exons and CNEs (Halligan et al. 2013; Chapter 3). Because of this, we required estimates of the effect of BGS on neutral diversity, B, to fit as a covariate when fitting Equation 1 to diversity troughs. There are formulae for calculating B given the DFE as well as mutation and recombination rates (Nordborg et al., 1996; Hudson and Kaplan, 1995), but these over-predict the effects of BGS when purifying selection is weak ($\gamma_d < 1$) (Good and Desai 2014; Gordo et al. 2002). Since weakly deleterious mutations comprise a large portion of the DFEs previously obtained for mice (Halligan et al. 2013, Chapter 3), we opted to obtain estimates of B from simulations. In Chapter 3, we used simulations to estimate the contribution of BGS to patterns of nucleotide diversity around both protein-coding exons and CNEs. These simulations incorporated recombination rate variation, the actual distribution of functional elements in the genome and dDFEs specific to each of the functional elements analysed. By extracting $\frac{\pi}{\pi_0}$ as a function of genetic distance to both protein-coding exons and CNEs from these simulated data, we obtained estimates of B that can be used when fitting Equation 1.

The simulations we used to estimate B were the same as those we used in Chapter 3, except that we increased the number of simulation replicates from 2,000 to 6,000. To obtain smoothed B values we fit Loess curves to the simulation data using R (v3.4.2) using a span of 0.2 and using the number of sites contributing to each analysis bin as weights.

Simulations to obtain uSFS

In Chapter 3, we argued that strongly selected advantageous mutations are difficult to detect by analysis of the uSFS. To these this hypothesis, we generated simulated datasets using the forward-time simulation package SLiM (v1.8; Messer 2013). We simulated the evolution of 1Mbp chromosomes containing 20 evenly spaced out 'genes'. Each 'gene' consisted of 10 100bp exons, separated by 1Kbp of neutrally evolving intronic sequence. Nonsynonymous mutations were modelled as 75% of mutations occurring in exons, the remaining 25% were strictly neutral (i.e. synonymous sites). We varied the γ_a and p_a parameters across simulations, but kept the product $\gamma_a p_a$ equal to 0.1. We based this value of $\gamma_a p_a \approx 0.1$ on a recent study in *D. melanogaster* (Keightley et al., 2016). All simulations incorporated the same gamma dDFE (shape parameter $\beta = 0.2$ and mean $\hat{\gamma_d}$ = -1,000), but the advantageous mutation parameters varied, these are listed in Table 2. The population-scaled mutation and recombination rates (i.e. $\theta = 4N_e\mu$ and $\rho = 4N_e r$, respectively) were set to 0.01. Populations of N = 1,000 diploid individuals were simulated for an initial burn-in of $10N$ generations to establish equilibrium conditions. After the burn-in, 20 haploid chromosomes were sampled every $2N$ generations for a further $100N$ generations. We performed 10 replicate simulations for each set of selection parameters (Table 2). Across simulation replicates, time-points and loci we extracted the simulated nonsynonymous and synonymous sites, giving uSFS data for 10,000 'genes'. We sampled the set of 10,000 'genes' with replacement 100 times, collating the nonsynonymous and synonymous site uSFSs for each replicate.

We estimated our simulated DFEs by analysis of the uSFS using the methods of Tataru et al. (2017), as implemented in the polyDFE (v1.1) package. polyDFE fits an expression for the uSFS expected in the presence of both advantageous and deleterious mutations to data from putatively neutral and selected classes of sites, and estimates parameters by maximum likelihood. The neutral class is used to determine distortions to the uSFS caused by processes such as selection at linked sites and a history of population size change. In addition, polyDFE corrects for polymorphism misattributed to divergence, mutation rate

variability and error in assigning sites as ancestral/derived. Tataru et al. (2017) performed extensive simulations and showed that accurate estimates of the parameters for both deleterious and advantageous mutations can be obtained using their methods. However, there are a range of parameters that they did not test which may be biologically relevant, specifically when advantageous mutations are strongly selected, but infrequent.

We analysed the simulated uSFS data using polyDFE choosing Model C (a gamma dDFE and a discrete class of advantageous mutations) and either including or not between-species divergence. We analysed the uSFS for simulated nonsynonymous sites using synonymous sites as the neutral reference class. For each of the advantageous mutation parameter sets tested, we analysed 100 bootstrap samples of the simulation data.

Results

Patterns of genetic diversity around protein-coding exons and conserved non-coding elements

Recombination rates can be estimated in various ways, which have different pros and cons. For instance, the population-scaled recombination rate (ρ) can be inferred from a relatively small sample of unrelated individuals at very fine-scales using patterns of linkage disequilibrium (LD). However, selection at linked sites influences local LD and may therefore affect recombination rate estimates obtained in this way (Clark REVIEW). Alternatively, direct estimates of the recombination rate (r) can be obtained from crossing experiments, but to achieve sufficient power to generate recombination maps a very large number of individuals need to be genotyped, which has typically precluded the use of whole-genome re-sequencing, limiting resolution. In summary, high resolution recombination maps can be generated using patterns of LD, but these may be biased by selection at linked sites, while unbiased recombination maps may be

