Estimating the parameters of selective sweeps from patterns of genetic diversity in the house mouse genome

Tom R. Booker[1,*], Brian Charlesworth[1], and Peter D. Keightley[1]

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

[*]*t.r.booker@sms.ed.ac.uk*

May 7, 2018

Abstract

Introduction

In the past 30 years of population genetic research it has become clear that natural selection shapes patterns of nucleotide diversity across the genomes of many species (Corbett-Detig *et al* 2015; Cutter and Payseur 2012). Because genetically linked sites do not evolve independently, selection acting at one site may have consequences for 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 (REF DUMP). Two main modes of selection at linked sites have been identified; selective sweeps caused by the spread of advantageous mutations and background selection 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). However, the proportion of nonsynyonomous substitutions attributable to adaptive evolution (α) is typically high (50%)(Galtier; but see Booker et al 2017 for caveats), suggesting that selective sweeps may play a substantial role in shaping nucleotide diversity across the genomes of many species.

Selective sweeps have been subject to rigorous population genetic research (Maynard-Smith and Haigh, Hudson and Kaplan, Coop and Ralph, Hermisson and Pennings, Barton 2000). The classic footprint of a selective sweep is a trough in nucleotide diversity at neutral sites surrounding substitutions. Reductions in nucleotide diversity caused by selective sweeps 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 selective sweeps to estimate the frequency and strength of advantageous mutations in *Drosophila melanogaster* using the positive correlation between recombination rate and nucleotide diversity. At the time of their analysis, the theory of background selection was in its infancy and theory connecting background selection and sweeps had not been developed. However, the effects of background selection are expected to be ubiquitous across the genome (Comeron; McVicker, Comeron REVIEW, Elyashiv), and studies, conceptually similar to Wiehe and Stephan's (1993), have shown that controlling for background selection is highly important when parametrizing sweep models from patterns of nucleotide diversity (Campos et al 2017; Elyashiv et al 2015).

Because both selective sweeps and background selection act to reduce nucleotide diversity, it has proven difficult to distinguish their effects using population genetic data (Stephan review?). 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 selective sweeps, so they (Sattath et al. 2011) fitted a sweep model to the trough they observed and estimated that strongly selected mutations ($2N_e s \approx 5{,}000$) act 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). Indeed, there are reductions in average diversity rextending beyond the flanks of both protein-coding exons and conserved non-coding elements in mice (CNEs) (Halligan *et al* 2013). For both classes of elements, however, values of $\alpha \geq 0.19$ have been reported for both classes of elements (Halligan *et al* (2013) and background selection alone cannot fully explain the troughs in diversity (Halligan et al 2013; Booker and Keightley Unpublished), suggesting that selective sweeps do contribute to the observed patterns.

We recently estimated distributions of fitness effects (DFEs) for both harmful and advantageous mutations occurring in multiple classes of functional elements in house mice (Booker and Keightley *Unpublished*). We obtained our estimates of the DFE by analysing the distribution of derived allele frequencies,

referred to as the unfolded site frequency spectrum (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 showed that the parameters of the DFE we obtained were unable to explain the troughs in diversity around protein-coding exons, but were able to explain the troughs around conserved non-coding elements. A possible explanation for this is that advantageous mutations that occur in protein-coding regions have, on average, larger effects on fitness that those occurring in regulatory regions which may affect the power to It has been suggested that if advantageous mutations are strongly selected, they may go undetected by analysis of the uSFS since they may Estimates of the DFE that had we obtained by analysis of the uSFS were not able to explain the trough in diversity observed around protein-coding exons in mice, but were able to explain the reductions in diversity observed around conserved non-coding elements.

