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 17, 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, 2013). 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, 2013). However, the proportion of nonsynyonomous substitutions attributable to adaptive evolution (α) is typically high (50%) (Galtier 2016; but see Booker et al. 2017a 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, 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. 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 models combining the effects of background selection and sweeps had not been developed. However, the effects of background selection 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 background selection is highly important when parametrizing sweep models from patterns of nucleotide diversity (Campos et al., 2017; Elyashiv et al., 2016).

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, 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 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 extending 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., 2011; Tataru et al., 2017). The McDonald-Kreitman test (McDonald and Kreitman, 1991) and its extensions, 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

Simulations

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 introns. Nonsynonymous mutations were modelled as 75% of mutations occurring in exons, the remaining 25% 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 $N = 1{,}000$ diploid 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. We performed 10 such simulations for each set of selection parameters (Table

??). For each locus, time-point and simulation replicate we extracted the simulated nonsynonymous and synonymous sites, giving uSFS data for 10,000 loci. We sampled this set of 10,000 uSFSs, with replacement, 100 times and collated the uSFSs to obtain bootstrap samples.

Across simulated datasets the γ_a and p_a parameters varied, but the product $\gamma_a p_a$ was set to 0.1. Estimates of $\gamma_a p_a \approx 0.1$ have been obtained for *Drosophila melanogaster* in studies that used different methods (Campos et al., 2017; Keightley et al., 2016). All simulations incorporated the same dDFE (β = 0.2 and $\hat{\gamma}_d$ = -1000). The advantageous mutation parameters we simulated are listed in Table **??**.

Analysis of the uSFS

the distribution of fitness effects for new mutations by analysis of our simulated uSFSs using the methods of Tataru et al. (2017) as implemented in their program polyDFE. Briefly, polyDFE uses results from SAWYER and CLARK's PRF theory to obtain an expression for the uSFS expected in the presence of both advantageous and deleterious mutations. This expression is fitted to the uSFS for both putatively neutral and selected classed of sites (e.g. synonymous and nonsynonymous sites, respectively), by maximum likelihood. Tataru et al. (2017) extended the basic PRF framework to account for polymorphism misattributed to divergence, mutation rate variability, error in assigning sites as ancestral/derived and uses the putatively neutral uSFS to correct for non-neutral processes other than selection. 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 data using polyDFE choosing Model C and either including or not including between-species divergence. We analysed the uSFS for simulated nonsynonymous using simulated synoymous sites as the neutral reference class. Selection at linked sites causes distortions to the uSFS for both selected sites and the neutral reference class, we used the nuisance parameters in We analysed 100 bootstrap samples of the simulation data.

Obtaining estimates of B

Background selection contributes to the troughs in diversity around both protein-coding exons and CNEs (Halligan et al 2013; Booker and Keightley Unpublished). Because of this, we required estimates of the effect of background selection on neutral divresity, B, to fit as a covariate when fitting Equation 2 to the diversity troughs. There are formulae for calculating B given the mutation and recombination rates as well as the DFE (Nordborg et al., 1996; Hudson and Kaplan, 1995), but these over-predict the effects of BGS when selection is weak ($\gamma_d < 1$) (Good and Desai; Gordo et al). Since weakly selected mutations comprise a large portion of the DFEs we obtained previously, we opted to obtain estimates of B from simulations. In our earlier study (Booker and Keightley *Unpublished*), we, effectively generating a map of background selection's effects across the mouse genome. By extracting diversity as a function of genetic distance from protein-coding elements and CNEs from these simulations, we obtained estimates of B that can be used when fitting

We incorporated simulation results from our earlier study (Booker and Keightley Unpublished). Briefly, those simulations incorporated the actual distribution of functional elements (the exons and untranslated regions of protein-coding genes as well as conserved non-coding elements) in the mouse genome.

The simulations assumed dDFEs that were estimated by analysis of the uSFS and the recombination map inferred for *Mus musculus castaneus* by Booker et al. (2017b). We increased the number of simulation replicates performed by Booker and Keightley from 2,000 to 6,000 to obtain smoother estimates of B.

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. Reads were mapped and variants called using a BWA and Samtools pipeline (see Halligan et al. 2013 for details).