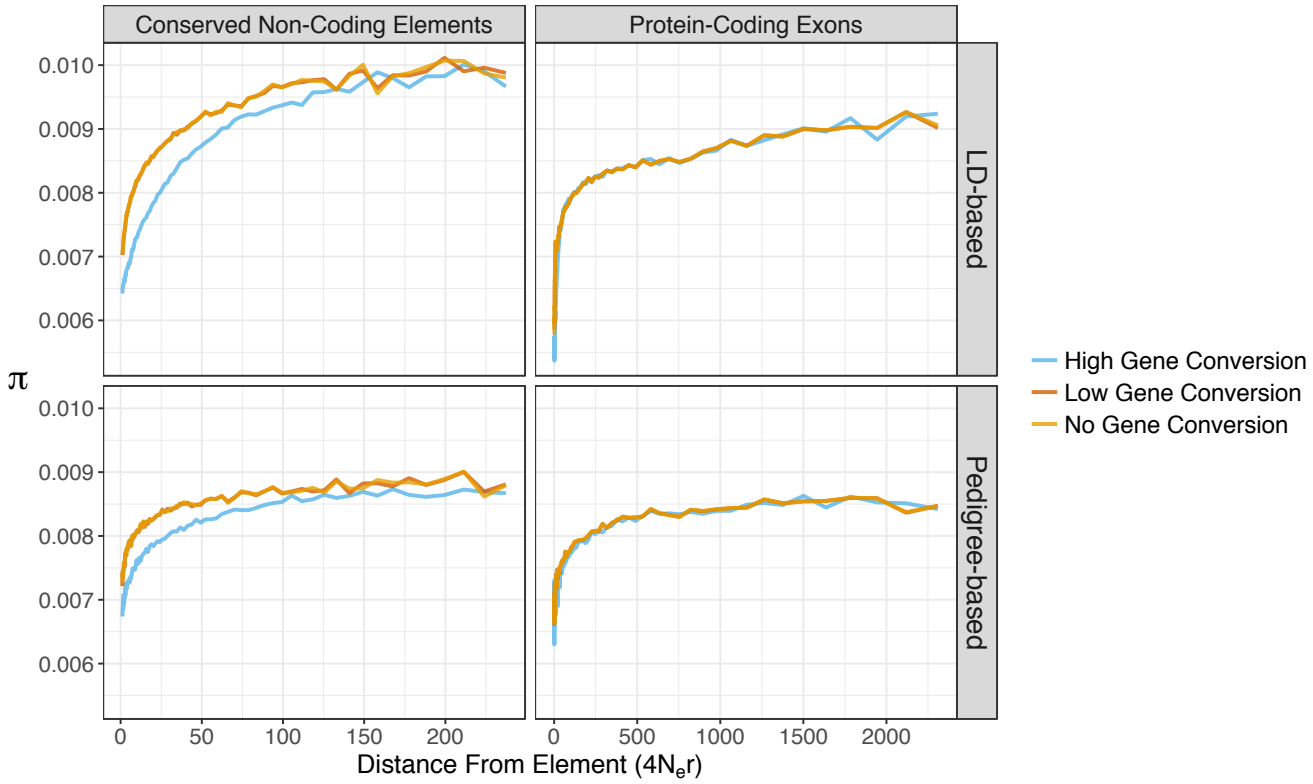

Figure 1: Nucleotide diversity in regions surrounding protein-coding exons and conserved non-coding elements in wild mice. Population-scaled genetic distances($4N_er$) were calculated using either an LD-based recombination map constructed for *M. m. castaneus* or the pedigree based *M. musculus* genetic map constructed by Cox et al. (2009). Gene conversion was included assuming either the gene conversion rates

.

generated using crosses, though these typically have low resolution. When analysing patterns of genetic diversity using a model of selection at linked sites, the way in which recombination rate estimates were obtained may, therefore, affect parameter estimates.

In this study, we analysed the relationship between nucleotide diversity and genetic distances from funcitonal eleemnts in *M. m. castaneus* assuming either the high resolution recombination map constructed from LD by Booker et al. (2017b) (the *castaneus* map) or the pedigree-based map of Cox et al.

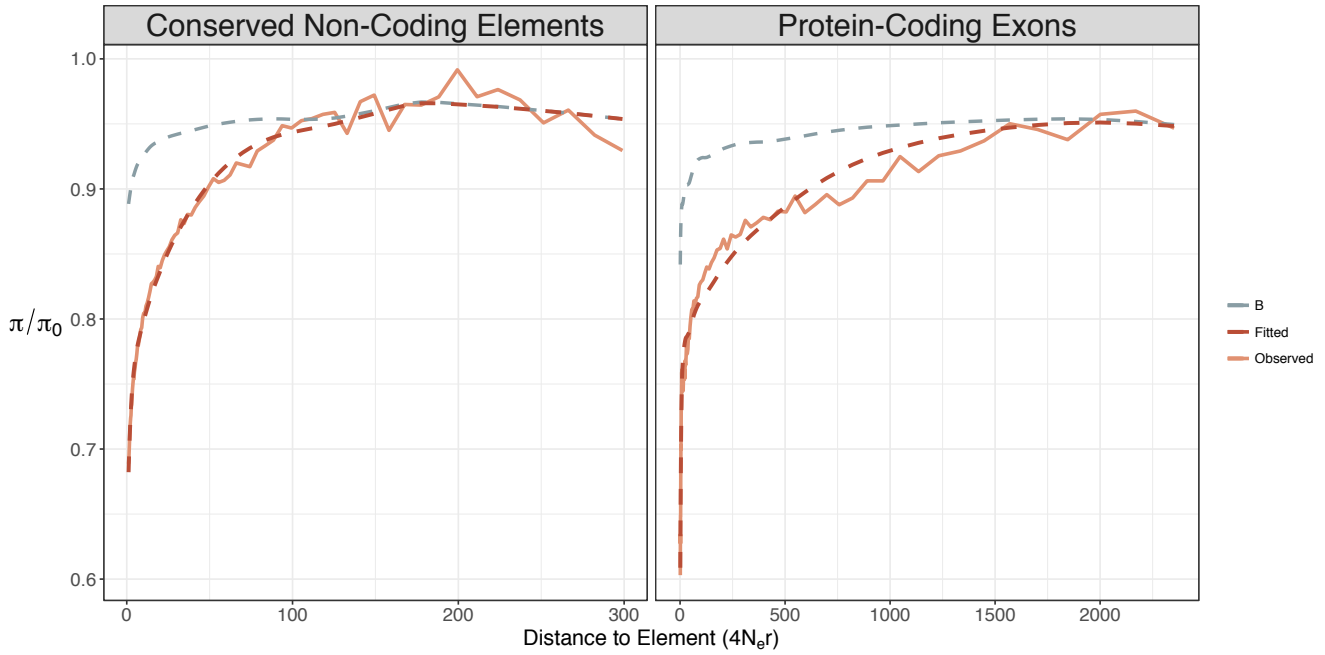

Figure 2: The pattern of scaled nucleotide diversity around protein-coding exons and CNEs in *M. m. castaneus*.

(2009) (the Cox map). The choice of recombination map had a substantial effect on patterns of nucleotide diversity. We found that, in the immediate flanks of both exons and CNEs, diversity was lower when assuming the LD-based *castaneus* map than when assuming the pedigree-based Cox map (Figure 1). This difference is consistent with the idea that regions of the genome close to functional elements, where the effects of BGS and/or SSWs are strongest, and thus exhibit reduced diversity, may yield downwardly biased estimates of the recombination rate obtained from LD. An alternative explanation is that the Cox map, which lacks resolution, does not fully capture regions of low recombination rate, causing analysis windows that are tightly linked to functional elements to appear less tightly linked. However, genetic diversity plateaus at a higher level when assuming the *castaneus* map, suggesting that the Cox map may not capture some of the highly recombining portions of the genome. The choice of recombination map will, therefore, have an impact on the parameters of selection inferred from the patterns of diversity. Throughout the rest of this chapter, we present, in parallel, the results of analyses based on the *castaneus* map with those based on the Cox map.

Diversity expected in the absence of selection, π_0

A key parameter in Equation 1 is π_0, the nucleotide diversity expected in the absence of the effects of selection at linked sites. This parameter is very difficult to estimate and may even prove unobservable in real data given the ubiquity of the effects of selection at linked sites (Kern and Hahn, 2018), yet it is required. One strategy for estimating π_0 would be to divide the mean π in regions sufficiently distant to functional elements by the corresponding value of B. B plateaus at approximately 0.95 in regions surrounding both protein-coding exons and CNEs, but the level at which observed diversity plateaus is different for the two classes of elements (Figure), so it may be that SSWs at elements linked to exons have reduced overall diversity more than for CNEs. Simply dividing observed π by B, would therefore give an underestimate of π_0 as does not incorporate the reduction in diversity caused by SSWs at linked elements. When analysing patterns of diversity around protein-coding exons, we assumed pi_0 values of 0.00955 and 0.00895 when using the *castaneus* and Cox maps, respectively. When analysing patterns of diversity around CNEs, we assumed pi_0 values of 0.0102 and 0.00915 when using the *castaneus* and Cox maps, respectively.