It has been suggested that such analysis methods may not be able to detect strongly selected advantageous mutations if they contribute little to polymorphism

sing our DFE estimates in simulations we showed that our estimates of the DFE obtained by analysis of the uSFS do not fully explain the diversity troughs around protein-coding exons in mice (Booker and Keightley Unpublished).

is to understand the fitness consequences of new mutations. There is undoubtedly a distribution of fitness effects (DFE) for new mutations, but this distribution may vary across the genome, for instance the average selective effect of a new mutation in a protein-coding gene may differ from a mutation occurring within a regulatory element (Peter's Review?). Experimental approaches for estimating the DFE are limited to organisms which can be maintained in lab populations, which typically excludes mammals. Statistical methods have been developed for estimating the DFE from population genetic data. Such methods rely on the assumption that selected alleles segregate within populations of interest, such that they affect the shape of the distribution of allele frequencies (a vector known as the site frequency spectrum). Since natural selection should push advantageous alleles to high frequencies and maintain deleterious alleles at low frequencies, analysis of the SFS for derived alleles (the *unfolded* SFS, henceforth uSFS) can be used to

The rate of advantageous substitutions is determined, largely, by the product of the effective strength of selection acting on advantageous mutations ($2N_e s_a$) and the proportion of new mutations that are advantageous (p_a) (Eq. 2.14 Kimura and Ohta 1972). This implies that any given rate of advantageous substitutions could be driven by infrequent strongly selected mutations, or frequent weakly

selected mutations. Data analyses reflect this; the product $2N_e s_a p_a$ can be accurately estimated by analysis of the uSFS, but the selection coefficient and frequency parameters can be difficult to tease apart (Schneider et al 2012; Tataru et al 2017). The McDonald-Kreitman test (MK 1992) and its extensions (Keightley and Eyre-Walker etc.), that are used to estimate α rely on the assumption that focal species and the outgroups used to estimate divergence are subject to the same distribution of fitness effects (DFE). Because this assumption may not be met in practice, Tataru *et al.* proposed estimating selection parameters from polymorphism alone, ignoring between-species divergence. Tataru *et al* (2017) recently developed a method for estimating selection parameters from the uSFS and demonstrated that accurate parameter estimates can be obtained when excluding between-species divergence (typically the final value in the uSFS). However, if advantageous mutations are strongly selected and infrequent, a scenario that Tataru *et al.* (2017) did not test in their study, then they may not contribute substantially to polymorphism and thus cannot be accurately estimated by analysis of the uSFS.

When analysis of the uSFS includes between-species divergence, estimates of Because of this, p_a and $2N_e s$ are difficult to estimate from the uSFS. This my especially be the case when $N_e s_a$ is large and p_a is small.

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

Materials and Methods

Model of Recurrent Sweeps and Background Selection

Background selection (BGS) is often modelled as the reduction in diversity experienced by a focal neutral site caused by deleterious mutations occurring at linked selected sites. An approximation for the

reduction in diversity caused by background selection:

$$B = \frac{N_e}{N_0} \approx exp\Bigg[-\sum_x \int_0^1 \frac{u_x f_x(t)\, dt}{t\Big(1 + \frac{(1-t)r_{i,j}}{t}\Big)^2} \Bigg] \tag{1}$$

Where the sum is over all linked selected sites, the integral is over the distribution of fitness effects for deleterious mutations, u_x is the deleterious mutation rate, t is the reduction in fitness for heterozygotes (assumed to be $\frac{s}{2}$), $r_{x,y}$ is the recombination distance between the focal neutral site and the selected site and $f_x(t)$ is the proportion of sites in the DFE with selection coefficient t.