From the edges of exons (CNEs), polymorphism data was extracted in windows of 1Kbp (100bp) extending to distances of 100Kbp (5Kbp). All non-CpG sites in these windows were extracted, and mouse-rat divergence was calculated. Analysis windows were then binned based on genetic distance to the focal element based on either the LD-based recombination map for *M. m. castaneus* (Booker et al., 2017b) or a pedigree-based genetic map constucted using common lab strains of *M. musculus*. Because LD-based and pedigree based recombination maps have different benefits and drawbacks (discussed below), we perform all analyses in parallel, assuming either both of these recombination maps.

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 Campos et al. (2017) 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 Frisse et al. (2001)

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

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-fold 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.

Model of Recurrent Sweeps with Background Selection

Background selection (BGS) and selective sweeps (SSWs) are processes that induce coalescence at linked sites. 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 ($2N_es_a$).

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

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

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).

We estimate γ_a and p_a by fitting Equation 2 to 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.

One strategy would be to take the value of π at non-constrained sites far from funcitonal elements. In principal, there should be a single π_0 for *M. m. castaneus*, but in practice, genetic diversity varies across the genome.

Results

Estimating selection parameters from the uSFS of simulated data

Parameters of the DFE can be estimated directly from unfolded site frequency spectra (uSFS) if selected mutations are segregating in populations of interest (REFS). It has been repeatedly demonstrated that parameters of the DFE for deleterious mutations (dDFE) can be accurately estimated from population genetic data. It has also been shown that the parameters of advantageous mutations can also be estimated from the uSFS, 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). In this study, we performed simulations to show that accurate estimation of positive selection parameters do indeed depend on the strength and relative frequencies of advantageous mutations.

We used forward-in-time simulations that incorporated linkage, as selection at linked sites causes distortions to the uSFS that likely affect real data and thus cannot be ignored. For each set of 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 quantity of data is fairly typical of population genomic studies (REFS). Using these data we estimated the parameters of selection using polyDFE, an implementation of the methods of Tataru *et al* (2017). These methods allow the simultaneous estimation of the dDFE and positive selection parameters, taking into account distortions in the uSFS caused by, for example, demographic effects and selection at linked sites.

Consistent with Tataru *et al* (2017) we found that estimates of the dDFE obtained by analysis of the uSFS were very accurate, but only when the full DFE is assumed

Across different sets of 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. Simulations all incorporated the same dDFE, so the extent of background selection

Table 1: 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

should be fairly similar, while the effect of selective sweeps varied. We found that selection at linked sites reduced synonymous site diversity below the expectation value of 0.01 in all simulations (Table 0.1), 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 0.1). The relative fixation rate of nonsynonymous mutations did not vary systematically across simulations, dN/dS was fairly constant across simulations as expected (Table 0.1).

We analysed the uSFS from our simualted populations and found that when advantageous mutations are relatively frequent ($p_a > 0.0005$), but weakly selected ($\gamma_a < 100$), then both γ_a and p_a parameters could be estimated with precision (Table REF). 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 scaled selection on advantageous mutations was ≥ 100. Furthermore, likelihood ratio tests typically failed to detc the presence of advantageous mutations contributing to the uSFS.

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 0.1). This all suggests that when advantageous mutations are strongly selected, but rare, patterns of nucleotide diversity carry information that is not present in the unfolded site frequency spectrum.

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

We compared the the relationship between nucleotide diversity and genetic distance to functional elements assuming two different genetic maps constructed 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 and this may be most pronounced close to functional elements.

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 *castaneus* 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.