Parameters of selective sweep obtained from patterns of nucleotide diversity

When fitting a model of SSWs and BGS to the reductions in diversity around exons and CNEs, we found that a two classes of advantageous mutational effects typically gave a substantially better fit to the data than did a single class or an exponential distribution (as judged by AIC). This result held regardless of the recombination map assumed and whether or not gene conversion was included in the analysis. The only exception was that an exponential distribution gave the best fit when protein-coding exons were analysed using the *castaneus* map and the high gene conversion rate was assumed, but in this case the difference in fit was fairly small (Table X).

For both protein-coding exons and CNEs we estimated a class of strongly advantageous mutations

and a class of more weakly selected mutations (Table 1). When assuming the *castaneus* map and the low gene conversion rate, we estimated advantageous mutations with scaled effects (γ_a) of 8,470 and 432 for protein-coding exons and CNEs, respectively. The frequencies of these mutations were substantially different between the site classes For both classes of sites, we inferred a class of more frequent, more mildly advantageous mutations as well.

inferred positively selected mutations with large fitness. We estimated stpositively selected mutations with γ_a = 8,470 and

Ignoring the contribution of BGS by setting B to 1.0 when fitting Equation 1 to the diversity troughs resulted in a much poorer model fit (Table 0.3). The strength of selection acting on advantageous mutations estimated when ignoring BGS is far greater than when BGS is accounted for. This is presumable because selection strength of selection acting on advantageous mutations was estimated to be higher and the frequency lower when BGS was ignored, for both classes of elements, for both recombination maps (Table 0.3).

In the absence of BGS effects, the strength of advantageous mutations required to explain the observed data is more far higher (Table 0.3), consistant with Campos et al. (2017). Furthermore, ignoring the contribution of BGS by setting B to 1.0 when fitting Equation 1 to the diversity troughs resulted in a much poorer model fit.

The gene conversion parameters of Paigen *et al* did not substantially influence the analysis.

Estimating selection parameters from the uSFS of simulated data

Parameters of the DFE can be estimated directly from polymorphism data if selected mutations are segregating in populations of interest (*Reviewed in* Eyre-Walker and Keightley 2007). It has been repeatedly demonstrated that parameters of the DFE for deleterious mutations (dDFE) can be accurately

Table 1: Parameters of positive selection in *M. m. castaneus* estimated by fitting a model of selective sweeps to troughs in diversity around functional elements. Parameters were obtained assuming the gene conversion parameters estimated by Paigen et al. (2008). Standard errors are shown below in square brackets below point estimates

Element	$\gamma_{a,1}$	$p_{a,1}$	$\gamma_{a,2}$	$p_{a,2}$	
Protein-Coding Exons	8,470 [672]	2.22 x 10^{-5} [2.21 x 10^{-6}]	22.3 [3.39]	0.0202 [4.38x 10^{-3}]	*castaneus* map
Conserved Non-Coding Elements	432 [21.2]	1.12 x 10^{-3} [8.86 x 10^{-5}]	14.5 [3.17]	0.0298 [8.22 x 10^{-3}]	
Protein-Coding Exons	4,100 [640]	2.45 x 10^{-5} [5.56 x 10^{-6}]	117 [49.9]	6.16 x 10^{-4} [2.78 x 10^{-4}]	*Cox* map
Conserved Non-Coding Elements	357 [46.8]	4.77 x 10^{-4} [9.37 x 10^{-5}]	5.95 [3.533]	0.0454 [0.0358]	

estimated from the SFS (Boyko et al., 2008; Keightley and Eyre-Walker, 2007; Kousathanas and Keightley, 2013; Tataru et al., 2017). It has also been shown that the parameters of advantageous mutations can also be estimated from the uSFS (Schneider et al., 2011; Tataru et al., 2017), but it has been argued that strongly selected advantageous mutations, which may contribute little to standing variation, will be undetectable by such methods (Campos et al. 2017; Chapter 3). In this study, we confirm this verbal argument using simulations, showing that accurate estimation of positive selection parameters does indeed depend on the strength and relative frequencies of advantageous mutations. We used forward-in-time simulations that incorporated linkage, because selection at linked sites can distort the uSFS in ways that likely affect real data and thus cannot be ignored. For each set of advantageous mutation parameters, we simulated 10Mbp of gene-like sequences giving a total of 7.5Mbp of nonsynonymous sites and 2.5Mbp of synonymous sites which we used to construct the uSFS for 20 haploid individuals. This sample size and quantity of data is fairly typical of population genomic datasets (e.g. Williamson et al. 2014; Kousathanas et al. 2014; Laenen et al. 2018).

Across simulations, the strength of selection differed (ranging between $\gamma_a = 10$ and $\gamma_a = 800$), but the product $\gamma_a p_a$, which is expected to be directly proportional to the rate of sweeps, was always equal to 0.1. All simulations were subject to the same dDFE, so the extent of BGS should be fairly similar. We found that selection at linked sites reduced synonymous site diversity below the expectation value

Table 2: Positive selection parameter estimates obtained by analysis of the uSFS for simulated poplations.

Divergence [a]	γ_a		p_a		$\gamma_a p_a$	Prop. Significant [b]
	Simulated	*Estimated*	*Simulated*	*Estimated*		
+	10	11.2 [5.60 - 20.0]	0.010000	0.00856 [0.00440 - 0.0199]	0.0954 [0.0838 - 0.115]	1.00
-		3.97 [1.13 - 27.2]		0.0201 [0.00472 - 0.0706]	0.0828 [0.0616 - 0.155]	1.00
+	20	16.6 [9.20 - 37.4]	0.005000	0.00568 [0.00241 - 0.0107]	0.0949 [0.0822 - 0.108]	1.00
-		19.9 [2.90 - 37.4]		0.00532 [0.00289 - 0.0207]	0.106 [0.0454 - 0.193]	0.97
+	50	37.4 [21.6 - 41.8]	0.002000	0.00257 [0.00202 - 0.00467]	0.0951 [0.0809 - 0.106]	1.00
-		37.3[1.87 - 65.5]		0.00266 [0.00125 - 0.0146]	0.0717 [0.0112 - 0.145]	0.86
+	100	37.43 [37.4 - 1530]	0.001000	0.00249 [0.0000738 - 0.00283]	0.0938 [0.0795 - 0.107]	1.00
-		0.323 [0.0371 - 1.25]		0.00259 [0.000525 - 0.0941]	0.00102 [0.0000620 - 0.0137]	0.00
+	200	37.4 [37.4 - 1,700]	0.000500	0.00251 [0.000220 - 0.00283]	0.0947 [0.0738 - 0.106]	1.00
-		0.272 [0.00546 - 1.911]		0.0122 [0.000690 - 0.138]	0.00310 [0.000104 - 0.0294]	0.07
+	400	37.4 [32.7 - 37.4]	0.000250	0.00245 [0.00199 - 0.00283]	0.0919 [0.0776 - 0.102]	1.00
-		12.3 [0.287 - 66.6]		0.00212 [0.000783 - 0.0104]	0.0338 [0.000250 - 0.0984]	0.22
+	800	37.4 [32.9 - 37.4]	0.000125	0.00222 [0.00186 - 0.00264]	0.0831 [0.0701 - 0.0936]	1.00
-		1.75 [0.111 - 43.0]		0.00240 [0.000343 - 0.0293]	0.0134 [0.0000515 - 0.0649]	0.12

[a] +/- indicates whether or not divergence was included when analysing the uSFS

[b] The proportion of bootstrap replicates where a full DFE gave a significantly better fit than a model containing just deleterious mutations