Background selection (BGS) and selective sweeps (SSWs) are processes that induce coalescence at linked neutral sites. Provided that the effects of both processes are relatively weak as compared to drift, the effects of the two can simply be summed (Kim and Stephan 2001). Furthermore, if it is assumed that harmful mutations and is far more frequent than SSWs, BGS may reduce the population scaled mutation rate, and thus the rate of selective sweeps. Under the above assumptions, the combined effect of the two processes have on levels of genetic diversity can be modelled as the following:

Background selection (BGS) and selective sweeps (SSWs) are processes that induce coalescence. If we assume that the two are independent exponential processes, then the rates at which they induce coalescence can simply be summed (KIM AND STEPHAN 2000). While this assumption has been shown to hold reasonably well, in reality BGS may influence the effects of selective sweeps and *vice versa* The assumption that selective sweeps and background selection are independent has been made before (KIM AND STEPHAN 2000; CORBETT-DETIG et al. 2015; ELYASHIV et al. 2016; CAMPOS et al. 2017) but in reality, BGS may influence the fixation probabilities of new advantageous mutations (REF?).

The model we use here is an extension to the model used by CAMPOS et al. (2017) suggested by Charlesworth (unpublished).

$$\frac{\pi_j}{\pi_0} \approx \frac{1}{B_j^{-1} + B2N_eP_{sc,j}} \tag{2}$$

Where $\frac{\pi_j}{\pi_0}$ is the reduction in neutral genetic diversity at site j relative to the expectation in the absence of selection at linked sites. The differences between our model and that used by Campos et al (2017) is that B is in the second term in the denominator of Equation 1. B is the reduction in pairwise coalescence times due to the effects of background selection which is calculated using Equation 1. Multiplying the rate of sweep induced coalescence (P_{sj}) by B reflects the reduction in fixation probability of new mutations caused by background selection.

$$P_{sc,j} \approx V_a \tau \gamma_a^{\frac{-4r_{i,j}}{s}} \tag{3}$$

The term $V_a = 2\mu\ p_a\gamma_a$ is the rate of sweeps per generation, where μ is the per-base pair per generation mutation rate, p_a is the fraction of new mutations occurring within a focal element that are advantageous and γ_a is the scaled selection coefficient of a new mutation ($2N_e s_a$).

Assuming that all new advantageous mutations have the same selective effect is an assumption that is difficult to justify in light of experimental studies (REFS). There is evidence that the distribution of fitness effects for advantageous mutations is beneficial from both theoretical (Orr 2003; Griffiths REFS) and empirical studies (DATA Papers?). It is straightforward to incorporate a distribution of advantageous mutation effects to Equation 3

$$P_{sc,j} \approx \int_0^1 f_x(\gamma) V_a \tau \gamma_a^{\frac{-4r_{i,j}}{s}}\ d\gamma \tag{4}$$

When analysing the mouse data, see below, we compared the fit of Equation 2 incorporating either single class of beneficial mutations (Equation 3) or the exponential distribution (Equation 4) using Aikieke's Information Criterion (AIC) by assumig that the residuals about the curve are normally distributed.

We estimate γ_a and p_a by fitting the relationship between nucleotide diversity and distance to functional elements by non-linear least squares using the *lmfit* (0.9.7) package for Python 2.7.

Simulations

We generated simulated datasets using the forward-time simulation package SLiM (v1.8; Messer 2012) to assess how well $N_e s_a$ and p_a are estimated by analysis of the uSFS. Our simulations were based on those used by Kousathanas and Keightley (2012). We simulated chromosomes containing 20 'genes' spaced out evenly on a 1Mbp chromosome. Each 'gene' consisted of 10 100bp exons, separated by 1Kbp introns. Nonsynonymous mutations were modelled as 75% of mutations occurring in exons, the remainder were strictly neutral (i.e. synonymous sites). 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 individuals were simulated for $10N$ generations to establish equilibrium conditions, after which 20 haploid individuals were sampled every $2N$ generations for a further $100N$ generations. For a given set of positive selection parameters (see below), we performed 10 replicate simulations, giving a total of 10,000 simulated 'genes' with which to construct the uSFS. Bootstrap samples were generated by sampling 10,000 loci, with replacement, from all 10 simulation replicates, time points and physical positions.