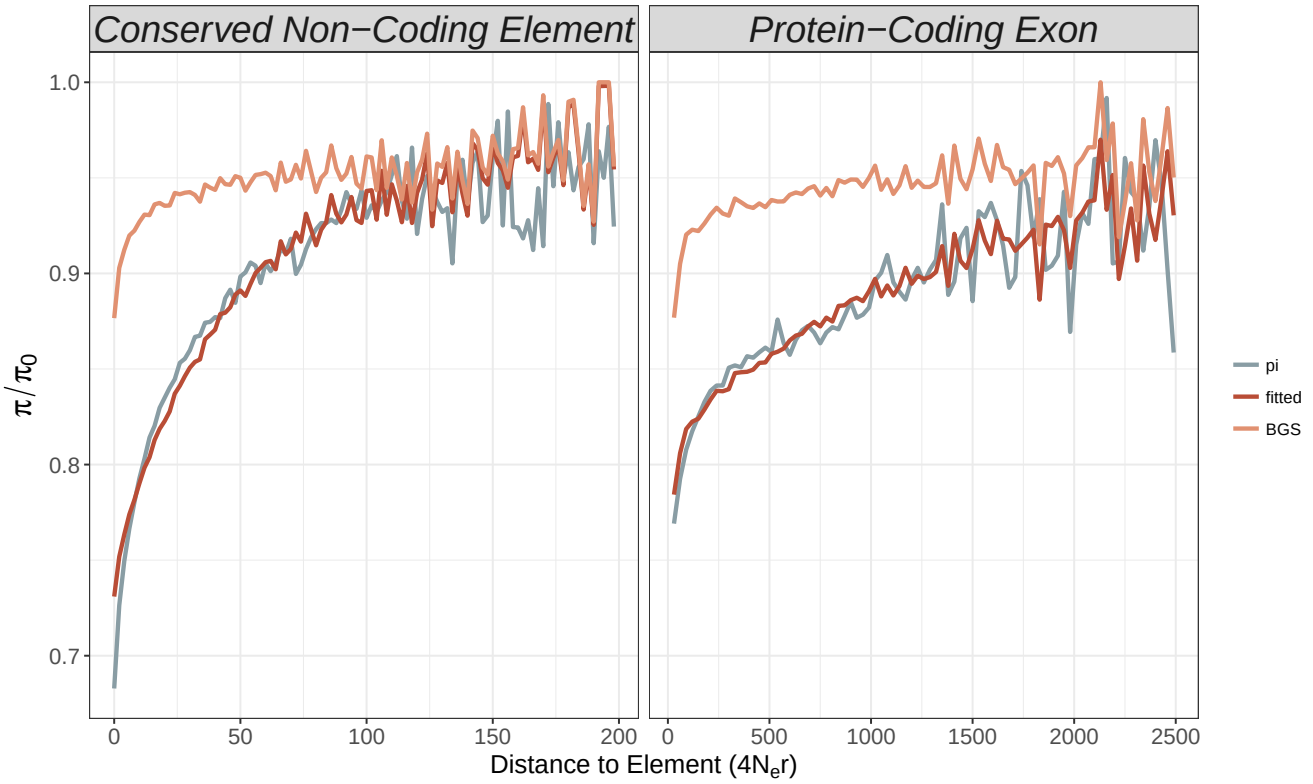

Figure 1:

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.

Parameters of selective sweep obtained from patterns of nucleotide diversity

By fitting Equation 2 to the troughs in diversity surrounding protein-coding exons and CNEs, we were able to estimate that very strongly selected mutations may occur in both elements. Regardless of which recombination map we assume, selection coefficients for mutations occurring in exons were order of magnitude greater than CNEs. Comparing selection parameters obtained assuming the Cox ans *castaneus* maps highlight a

If we assume a long-term effective population size of 420,000 for *M. m. castaneus*, we estiamte that selection coefficients in natural popuations of ≈ 0.01

We compared the fit of different models for the DFE for advantageous mutations. We compared the fit of one or two classes of discrete effects as well as the exponential distribution. In the case of protein-coding exons a single class of effects or an exponential distribution gave similar fits to the data, as judged by differences in AIC, regardless of whether we used the *castaneus* or Cox maps to estimate genetic distances. In the case of CNEs, on the other hand, a single class of advantageous mutations was

supported in when analysing using the Cox distances, but two class of effects were strongly supported when using the *castaneus* map.

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

A key parameter in Equation 2 is π_0, the nucleotide diversity expected under strict neutrality. This parameter is very difficult to estimate and may even prove unobservable in real data (Kern and Hahn?). However, an estimate of π_0 is required to fit the troughs in diversity. When fitting the data, the value of this parameter we assumed depended on which recombination map we assumed and which functional element was being analysed. The distribution of functional elements surrounding protein-coding exons and CNEs differs, which will affect the level at which nucleotide diversity plateaus surrounding those elements, as the effects of selection at linked sites will differ between the two. This may explain why the level at which diversity plateaus around the two classes of elements, as can be seen in Figure 1. The reductions in diversity caused by selective sweeps occuring at linked elements will differ around CNEs and protein-coding exons as the distribution of function unobservable in the patterns of nucleotide diversity around both classes of elements analysed in this study, as even where neutral diversity plateaus, it is reduced below its expected

Table 2: 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	1.24×10^{-5}	228	2.27×10^{-3}
	[1,914]	[3.90×10^{-6}]	[12.8]	[2.40×10^{-4}]
-	20,200	8.61×10^{-6}	504	1.27×10^{-3}
	[1,460]	[9.52×10^{-7}]	[18.2]	[7.12×10^{-5}]

Table 3: 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 estimates of α using polymorphism 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) beneficial mutations were not detected in standing variation (based on a likelihood ratio test) and b) that while γp_a is reliably estimated 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, TTeschke et al. (2008) analysed patterns of variation at microsatellite loci across 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 $s \approx 0.00027$.