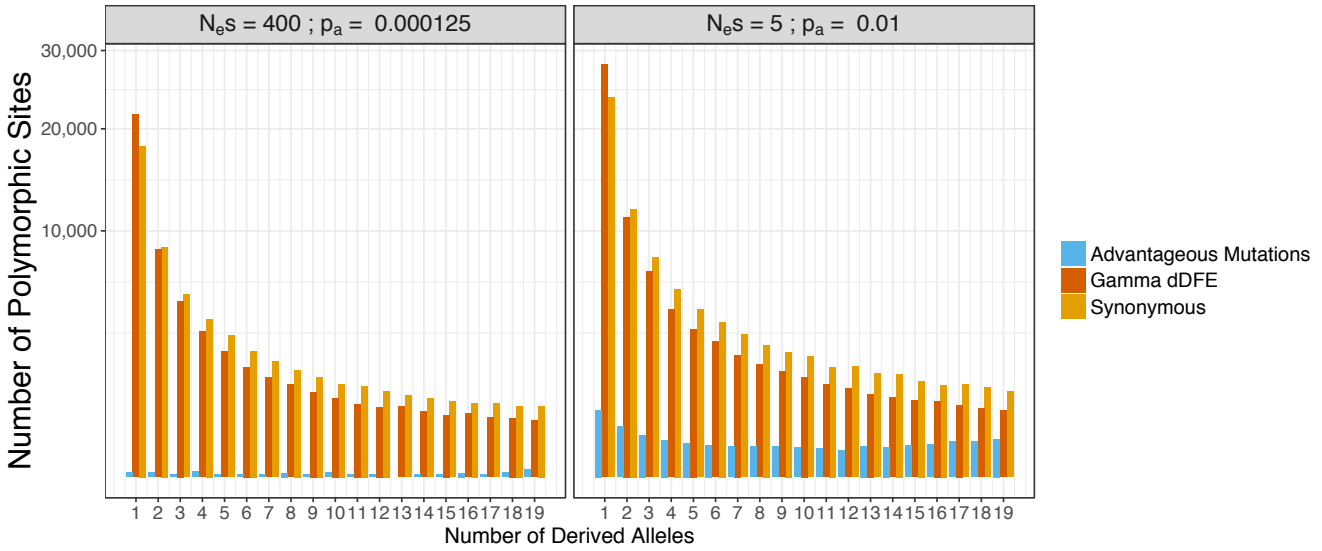

Figure 3: An example of the uSFSs for deleterious (Gamma dDFE) and advantageous nonsynonymous sites with neutral synonymous sites from simulated populations. Results shown are from simulations modelling strongly or weakly advantageous mutations. The uSFS model the same rate of synonymous substitutions $\gamma_a p_a = 0.1$. Simulated datasets included 10Mbp of exonic sites (3:1 nonsynonymous:synonymous sites).

of 0.01 in all simulations (Table 0.3), but as the strength of selection acting on advantageous mutations increased, diversity at linked sites decreased (reflected in in the decreasing values π/π_0 shown in Table 0.3). As expected, the relative fixation rate of nonsynonymous mutations (meausured using dN/dS) did not vary systematically across simulations (Table 0.3). From a visual inspection of the frequency spectra, it is clear that there is little information in standing variation for inferring the positive selection parameters when selection is strong (Figure 3).

We analysed the uSFS from our simulated populations and found that when advantageous mutations are relatively frequent ($p_a > 0.0005$), but weakly selected ($\gamma_a < 100$), both γ_a and p_a parameters can be estimated with precision (Table 1). However, we found that when advantageous mutations were infrequent but strongly selected ($\gamma_a \geq 100$ and $p_a \leq 0.0005$) the parameters were very poorly estimated. Across all simulated datasets, when we included divergence in the analysis, the product $\gamma_a p_a$ was accurately estimated (Table REF) and likelihood ratio tests never failed to detect the presence of advantageous mutations in the uSFS. When we excluded divergence from the analysis, however, the product $\gamma_a p_a$ was

poorly estimated when $\gamma_a \geq 100$ and likelihood ratio tests typically failed to detect positive selection (Table REF). Across all simulations, we found that polyDFE gave estimates of the dDFE which were highly accurate (Table 0.3). polyDFE performed most poorly when divergence was included in the analysis, but only a dDFE was inferred. These results replicate the findings of Tataru et al. (2017) and further emphasize the importance of specifying a full DFE model when making inferences of selection from the uSFS.

Discussion

BGS and SSWs both operate in the mouse genome. By fitting a model combining the effects of both processes to the troughs in diversity around protein-coding exons and CNEs in *M. m. castaneus*, we estimated parameters of positively selected mutations occurring in the two classes of element that predict the observed patterns. This relied on estimates of the reduction in diversity caused by BGS (B), which, in turn, relied on estimates of the DFE for harmful variants. The dDFE can be accurately estimated from the uSFS as long as the dDFE is estimated using a model that includes advantageous mutations (Table 0.3) (Tataru et al., 2017).

Here we have estimated parameters of positive selection from patterns of genetic diversity in the house mouse genome. We found that protein-coding regions experience more strongly selected mutation than do conserved non-coding elements. For both classes of sites, we found statistical support for two classes of advantageous mutational effects, one strong ($\gamma_a > 200$) and one comparatively weak. Using simulations, we demonstrated that the parameter estimates that we obtained are likely out of the range detectable by analysis of the uSFS.

The relative contribution of adaptive substitutions in protein-coding and regulatory regions to fitness change in mice

An enduring goal of evolutionary biology has been to understand the extent to which protein-coding and regulatory regions of the genome contribute to phenotypic evolution (King and Wilson, 1975; Carroll, 2005). King and Wilson (1975) posited that, since identity between human and chimpanzee proteins is around 99%, changes in gene regulation may explain the plethora of phenotypic differences between humans and chimps. Using a simple model of the fitness change brought about by the substitution of advantageous mutations, we can use the parameter estimates we obtained in this study to try and understand the contribution of protein-coding and regulatory regions of the genome make to phenotypic evolution.

Consider the following model of the fitness change brought about by the fixation of advantageous mutations (ΔW). New mutations occur at a particular class of sites with rate μ per base-pair, per generation. A proportion of these new mutations, p_a, are advantageous with an expected selection coefficient of s_a. The advantageous mutations fix with probability $u(s_a)$ and once fixed contribute s_a to the change in fitness. If it is assumed that selection is strong relative to genetic drift, then $u(s_a)$ is approximately s_a, giving the following expression:

$$\Delta W \propto \mu p_a n_a E(s_a)^2, \tag{5}$$

We parametrized Equation 5 using the estimates of selection we estimated in this study, summing the fitness contribution for the two classes of fitness effect. We assume that the average point mutation rate is the same for CNEs and protein-coding exons, so we ignore μ in Equation 5.

Based on our parameter estimates, we find that adaptation is far more frequent in CNEs than it is in protein-coding exons, but that protein-coding regions contribute more to fitness change. We inferred that

Table 3: Estimates of the changes in fitness caused by new mutations occurring in protein-coding exons and CNEs. and a per base-pair per generation mutation rate of 5.4 x 10^{-9} (Uchimura *et al.* 2015).

