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

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*

April 4, 2018

Abstract

Woah! Selective sweeps are neat. Let's look at a model for estimating their strength and frequency from pop. gen. data using a more developed version of the approach adopted by Wiehe and Stephan in 1993. Let's go on to compare the strength and frequnecy of new mutations occuring in protein-coding versus regulatory regions and see which one will contribute more to phenotypic change.

Introduction

Random mutation is a fundamental force in evolution, it the ultimate source of all biodiversity Populations may become better adapted to their environments (REF), experience 'mutational meltdown' (REF), undergo evolutionary rescue (REF) and become speciated all through the effects of new mutations (REF). In each of these processes, the frequencies and fitness effects of new mutations is a key parameter.

In order to answer this question, however, methods are required for esitmating the rates of occurence and fitness effects of new mutations.

One class of methods for inferring the strength and frequnecies of new mutations relies on the assumption that selected mutations segregate in natural populations. The distributon of allele frequencies in a sample, the site frequency spectrum, for a class of sites subject to natural selection

In the past 30 years of population genetic research it has become clear that natural selection shapes patterns of genetic diversity across the genomes of many species (Corbett-Detig *et al* 2015; Cutter and Payseur 2012). Because genetically linked sites in the genome do not evolve independently, selection acting at one site may have consequences for linked sites. This process, which we shall refer to as selection at linked sites, is intrinsically linked to the rate of recombination both by crossing-over and gene conversion (REF DUMP).

One of the biggest goals of evolutionary genetics is to understand the fitness consequences of new mutations. There is undoubtedly a distribution of fitness effects for new mutations, but this distribution may depend on a number of factors such as the genomic region where the mutations occur and the fitness of the population in question (Peter's Review?). Experimental approaches for estimating the DFE are limited to organisms which can be maintained in lab populations, which typically excludes mammals.

When advantageous mutations are driven to high frequencies by selection, they drag with them portions of the haplotype on which they are present. This process, termed a selective sweep, has been the subject of rigorous population genetic research (Maynard-Smith and Haigh, Hudson and Kaplan, Coop and Ralph, Hermisson and Pennings, Barton 2000). There are a number of models decribing the ways in which selective sweeps may proceed (reviewed in Booker *et al.* 2017). One of the consequences of selective sweeps is a reduction in neutral diversity in the regions surrounding advantageous mutations.

The time it takes an advantageous mutation to sweep through a population is proportional to the strength of selection and this has an effect on the level of neutral diversity in genomic regions linked to the mutation. As a mutation sweeps, linked neutral variants may get trapped by the sweep and carried to high freuqency with the advantageous mutation. Corssing-over can break apart associations between neutral variants and sweeping mutations, but obviously if a mutation is swept rapidly, there are fewer chances for crossing-over to occur. Because of this, selective sweeps generate troughs in genetic diversity, centred on the site of the advantageous mutation. The width and depth of the trough are dependant on the rate of crossing-over but also, crucially, the strength and frequency of advantageous mutations.

In wild mice, there are troughs in diversity surrounding functional elements. In a recent analysis, we estimated the frequnecy and selection coefficeients 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, 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.

In this study, we use a model of selective sweeps to estimate the strength and frequency of advantageosu mutations that occur within protein-coding exons and regulatory elements. The model we use incorporates the confounding effects of background selection as well as gene conversion as both processes are known to influence estimates of selection obtained from patterns of nucleotide polymorphism. We use simulations to validate our approach and to also demonstrate that uSFS-based methods fail to detect the strength and frequnecy of new mutations when they are rare.

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 fx(t) is the proportion of sites in the DFE with a selection coefficient of t.

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 is the per-base pair per generation mutation rate (assumed to be 5.4×10^{-9} (UCHIMURA et al. 2015)), 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. It is straightforward to incorporate a distribution of advantageous mutation effects to Equation 3:

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

In this study, we assume an exponential distribution for the distribution of fitness effects for advantageous mutations.

Gene conversion can be included in the above formulae if we assume that the distribution of gene conversion tract lengths is exponential by substituting the following in for the recombination distance:

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

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. If gene conversion is present, but not accounted for, one would underestimate the strength of selection, it is

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.