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 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,

Table 4: 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 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

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 differenes between humans and chimps. Furthermore, Carroll (2005) suggested that pleiotropy may place a burden on protein-coding genes such that adaptation most often occurs in regulatory regions. Using a simple model of adaptive fitness change, our selection parameter estimates allow us to examine this problem in mice.

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 obtained assuming the *castaneus* map. We assume that the mutaiton rate is the same for both CNEs and protein-coding exons, so we can ignore μ in Equation **??**.

Our parameter estimates suggest that substitutions in protein-coding regions contribute more to fitness change than do substitutions in regulatory regions. The target size for advantageous mutations in CNEs is far larger than for protein-coding exons (there are approximately three times as many CNE sites than there are nonsynoynmous sites in the mouse genome and p_a is approximately an order of magnitude higher). However, since the change in fitness is dependant on the square of the selection coefficient (it is related to the additive genetic variance in fitness), so the ten-fold difference in selection coefficient for protein-coding mutations versus regulatory mutations makes a hundred-fold difference to the change in fitness.

There are a number of factors that should, perhaps, temper these conclusions. Firstly, the selection coefficient that appears in Equation **??** is the expectation of the DFE for advantageous mutations. If the shape of the DFE for advantageous mutations were, for instance, highly leptokurtic or bimodal then using the expectation value, rather than integrating over the full DFE, may give misleading results. While we found that discrete classes gave a better fits to the data than the continuous exponential distribution (TABLE REF), we do not suppose that the DFE for advantageous mutations is, in reality so simple. Secondly, we have assumed that all CNEs share the same DFE. This is slightly problematic since there will is likely a large number of sub-categorisations that could be applied to the set of CNEs we analysed (e.g. promoters and enhancers may be subject to differnet selective pressures). Indeed, sub-categorisations of protein-coding genes may also be subject to different selection pressures; in humans for example, virus interacting proteins seem to be subject to a unique suite of selection parameters (Enard eLife paper).

Whether or not the conclusions we have drawn in this study can be generalised to other organisms remains to be seen. Brown rats, *Rattus norvegicus*, provide a compelling first case for comparison, as in that species 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). Additionally, broad-scale recombination rates are strongly correlated between mice and rats (Jensen-Seaman et al., 2004), suggesting then, that qualitatively similar conclusions regarding the contribution of protein-coding versus regulatory change to adaptive evolution may be reached when analysing rats.

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 occurring 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.

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.

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*, 114(24):E4762–E4771.

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.

Corbett-Detig, R. B., Hartl, D. L., and Sackton, T. B. (2015). Natural selection constrains neutral diversity across a wide range of species. *PLoS Biol*, 13(4):e1002112.

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.

Frisse, L., Hudson, R. R., Bartoszewicz, A., Wall, J. D., Donfack, J., and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet*, 69:831–843.

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

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.

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

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

McDonald, J. M. and Kreitman, M. (1991). Adaptive protein evolution at the adh locus in drosophila. *Nature*, 351.

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.

Orr, H. A. (2003). The distribution of fitness effects among beneficial mutations. *Genetics*, 163:1519–1526.

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.

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.

Table S1: Summary statistics for simulated populations

γ_a	p_a	π/π_0	dS	dN	dN/dS
5	0.010000	0.940	0.00533	0.00159	0.299
10	0.005000	0.914	0.00526	0.00157	0.299
25	0.002000	0.880	0.00527	0.00159	0.302
50	0.001000	0.862	0.00525	0.00162	0.309
100	0.000500	0.844	0.00531	0.00159	0.300
200	0.000250	0.819	0.00523	0.00155	0.296
400	0.000125	0.795	0.00527	0.00156	0.295

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-

Table S2: Parameters of the distribution of fitness effects for harmful mutations obtained by analysis of the uSFS

γ_a	p_a	Divergence[a]	Full DFE[b]	β[c]	$\hat{\gamma_d}$ [c]
		+	+	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]
		+	+	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]
		+	+	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]
		+	+	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]
		+	+	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]
		+	+	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]
		+	+	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