	μ_a	n_a (x 10^6)	$s^2_{a,1}$	ΔW x (10^{-12})
Exons	6.70 x 10^{-14}	24.0	1.39 x 10^{-4}	224
CNEs	1.23 x 10^{-11}	54.2	7.36 x 10^{-9}	4.91

the proportion of mutations occurring in CNEs that are advantageous is more than an order of magnitude higher than for protein-coding exons. Because of this and the fact that there are about three times as many CNE bases in the genome as there are protein-coding sites, the rate of advantageous mutations occurring in CNEs likely exceeds the rate in protein-coding regions. However, the average strength of selection acting on a new advantageous mutation in a protein-coding exon far exceeds that of a mutation in a CNE (Table 1) and since the change in adaptive fitness is dependant on the square of the selection coefficient (it is related to the additive genetic variance in fitness), the change in population mean fitness brought about by the fixation of advantageous mutations is higher for protein-coding exons than for CNEs (Table 3. The difference we found was small (around 3 times higher for protein-coding regions than for regulatory regions), but was robust to the choice of recombination map (Table REF).

There are a number of factors that should, perhaps, temper these conclusions. Firstly, the selection coefficient that appears in Equation 5 is the expectation of the DFE for advantageous mutations. If there were a continuous distribution of s values around the point estimates we obtained, integrating over this distribution may yield a different result. Secondly, we have assumed that all elements of a particular class share a common set of selection parameters. This is slightly problematic since there are a number of sub-categorisations that could be applied to the set of CNEs we analysed (e.g. promoters and enhancers may be subject to different selective pressures). Indeed, different categories of protein-coding genes may also be subject to different selection pressures. For instance, virus interacting proteins and highly expressed genes have been estimated to have higher rates of adaptive substitutions in different organisms (Enard eLife paper; Williamson).

Whether or not the conclusions we have drawn in this study can be generalised to other organisms remains to be seen, but the brown rat, *Rattus norvegicus*, provides a compelling first case for comparison. In *R. norvegicus* there are troughs in nucleotide diversity around protein-coding exons and CNEs that are very similar to those observed in *M. m. castaneus* (Deinum et al., 2015). Since broad-scale recombination rates are similar in mice and rats (Jensen-Seaman et al., 2004), qualitatively similar conclusions regarding the contribution of protein-coding versus regulatory change to adaptive evolution may be reached when analysing patterns of genetic diversity in rats.

0.1 Analysis of the uSFS

By analysing the uSFS of simulated populations, polyDFE yielded accurate estimates of the dDFE from simulated data, even when positive selection was very strong. Consistent with Tataru et al. (2017), we found that if advantageous mutations are present, but unaccounted for, estimates of the dDFE become inaccurate.

The number of fixed, advantageous mutations carries information on the compound parameter $\gamma_a p_a \mu$ (Kimura and Ohta 1971), which will be embedded within between species divergence at selected sites. Without further information from polymorphism data, this compound parameter cannot be disentangled by analysis of the uSFS. Across our simulations, the rate of sweeps did not vary, but nucleotide diversity at neutral, synonymous sites did; as the scaled strength of selection increased, synonymous site diversity decreased (Table 0.3). This all suggests that when advantageous mutations are strongly selected, but sufficiently rare, patterns of nucleotide diversity carry information that is not present in the unfolded site frequency spectrum.

Estimating parameters of positive selection from the uSFS versus patterns of diversity

To our knowledge, there are currently no methods that estimate the DFE using the site frequency spectrum expected under either BGS or SSWs. Rather, nuisance parameters or demographic models are used to account for the contribution of selection at linked sites to the shape of the SFS while assuming that selected mutations also shape the SFS. However, we have shown that advantageous mutations occurring in *M. m. castaneus* may be far stronger and infrequent than those that can reliably be detected by analysis of the uSFS. Interestingly, when we fit a bimodal DFE for advantageous mutations to the pattern of diversity around CNEs, one of the modes we inferred very closely matched the selection parameters we obtained by analysis of the uSFS in a previous study (Booker and Keightley BioRXiv).

there is potentially information present in the uSFS that may be useful for estimating the fitness effects of new mutations. Approximations for the uSFS expected under both BGS and SSWs have been developed (REFS), so a potential avenue for further research would be to incorporate these for making inferences from population genetic data.

In an earlier study, Teschke et al. (2008) analysed patterns of variation at microsatellite loci across the *M. m. domesticus* genome. In their study they estimated that SSWs driven by mutations with a selection coefficient of $s \approx 0.008$ occur at least every hundredth generation. If we assume an N_e of 420,000, we estimate that SSWs in protein-coding exons are driven by mutations with $s \approx 0.0099$ and in CNEs $s \approx 0.00027$.

We assumed that all new advantageous mutations are semi-dominant, which is something of a problem. Haldane's sieve predicts that most advantageous mutations that become fixed are dominant. There are a number of examples of SSWs being driven by recessive mutations in mammals, particularly humans (REFS). If advantageous mutations are fully recessive, where the dominance coefficient (h) is 0, the chance of stochastic loss exceeds that of mutations that have $h > 0$. As long as mutations are neither

fully recessive nor fully dominant ($0 < h < 1$), the troughs in diversity resulting from mutations with the compound parameter *2hs* are similar (Greg Ewing paper). Because of this, as long new mutations are neither fully recessive nor dominant, the selection coefficients we estimated should be directly proportional to the true values

0.2 Limitations

In collating the patterns of genetic diversity around either CNEs or protein-coding exons across the entire genome, it is likely that we have lost some valuable information. In particular, we set the π_0 values when fitting Equation 1 using values that gave a reasonable fit to the data, but did not explicitly model the reduction in genetic diversity caused by SSWs at linked elements. An alternative approach would be to fit Equation 1 to genome-wide variation in nucleotide diversity, conditioning on the locations of functional elements and a genetic map. Elyashiv et al. (2016) performed such an analysis on polymorphism data from *D. melanogaster* using a model that condiditoned the effects of SSWs on the locations of recent substitutions and the effects of BGS on the locations of functional elements. However, applying their methods to mice, where there is little information in the patterns of mean diversity around putatively selected/neutral nucleotide substitutions Halligan et al. (2013), would likely result in spurious parameter estimates.

The model of SSWs that we used in this study is of so-called 'hard' (or 'classic') sweeps, whereas studies in both humans and *Drosophila* suggest that 'soft' sweeps are common (Garud and Petrov, 2016; Garud et al., 2015; Schrider and Kern, 2017). A 'soft' selective sweep differs from the model outlined in the Methods section of this paper in that multiple haplotypes reach fixation, this can occur if selection acts on standing genetic variation or if multiple copies of the selected mutation arise independently (REVIEW). Additionally, adaptation acting on quantitative traits subject to stabilising selection may generate partial sweeps, as abrupt changes in allele frequency at many loci can rapidly affect mean

phenotypes, withoutnecessarily causing fixations. Such partial sweeps may be common in humans (Field et al., 2016). In their paper, Elyashiv et al. (2016) thoroughly discussed how assuming a model of hard sweeps when either soft or partial sweeps are common would affect estimates of positive selection parameters obtained from patterns of diversity. Briefly, if soft sweeps were common, it would likely have caused us to overestimate the strength of selection. Whereas if partial sweeps were common then we would likely underestimate the positive selection parameters. It seems likely that adaptation does not fit any one category, rather different functional elements will be subject to a mixture of different types of sweep. In the case of a soft sweep from standing variation, for example, the effect on neutral diversity is related to the frequency with which the focal allele was segregating before the onset of selection. Since nonsynonymous variants are maintained at lower frequencies than variants within CNEs (Halligan et al., 2013), the effects of soft sweeps may differ between the two classes of sites.

