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*

March 6, 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

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:

$$B = \frac{N_e}{N_0} \approx exp\left[-\sum_x \int_0^1 \frac{u_x f_x(t)\, dt}{t\left(1 + \frac{(1-t) r_{x,y}}{t}\right)^2} \right] \tag{1}$$

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).

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

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.

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

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:

$$P_{sc,j} \approx \int_{0}^{1} f_x(\gamma) V_a \tau \gamma_j^{\frac{-4r_{i,j}}{s}}\, d\gamma \tag{4}$$

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

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 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 ($4N_e\mu$) was set to 0.01 and the population-scaled recombination rate ($4N_e r$) was set to either 0.009, 0.0045 or 0.001. For a given distribution of fitness effects, see below, 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

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

Table 1: Distributions of fitness effects in simuations. In all simulations, deleterious mutations were drawn from a gamma DFE with $\gamma_d = 48.50$.

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	4
	200	0.0001	5
	20	0.001	6
	20	0.0001	7
Fixed	200	0.001	8
	200	0.0001	9
	20	0.001	10
	20	0.0001	11

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.

DFE-alpha does not currently allow the user to estimate an exponential distribution of advantageous mutational effects. We use the PRF-based method polyDFE to infer the parameters of the DFE from the uSFS of simulations incorporating the

1. DFE-alpha

The methods of Schneider et al (2011) are implemented in the program DFE-alpha. We analyse the simulation data using DFE-alpha

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 ($4N_er$) 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. A difficulty is that genetic

Results

These are the results

Discussion

And this is what I think of them

One of the enduring questions in evolutionary genetics concerns the contribution of protein-coding versus gene-regulatory variation to a species' fitness (REF DUMP). Evidence in multiple species suggests that selection is frequent in non-coding portions of the mammalian of the genome (Halligan *et al.* 2013; Booker and Keightley *submitted*; MORE MORE MORE). With the