Advnatgeous mutation parametrs were set in simulations to match the levels of diversity observed in Campos et al (2017) analysed patterns of neutral variation in the *Drosophila melanogaster* genome and estimated $N_e s_a p_a \approx 0.05$. We sought to establish whether uSFS analysis, which can accurately estimate p_a and $N_e s_a$ when advantageous mutations are relatively frequent, could accurately estimate positive selection parameters when advantageous mutations are rare. To this end, we modelled

Analysing simulation data using DFE-alpha and polyDFE

We estimate the strength and frequency of new mutations using the uSFS. We analyse the uSFS using either the method of Schneider et al. (2011) as implemented in DFE-alpha, or the methods of Tataru et al. (2017) as implemented in polyDFE. Both methods estimate the rate and strength of advantageous mutations using the unfolded site frequency spectrum (uSFS). However, the models implemented in the two differ in their underlying assumptions. The Schneider et al. (2011) approach builds upon the Wright-Fisher transition matrix methods developed by Keightley and Eyre-Walker (2007) to estimate the distribution of fitness effects for harmful mutations. The methods implemented by Tataru et al. (2017) build upon Sawyer and Hartls Poisson random field model. Throughout the rest of the paper, we will refer to these methods by the names of the programs in which they are implemented.

The methods of Schneider et al (2011) are implemented in the program DFE-alpha. We analyse the simulation data using DFE-alpha. Because the 3-epoch model takes a long time to converge, we performed 10 bootstraps per DFE estimate and take the mode of each parameter for the selected site analysis. Selective sweeps affect thte frequencies of linked alleles, distorting the uSFS in ways not necesaassrily captured under the demographic models implemented by DFE-alpha. Because of this, we correct the selected site uSFS prior to estimating selection parameters using the fit of the demographic model following Keightley *et al.* (2016) and Booker and Keightley (*Unpublished*)

DFE-alpha does not currently allow the user to estimate an exponential distribution of advantageous mutational effects, so when estimating the DFE for simulations modelling an exponential distribution of fitness effects for advantageous mutations, we used the program polyDFE (v1.0; Tataru et al. 2017). polyDFE implements the Poisson-random field methods of Tataru et al. 2017. Like DFE-alpha, polyDFE contrasts the uSFS for selected and putatively neutral sites in order to infer the full DFE. polyDFE does not explicitly model the population's demographic history, rather it uses the neutral site uSFS to obtain a set of nuisance parameters which encapsulate devations from a purely neutral model (e.g. demographic history and selection at linked sites).

When analysing data using polyDFE we used the following options: Model C, including between-species divergence. As with the DFE-alpha analysis, we analysed 1,000 replicate uSFSs from the simulation data.

Analysis of Mouse Data

Halligan et al. (2013) sequenced the genomes of 10 wild-caught *Mus musculus castaneus* individuals to high coverage using Illumina paired-end reads. We used the variants called in that study to obtain estimates nucleotide diversity.

From the edges of exons (CNEs), I extracted the SFS in windows of 1Kbp (100bp) extending to distances of 100 Kbp (5Kbp). All non-CpG sites in these windows were extracted, and mouse-rat divergence was calculated.

Analysis windows that are physically distant to protein-coding exons may be tightly linked genetically if the recombination rate is low and conversely, windows that are physically close may be genetically distant if the recombination rate is high. Recombination rates vary substantially across the mouse genome so incorporating this is likely very important Using either the LD-based map or the Cox-map I calculated the genetic distance between an analysis window and the centre of the focal element.