0.3 Recombination rate issues

The *castaneus* map that we inferred previously (Booker et al., 2017b) was constructed using a model of crossovers that does not explicitly model gene conversion. However, gene conversion will affect LD and may be reflected in the estimates of *rho*/bp obtained using LDhelmet (see Comeron 2017 review for discussion of this).

The estimates based on the Cox map, on the other hand, are not subject to this issue, but are Estimates of the relative rate of gene conversion and crossing-over In this study, gene conversion made little to no difference to parameter estimation, but this depends on the gene conversion parameters assumed.

The LD-based methods to infer recombination rates assume that patterns in

We assumed the estiamtes obtained by Paigen et al. (2008) when performing our analyses, which

yielded little difference in the parameter estimates. Assuming the high rate of gene conversion estimated for humans (XXX) made a much larger difference to parameter estimates.

Conclusions

In this study we have shown that that strong positive selection explains the diversity dips around protein-coding exons and CNEs. Using simulations, we showed that the parameters of these mutations are out of the range detectable by the uSFS, thus reconciling the present results with those obtained in Chapter 3. Furthermore, the parameters we estimated suggested that mutations in protein-coding regions contributes more to phenotypic change than do regulatory mutations.

Increasing the impact of the paper

I am happy with the way this chapter has turned out and I think that I could pursue a publication with it. However, there are a number of things that I/we could do to turn this from a fairly good low impact paper to a higher impact PNAS-level paper. By broadening the focus of the analysis to include all the species for which Ben has generated polymorphism datasets, we would likely garner more interest. The crucial question at the end of this paper is whether CNEs or exons contribute more to adaptive evolution. With an analysis of multiple species, we could ask whether general trends emerge from the analysis.

We would need the following to perform the analyses:

- Polymorphism and recombination maps for *M. m. castaneus*, *M. m. musculus*, *M. m. domesticus*, *R. norvegicus* and *M. spretus. I think, Ben has done all of this.*
- CNEs defined using progressive cactus placental alignment perhaps, of the kind Rory obtained for

flycatchers. *Me or Ben could do this, or Rory, if he were interested in being involved and had the time*

- We would also need dDFEs for each of the species (sub-species) above, which would be used to obtain values of B. I would do this part.
- We could then use the framework I have used in this chapter to estimate the sweep parameters in multiple species. I would also do this part.

Acknowledgements

Thanks to Bret Payseur, Sally Otto, Nathaniel Sharp and the Otto labgroup at UBC for discussions. TRB is supported by an EASTBIO BBSRC studentship. This project has received funding from the ERC.

References

Barton, N. H. (2000). Genetic hitchhiking. *Philos Trans R Soc Lond B Biol Sci*, 355(1403):1553–62.

Booker, T. R., Jackson, B. C., and Keightley, P. D. (2017a). Detecting positive selection in the genome. *BMC Biol*, 15(1):98.

Booker, T. R., Ness, R. W., and Keightley, P. D. (2017b). The recombination landscape in wild house mice inferred using population genomic data. *Genetics*, 207(1):297–309.

Boyko, A. R., Williamson, S. H., Indap, A. R., Degenhardt, J. D., Hernandez, R. D., Lohmueller, K. E., Adams, M. D., Schmidt, S., Sninsky, J. J., Sunyaev, S. R., White, T. J., Nielsen, R., Clark, A. G., and Bustamante, C. D. (2008). Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet*, 4(5):e1000083.

Braverman, J. M., Hudson, R. R., Kaplan, N. L., Langley, C. H., and Stephan, W. (1995). The hitchhiking effect on the site frequency spectrum of dna polymorphisms. *Genetics*, 140:783–796.

Campos, J. L., Zhao, L., and 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, S. B. (2005). Evolution at two levels: on genes and form. *PLoS Biol*, 3(7):e245.

Comeron, J. (2014). Background selection as a baseline for nucleotide variation across the drosophila genome. *PLoS Genetics*, 10(6).

Coop, G. and Ralph, P. (2012). Patterns of neutral diversity under general models of selective sweeps. *Genetics*, 192(1):205–24.

Cox, A., Ackert-Bicknell, C. L., Dumont, B. L., Ding, Y., Bell, J. T., Brockmann, G. A., Wergedal, J. E., Bult, C., Paigen, B., Flint, J., Tsaih, S. W., Churchill, G. A., and Broman, K. W. (2009). A new standard genetic map for the laboratory mouse. *Genetics*, 182(4):1335–44.

Cutter, A. D. and Payseur, B. A. (2013). Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet*, 14(4):262–74.

Deinum, E. E., Halligan, D. L., Ness, R. W., Zhang, Y. H., Cong, L., Zhang, J. X., and Keightley, P. D. (2015). Recent evolution in rattus norvegicus is shaped by declining effective population size. *Mol Biol Evol*, 32(10):2547–58.

Elyashiv, E., Sattath, S., Hu, T. T., Strutsovsky, A., McVicker, G., Andolfatto, P., Coop, G., and Sella, G. (2016). A genomic map of the effects of linked selection in drosophila. *PLoS Genet*, 12(8):e1006130.

Eyre-Walker, A. and Keightley, P. D. (2007). The distribution of fitness effects of new mutations. *Nat Rev Genet*, 8(8):610–8.

Field, Y., Boyle, E. A., Telis, N., Gao, Z., Gaulton, K. J., Golan, D., Yengo, L., Rocheleau, G., Froguel, P., McCarthy, M. I., and Pritchard, J. K. (2016). Detection of human adaptation during the past 2000 years. *Science*, 354(6313):760–764.

Galtier, N. (2016). Adaptive protein evolution in animals and the effective eopulation size hypothesis. *PLoS Genet*, 12(1):e1005774.

Garud, N. R., Messer, P. W., Buzbas, E. O., and Petrov, D. A. (2015). Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps. *PLoS Genet*, 11(2):e1005004.

Garud, N. R. and Petrov, D. A. (2016). Elevated linkage disequilibrium and signatures of soft sweeps are common in drosophila melanogaster. *Genetics*, 203(2):863–80.

Halligan, D. L., Kousathanas, A., Ness, R. W., Harr, B., Eory, L., Keane, T. M., Adams, D. J., and Keightley, P. D. (2013). Contributions of protein-coding and regulatory change to adaptive molecular evolution in murid rodents. *PLoS Genet*, 9(12):e1003995.

Hermisson, J. and Pennings, P. S. (2005). Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *Genetics*, 169(4):2335–52.

Hudson, R. R. and Kaplan, N. L. (1995). Deleterious background selection with recombination. *Genetics*, 141:1605–1617.

Jensen-Seaman, M. I., Furey, T. S., Payseur, B. A., Lu, Y., Roskin, K. M., Chen, C. F., Thomas, M. A., Haussler, D., and Jacob, H. I. (2004). Comparative recombination rates in the rat, mouse and human genomes. *Genome Res*, 14:528–538.