Table 1: Distributions of fitness effects in simuations. In all simulations, deleterious mutations were drawn from an exponential DFE with γ_d = -48.50. Where $\gamma = 2N_e s$

DFE Model	γ_a	p_a	Label
Bimodal	400 / 20	0.001 / 0.009	Bimodal
	400 / 20	0.0001 / 0.0009	Bimodal - div10
	400 / 20	0.00001 / 0.00009	Bimodal - div100
Exponential	200	0.001	Exp200
	200	0.0001	Exp200 - div10
	20	0.001	Exp20
	20	0.0001	Exp20 - div10
Fixed	400	0.001	Nes200
	400	0.0001	Nes200 - div10
	20	0.001	Nes10
	20	0.0001	Nes10 - div10

Simulations

We simulated background selection and selective sweeps using the forward-time simulation package SLiM (v1.8; Messer 2012). We performed simulations of a single 1Kbp protein-coding exon, flanked up and downstream by 70Kbp of strictly neutral sequence. 75% of sites in the simulated exon were subject to selection (i.e. nonsynonymous sites) and the remainder were strictly neutral (i.e. synonymous sites). The population-scaled mutation rate ($\theta = 4N_e\mu$) was set to 0.01 and the population-scaled recombination rate ($\rho = 4N_e r$) was set to either 0.009, 0.0045 or 0.001. For a given distribution of fitness effects (Table **??**) we performed 1,000 replicate simulations at each recombination rate resulting in 3,000 replicates per set of selection parameters. We ran simulations of 1,000 individuals for 20,000 generations to ensure that equilibrium conditions have been reached. At the final generation, 20 haploid genomes were sampled from the population. From these, we extracted the patterns of diversity around the exon or the uSFSs for nonsynonymous and synonymous sites within the exon itself.

Analysis of the uSFS 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. 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.

In mice, there is either non-crossover gene conversion, or gene conversion associated with crossing over events. It has been shown that the average gene conversion tract length differs in crossover or non-crossover gene conversion so we extended the recombination distance used

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 populiaton-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.

Rates of initiation of gene conversion in mice are known in mice. We incorporated non-crossover associated gene conversion A tract length of 145bp for non-crossover associated gene conversion was used (Paigen *et al.* 2008)

We incorporated background selection as a covariate when fitting Equation 3 using the simualtions results we obtained in an earlier study (Booker and Keightley *et al Unpublished*). These simulations incoporated 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.* (*Unpub-*

lished) 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.

Estimating the effects of background selection

We incorporated the effects of background selection into

Results

Estimating selection parameters from simulations

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.

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.

Additionally, if selective sweeps occur frequently, the assumpiton that multiple sweeps do not interfere with each other may be violated. We found that correcting for multiple, competing sweeps did not make much difference on parameter estimates. This sugegsts that the rate of sweeps in mice is too low for the

Modelling the effects of BGS and selective sweeps on patterns of diversity at linked

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

Table 2: Estimates of the DFE obtained from troughs in diversity around functional elements. Standard errors are shown in square brackets

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

Discussion

The rate of sweeps we estiamted for

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 frueqnecy spectrum expected under either backgorund selection or selective sweeps. Rather, nuisance parameters or demographic models are used to account for the conribution of selection at linked sites to the SFS while estimating the DFE. However, there is potentially information present in the uSFS that would be useul 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 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 regulatroy regions of the gneome contribute to fitness change. In this study, we have found that the strength of selection acting on new mutations occuring in protein-coding regions far exceeds that of those occuring in conserved non-coding elements (assumed to be regulatory) (Table 2). However, the total contribution that the two classes of sites make to fitness changes will depend upon the total number of the different site types in a species' genome. In mice, there is a much large number of non-coding sites in the genome inferred to be targets of selection than there are sites that encode proteins (Halligan et al 2013; MORE MORE MORE).

The rate of adaptive fitness change (ΔW) generated by mutations occurring at a particular class of sites can be modelled as

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

where μ_a is the rate of advantageous mutations occuring at a particular class of sites, n_a is the total

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

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).

A weakness of this study is that we did not estimate the strength of selection acting on the untranslated regions (UTRs) of protein-coding genes. The question as to whether adaptation on protein-coding or

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, 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; Matty Hartfield's paper). Because of this, as long new mutations are neither fully reccessive nor dominant, the selection coefficients we estiamted should be directly proportional to the true values.

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?).