Recombination rate maps can be estimated by performing crosses and observing where recombination events occur or by analysing patterns of There are two types of genetic maps available for mice, those constructed by performing crosses and those inferred from patterns of linkage disequilibrium. The two maps will have benefits and drawbacks. Firstly, maps based on pedigree information are unbiased. They give a description of the locations and rates of crossing over events in the genome,. However, pedigree-based maps require a large number of individuals to be genotyped, which has meant that researchers have often been limited to using a relatively small number of genetic markers. Recombination maps based on linkage disequilibrium, on the other hand, use patterns of linkage disequilibrium to infer the population-scaled recombination rates ($4N_e r$) across the genome. LD-based approaches can provide inferences of recombination rates at very fine-scales across the genome, enabling researchers to locate recombination hotspots (reference to a review? necessary?). A drawback of LD-based approaches is that the recombination rate estimates they produce are confounded with the level of genetic diversity, since both are functions of the effective population size (N_e). In this study, we incorporate genetic distances using both pedigree-based and LD-based recombination maps constructed for *Mus musculus*. We use the

Cox et al. (2009) genetic map, which was constructed with 10,195 SNPs genotyped in 3,546 meioses.

Recombination proceeds via crossing-over or gene conversion, but the above formulae (Equations 3 and 4) assume that genetic distance is solely a product of the local crossingover rate and the phyiscal distance. We incorporated gene conversion into Equation 2 by setting $r_{i,j}$ in Equations 3 and 4 using Equation 1 from [**?**]

$$r_{i,j} = d_{i,j} r_c + g_c d_g \left(1 - e^{-\frac{d_{i,j}}{d_g}} \right) \tag{5}$$

where $d_{i,j}$ is the physical distance between a focal neutral site and a selected site, r_c is the rate of recombination by crossing-over, g_c is the rate of non-crossing over gene conversion and d_g is the mean length of a gene conversion tract. This assumes that the distribution of gene conversion tract lengths is exponential. We assumed a mean tract length of 144bp and that the gene conversion rate was 10.5% of the local crossing-over rate (Paigen *et al.* 2008).

We incorporated background selection as a covariate when fitting Equation 2 using the simulation results we obtained in an earlier study (Booker and Keightley *et al Unpublished*). The simulations performed by Booker and Keightley (2018) incorporated the actual distribution of functional elements observed in the mouse genome. They assumed recombination rates These simulations incorporated the actual distribution of functional elements that is in the *M. musculus* genome as well as recombination rate variation. Theoretical models of the effects of background selection perform poorly when deleterious mutations have weak ($\gamma_d < 5$) effects on fitness. In practice, researchers have truncated the DFE for harmful mutations. Using simulation results allows us to include the effects of BGS without the need to truncate the DFE. In the case of 0-folde sites in protein-coding exons, Booker *et al.* (*Unpublished*) found that $\approx 20\%$ of new mutations occuring at nonsynsymous sites had $\gamma_d < 1$. Truncating the distribution may make a substantial difference to the effects of BGS.

We assume the point mutation rate to be 5.4×10^{-9} (UCHIMURA et al. 2015). The mean length of a protein-coding exon is 151bp. The mean length of a conserved non-coding exon is 51bp.

Estimating the effects of background selection

We incorporated the effects of background selection into

Results

Estimating selection parameters from simulations

We performed simulations to assess how accurately positive selection parameters are estimated by analysis of the uSFS. We performed simulations that incorporated linkage, as selection at linked sites causes distortions to the uSFS that likely affect real data and thus cannot be ignored. From our simulations we found that when advantageous mutations are relatively weakly selected and infrequent ($\gamma_a < 100$ and $p_a > 0.0005$), polyDFE estimated both γ_a and p_a with precision (Table REF). However, we found that when selection on advantageous mutations is strong but infrequent ($\gamma_a \geq 100$ and $p_a \leq 0.0005$) the positive selection parameters were poorly estimated regardless of whether divergence was used in the inference. However, when divergence was included, the product $\gamma_a p_a$ was accurately estimated (Table REF).

Strongly selected In our simulations, we modelled cases where $gamma_a p_a = 0.05$. This value of $\gamma_a p_a$ is based on values estimated in studies of *Drosophila melanogaster*.