Keightley, P. D., Campos, J. L., Booker, T. R., and Charlesworth, B. (2016). Inferring the frequency spectrum of derived variants to quantify adaptive molecular evolution in protein-coding genes of drosophila melanogaster. *Genetics*, 203(2):975–84.

Keightley, P. D. and 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(4):2251–61.

Kern, A. D. and Hahn, M. W. (2018). The neutral theory in light of natural selection. *Mol Biol Evol*, 35(6):1366–1371.

King, M.-C. and Wilson, A. C. (1975). Evolution at two levels in humans and chimpanzees. *Science*, 188(4184):107–116.

Kousathanas, A., Halligan, D. L., and Keightley, P. D. (2014). Faster-x adaptive protein evolution in house mice. *Genetics*, 196(4):1131–43.

Kousathanas, A. and Keightley, P. D. (2013). A comparison of models to infer the distribution of fitness effects of new mutations. *Genetics*, 193(4):1197–208.

Laenen, B., Tedder, A., Nowak, M. D., Torang, P., Wunder, J., Wotzel, S., Steige, K. A., Kourmpetis, Y., Odong, T., Drouzas, A. D., Bink, M., Agren, J., Coupland, G., and Slotte, T. (2018). Demography and mating system shape the genome-wide impact of purifying selection in arabis alpina. *Proc Natl Acad Sci U S A*, 115(4):816–821.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. *Genetical Research*, 23:23–25.

McVicker, G., Gordon, D., Davis, C., and Green, P. (2009). Widespread genomic signatures of natural selection in hominid evolution. *PLoS Genet*, 5(5):e1000471.

Messer, P. W. (2013). Slim: simulating evolution with selection and linkage. *Genetics*, 194(4):1037–9.

Nordborg, M., Charlesworth, B., and Charlesworth, D. (1996). The effect of recombination on background selection. *Genetical Research*, 67:159–174.

Paigen, K. and Petkov, P. (2010). Mammalian recombination hot spots: properties, control and evolution. *Nat Rev Genet*, 11(3):221–33.

Paigen, K., Szatkiewicz, J. P., Sawyer, K., Leahy, N., Parvanov, E. D., Ng, S. H., Graber, J. H., Broman, K. W., and Petkov, P. M. (2008). The recombinational anatomy of a mouse chromosome. *PLoS Genet*, 4(7):e1000119.

Sattath, S., Elyashiv, E., Kolodny, O., Rinott, Y., and Sella, G. (2011). Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in drosophila simulans. *PLoS Genet*, 7(2):e1001302.

Schneider, A., Charlesworth, B., Eyre-Walker, A., and Keightley, P. D. (2011). A method for inferring the rate of occurrence and fitness effects of advantageous mutations. *Genetics*, 189(4):1427–37.

Schrider, D. R. and Kern, A. D. (2017). Soft sweeps are the dominant mode of adaptation in the human genome. *Mol Biol Evol*.

Stephan, W. (2010). Genetic hitchhiking versus background selection: the controversy and its implications. *Philos Trans R Soc Lond B Biol Sci*, 365(1544):1245–53.

Tataru, P., Mollion, M., Glemin, S., and Bataillon, T. (2017). Inference of distribution of fitness effects and proportion of adaptive substitutions from polymorphism data. *Genetics*, 207(3):1103–1119.

Teschke, M., Mukabayire, O., Wiehe, T., and Tautz, D. (2008). Identification of selective sweeps in closely related populations of the house mouse based on microsatellite scans. *Genetics*, 180:1537–1545.

Uchimura, A., Higuchi, M., Minakuchi, Y., Ohno, M., Toyoda, A., Fujiyama, A., Miura, I., Wakana, S., Nishino, J., and Yagi, T. (2015). Germline mutation rates and the long-term phenotypic effects of mutation accumulation in wild-type laboratory mice and mutator mice. *Genome Res*, 25(8):1125–34.

Wiehe, T. and Stephan, W. (1993). Analysis of a genetic hitchhiking model, and its application to dna polymorphism data from drosophila melanogaster. *Mol Biol Evol*, 10(4):842–854.

Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M., and Wright,

S. I. (2014). Evidence for widespread positive and negative selection in coding and conserved noncoding regions of capsella grandiflora. *PLoS Genetics*, 10(9):e1004622.

Table S1: Summary statistics for simulated populations

γ_a	p_a	π/π_0	dS	dN	dN/dS
10	0.010000	0.940	0.00533	0.00159	0.299
20	0.005000	0.914	0.00526	0.00157	0.299
50	0.002000	0.880	0.00527	0.00159	0.302
100	0.001000	0.862	0.00525	0.00162	0.309
200	0.000500	0.844	0.00531	0.00159	0.300
400	0.000250	0.819	0.00523	0.00155	0.296
800	0.000125	0.795	0.00527	0.00156	0.295

Map	Element	$s^2_{a,1}$	$s^2_{a,2}$	Sites (Mbp)	$W_{a,1}$	$W_{a,2}$	W_a
castaneus	Exon	9.87×10^{-5}	6.87×10^{-10}	24.0	2.84×10^{-10}	1.79×10^{-12}	2.86×10^{-10}
	CNE	2.56×10^{-7}	2.90×10^{-10}	54.2	8.44×10^{-11}	2.53×10^{-12}	8.69×10^{-11}
Cox	Exon	2.31×10^{-5}	1.89×10^{-8}	24.0	7.34×10^{-11}	1.51×10^{-12}	7.50×10^{-11}
	CNE	1.76×10^{-7}	4.88×10^{-22}	54.2	2.45×10^{-11}	6.49×10^{-13}	2.51×10^{-11}

Table S2: Parameters of the distribution of fitness effects for harmful mutations obtained by analysis of the uSFS. Simulated values were $\beta = 0.20$ and $\hat{\gamma_d} = -2000$

γ_a	p_a	Divergence[a]	Full DFE[b]	β[c]	$\hat{\gamma_d}$ [d]
10	0.01000	+	+	0.203 [0.190 - 0.231]	-865 [-1120 - -561]
		+	-	0.135 [0.127 - 0.140]	-6860 [-10100 - -4850]
		-	+	0.217 [0.190 - 0.270]	-755 [-110000 - -483]
		-	-	0.175 [0.166 - 0.184]	-1550 [-2100 - -1180]
20	0.00500	+	+	0.199 [0.184 - 0.212]	-974 [-1390 - -744]
		+	-	0.132 [0.125 - 0.142]	-8480 [-13200 - -5030]
		-	+	0.199 [0.187 - 0.226]	-9831 [-1330 - -676]
		-	-	0.176 [0.168 - 0.183]	-1620 [-2040 - -1230]
50	0.00200	+	+	0.199 [0.179 - 0.210]	-979 [-1680 - -740]
		+	-	0.136 [0.130 - 0.144]	-7260 [-11100 - -4930]
		-	+	0.199 [0.187 - 0.215]	-944 [-1350 - -739]
		-	-	0.186 [0.177 - 0.195]	-1220 [-1640 - -986]
100	0.00100	+	+	0.195 [0.175 - 0.210]	-952 [-1780 - -661]
		+	-	0.137 [0.129 - 0.144]	-5980 [-9350 - -4140]
		-	+	0.193 [0.184 - 0.271]	-953 [-1270 - -637]
		-	-	0.189 [0.182 - 0.199]	-1040 [-1310 - -790]
200	0.00050	+	+	0.197 [0.174 - 0.210]	-1040 [-2060 - -748]
		+	-	0.136 [0.130 - 0.144]	-7470 [-10700 - -5100]
		-	+	0.207 [0.187 - 0.353]	-927 [-1320 - -498]
		-	-	0.190 [0.183 - 0.199]	-1160 [-1470 - -917]
400	0.00025	+	+	0.209 [0.192 - 0.224]	-745 [-1180 - -558]
		+	-	0.148 [0.141 - 0.156]	-4010 [-5910 - -2810]
		-	+	0.210 [0.199 - 0.229]	-727 [-939 - -541]
		-	-	0.202 [0.193 - 0.212]	-840 [-1040 - -660]
800	0.0001	+	+	0.210 [0.181 - 0.218]	-798 [-1500 - -592]
		+	-	0.148 [0.139 - 0.157]	-3890 [-6000 - -2720]
		-	+	0.205 [0.193 - 0.236]	-804 [-1020 - -543]
		-	-	0.198 [0.189 - 0.209]	-889 [-1130 - -693]