In simulations incorporating strongly advantageous mutations (i.e. $gamma > 100$), we found that positive selection could not be detected by analysis of polymorphism alone. Likelihood ratios between full DFE models and dDFE models fitted to the uSFS were not statistically significant when divergence was ignored (Table X). In such cases, advantageous mutations contribute little to standing variation, so there is little power to detect advantageous mutations from polymorphism alone.

Strongly selected, but rare advantageous mutations can be detected by analysis of the uSFS, if between-species divergence is included. data sets resulted in likelihood ratio tests, This is evidenced by significant likelihood ratio tests for the presence of advantageous mutations. However, in these cases, the parameter values obtained are

The number of fixed, advantageous mutations carries information on the compound parameter $\gamma_a p_a \mu$ (Kimura and Ohta 1971), but without further information from polymorphism data, this compound parameter cannot be disentangled by through analysis of the uSFS.

In such cases, advantageous mutations were detected when between-species did not substantially contribute to standing variation and so

Analysing simulated uSFSs with polyDFE yielded estimates of the dDFE that were extremely precise, but this depended on whether or not divergence was included or whether a full DFE was inferred. This is particularly evident for the case of $\gamma_a = 5$ and $p_a = 0.01$ (Table SX), where advantageous mutations contribute substantially to both standing variation and between-species divergence. In this case, by limiting the inference to just the dDFE, advantageous mutations that contribute to the shape of the uSFS are assumed to be deleterious, resulting in spurious dDFE inferences. In the simulations where γ_a = 400 and p_a 0.000125, advantageous mutations make little contribution to standing variation,

Adaptive evolution was fairly frequent in our simulations ($\alpha \approx 30\%$), ignoring the contribution of advantageous mutations to both standing variation and between-species divergence, either by modelling only the dDFE or by excluding divergence from calculations, led to biased parameter estimates, consistent with the findings of Tataru *et al* (2017).

Estimates of the parameters of the dDFE were recovered from simulated populations with high precision. Our simulations assumed a gamma dDFE, but also included of advantageous mutations. When divergence was included in the iner Ignoring the contribution of advantageous mutations to the uSFS (i.e. by estimating only the dDFE parameters) resulted in dDFE parameter estimates that were substantially biased (Table SX). This was particularly

The strength and frequency of new advantageous mutations can be estimated from both the uSFS and patterns of genetic diversity at linked sites.

In the case of weakly selected advantageous mutations, with effects of $\gamma_a = 20$, uSFS based inference methods outperform our method analysing patterns of diversity. This is presuably because under this selection regime, the fixation of advantageous mutaitons has little effect on patterns of genetic diversity. Furthermore, in such cases, if background selection is not corrected for, the parameters of selection inferred may be entirely spurious.

In our simulaitons, we modelled strongly selected mutations at different frequencies.

When analysing simulation data we used the actual DFE model assumed in the simulations. This is obviously not possible when analysing real data, where the true nature of the DFE is unknown. Estimates of the DFE for harmful mutations obtained using uSFS analysis methods can be biased if positive selection is present but not modelled

The fitness effects and

When advantageous mutaitons are frequent, the product γp_a is accurately estimated fromt the uSFS. However, the individual parameters are difficult to disentangle.

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

We compared the the relatoinship between nucleotide diversity and genetic distance to functional elements assuming two different genetic maps constraucted for *Mus musculus*. The first map was generated using linkage disequilibrium to infer the population-scaled recombination rate ($\rho = 4N_e r$) across the mouse genome. The second map, the Cox map, was constructed using

Recombination maps generated using LD be biased by selection at linked sites (REVIEW), so assuming the *castaneus* map when calculating genetic distances may exaggerate troughs in neutral diversity as regions of the genome that exhibit reduced diversity may have been inferred to exhibit reduced recombination rates.

We found that, in the immediate flanks of both exons and CNEs, levels of diversity were lower when assuming the LD-based *castaneus* map than when assuming the pedigree-based Cox map. Since LD-based recombination rate estimates can be biased by the effects of selection at linked sites (Review?), this difference could potentially be On the other hand, nucleotide diversity levels off at a higher level when assuming the *castneus* map, which suggests that the Cox map, which has limited resolution compared to the *castaneus* map, does not include many regions of the genome that have higher recombination rates.

The different reductions in diversity observed when assuming the different genetic maps will result in different parameters of selection. Because of this, we analysed both patterns.

Potentially consistent with this, genetic diversity in the immediate flanks of both exons and CNEs are lower when distances were calculated using the *castaneus* map than they were when the pedigree-based Cox map was used.

The patterns of nucleotide diversity obtained when were different depending on whether the LD-based or pedigree-based recombination maps were used.

Estimates of selection for *M. m. castaneus*

We estimated the parameters of a model of recurrent selective sweeps acting in two different classes of functional elements in *M. m. castaneus*. We compared parameters obtained when incorporating gene conversion and background selection.

Estimates of seelction obtained for protein-coding regions were an order of magnitude higher than those obtained for conserved non-coding elements.

Table 1: Parameters of positive selection in *M. m. castaneus* estimated by fitting model of selective sweeps to troughs in diversity around functional . Standard errors are shown in square brackets

Background Selection	Protein-Coding Exons		Conserved Non-Coding Elements	
	γ_a	p_a	γ_a	p_a
+	9,887	1.24 x 10^{-5}	228	2.27 x 10^{-3}
	[1,914]	[3.90 x 10^{-6}]	[12.8]	[2.40 x 10^{-4}]
-	20,200	8.61 x 10^{-6}	504	1.27 x 10^{-3}
	[1,460]	[9.52 x 10^{-7}]	[18.2]	[7.12 x 10^{-5}]

Table 2: Parameters of positive selection in *M. m. castaneus* estimated by fitting model of selective sweeps to troughs in diversity around functional assuming the Cox et al (2007) genetic map. Standard errors are shown in square brackets

Background Selection	Protein-Coding Exons		Conserved Non-Coding Elements	
	γ_a	p_a	γ_a	p_a
+	[ ]	[ ]	[ ]	[ ]
-	[ ]	[ ]	[ ]	[ ]

Discussion

Tataru *et al.* (2017) performed simulations to assess how accurately positive selection parameters can be obtained from the uSFS when excluding between-species divergence from their analysis. Previous methods to estimate α made the assumption that positively selected variants contribute little to standing genetic variation so can thus be ignored when correcting estiamtes of α using polymorpism data (Eyre-Walker and Smith 2002). Tataru *et al* (2017) showed that estimates of the dDFE can become biased if positively selected mutations contribute to standing variation and are ignored. However, the parameters that Tataru *et al.* (2017) used in their simulations may be fairly unrealistic. For example, to demonstrate

that α can be accurately estimated from polymorphism alone they simulated a population with $\gamma = 400$ (note that they used a different parametrisation of the selection model) and $p_a = 0.02$. This gives $\gamma p_a = 8$, whereas estimates of this parameter in other studies are not nearly so high. For example, Campos *et al.* estimated that $\gamma p_a = 0.055$ in *Drosophila melanogaster* by fitting a model of selection on linked sites to the correlation between synonymous site diversity and divergence at nonsynonymous sites, while Booker and Keightley (Unpublished) estimated $\gamma p_a = 0.0436$ in *M. m. castaneus* by analysis of the uSFS. We simulated populations where $\gamma p_a = 0.1$, but selection was strong ($\gamma = 400$). We found that a) benefical mutations were not detected in standing variation (based on a likleihood ratio test) and b) that while γp_a is relialby estiamtes when including divergence, that the individual parameters cannot be teased apart.