[a] +/- indicates whether or not divergence was included when analysing the uSFS

[b] +/- indicates whether or not advantageous mutation parameters were inferred

[c] The shape parameter of the gamma distribution of deleterious fitness effects

[d] Mean strength of selection of a new harmful mutation

Table S3: Comparison of advantageous mutational models. Dashes indicate the best-fitting model.

Map	GC	Model[a]	ΔAIC CNEs	ΔAIC Exons
castaneus	High	2	-	-
		e	-367	-284
		s	-85.6	-284
	None	2	-	-
		e	-262	-126.989377
		s	-87.3	-153
	Paigen	2	-	-
		e	-280	-134
		s	-113	-161
Cox	High	2	-	-3.91
		e	-177	-
		s	-92	-1.48
	No	2	-	-
		e	-150	-19.6
		s	-44	-38.9
	Paigen	2	-	-
		e	-147	-16.2
		s	-43.5	-32.1

[a] Denotes the model of advantageous mutations used. e - exponential, 2 - two classes of discrete effects and s - a single class of discrete effects.

Table S4: The effect of background selection (BGS) on estiamtes of positive selection parameters obtained by fitting the troughs in diversity. Values obtained assuming the *castaneus* map and the gene conversion parameters of Paigen et al. (2008) are shown. The difference in AIC between the full model and a model assuming that BGS does not contribute to the observed troughs

Element	BGS	ΔAIC	$\gamma_{a,1}$	$p_{a,1}$	$\gamma_{a,2}$	$p_{a,2}$
Exon	-	65.5	18,900 [1,510]	0.0000130 [0.000001]	62.8 [9.46]	0.00584 [0.00115]
	+	-	8,470 [672]	0.0000220 [0.000002]	22.3 [3.39]	0.02020 [0.00438]
CNE	-	109	1,460 [134]	0.000126 [0.000018]	68.2 [9.73]	0.00377 [0.000561]
	+	-	432 [21.2]	0.001120 [0.000089]	14.5 [3.17]	0.0298 [0.00822]

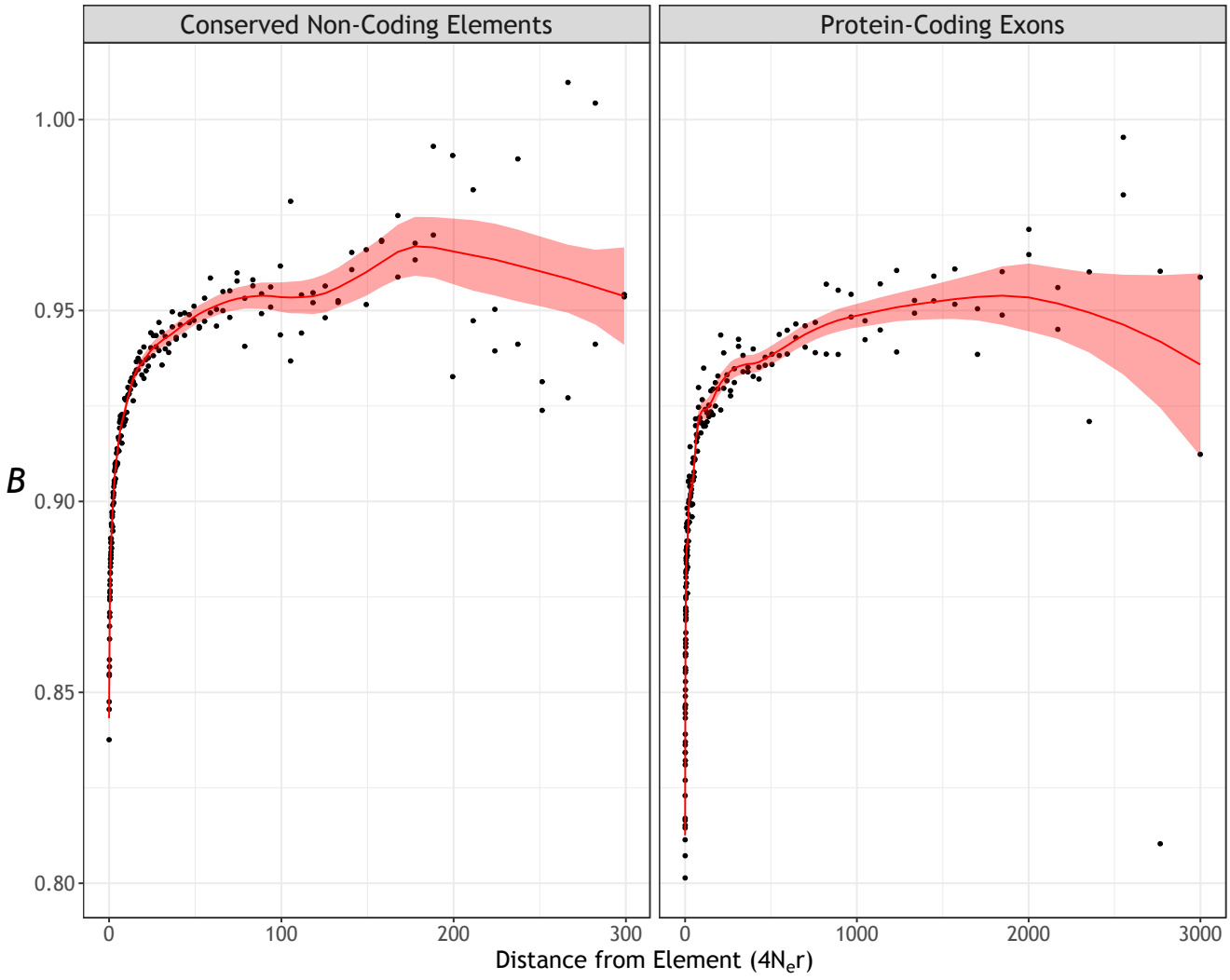

Figure S1: Reductions in diversity around protein-coding exons and CNEs in simulations. Red lines are Loess curves fitted to the data with a span of 0.2, the ribbons are an apprimate 95% confidence prediction intervals

.