0.1 Analysis of the uSFS

By analysing the uSFS of simulated populations, polyDFE yielded exquisitely accurate estimates of the dDFE from simulated data, even when positive selection was very strong. In these cases, ignoring the strength of

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 background selection or selective sweeps. 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 selective sweeps 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 microsatelite loci across

Table 3: Rough estimates of the changes in fitness caused by new mutations occuring in protein-coding exons and CNEs. Estimates were obtained assuming an effective population size of 420,000 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_a^2	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 *M. m. domesticus* genome. In their study they estimated that selective sweeps 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 selective sweeps in protein-coding exons are driven by mutations with $s \approx 0.0099$ and in CNEs $s \approx 0.00027$.

The contribution of adaptation in protein-coding and regulatory regions to phenotypic evolution in mice

An enduring question in evolutionary biology has been the extent to which protein-coding and regulatory regions of the genome contribute to fitness change (King and Wilson; Carroll). The rate of adaptive fitness change (ΔW) generated by mutations occurring at a particular class of sites can be modelled as the influx of new advantageous mutations multiplied by the square of the expected selection coefficient:

$$\Delta W \propto \mu_a n_a E(s_a^2), \tag{6}$$

where μ_a is the rate of advantageous mutations occuring at a particular class of sites, n_a is the total number of sites in the genome corresponding to that class and s_a^2 is the square of the selection coefficient for advantageous mutations (Halligan et al 2013).

In this study, we have estimated that the selection coefficients of advantageous mutations occurring in protein-coding regions far exceeds that of mutations occurring in conserved non-coding elements (assumed to be regulatory) (Table 2). However, the estiamted p_a was higher for CNEs than for exons (REFERENCE THE RESULTS TABLE), so the However, the total contribution that the two classes of sites make to fitness change will depend upon the total number of the different site types in a species' genome.

We assumed that all new advantageous mutations are semi-dominant, which is something of a prob-

lem. Haldane's sieve predicts that most advantageous mutations that become fixed are dominant. There are a number of examples of selective sweeps 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

Whether or not the conclusions we have drawn in this study can be generalised to other organisms or not remains to be seen. However, patterns of nucleotide diversity in rats are extremely similar to those seen in mice.

Conclusions

In this study we have shown that if advantageous mutations are infrequent and have, on average, strong effects on fitness, their parameters are very difficult to estimate from the site frequency spectrum. However, as has been shown previously (REF DUMP) the DFE for harmful mutations is estimated with precision from the SFS (RESULTS?). We estimated the strength of selection acting in two classes of functional sites in the mouse genome; protein-coding exons and conserved non-coding elements. Our parameter estimates suggest that selection is on average stronger in protein-coding regions of the genome than in regulatory regions, but that the influx of advantageous mutations occuring in into mouse populations is likely larger for regulatory regions. Using a simplistic model of the rate of change in fitness due to new advantageous mutations, we estimate that protein change contributes more to fitness than regulatory change.

Acknowledgements

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

junkyard

Patterns of genetic diversity in a number of species are cons In wild mice, there are troughs in diversity surrounding functional elements. In a recent analysis, we estimated the frequency and selection coefficients of advantageous mutations that occur in mice using distribution of derived allele frequencies (Booker and Keightley submitted). We showed that the parameters of selection obtained from the uSFS are unable to explain the patterns of selection observed in the genome.

Recently, Tataru *et al.* (2017) showed that accurate estimates of positive election parameters an be obtained by analysis of the uSFS. However, the range of selection parameters that analysed may not

Recently, we have estimated the DFE using the uSFS for wild mice and shown that the parameters of selection that we infer do not explain the reductions in diversity observed around protein-coding exons.

One of the long-standing goals of evolutionary biology has been to understand the contribution of coding versus non-coding change to adaptive evolution (King and Wilson; Carroll). Arguments have been made that regulatory regions, which may have a lower pleiotropic burden than protein-coding genes, may dominate phenotypic evolution. Indeed, regulatory regions are, on average, subject to weaker selective constraints than protein-coding regions. In mice, In mice, there are reductions in genetic diversity around both conserved non-