Understanding variation in nucleotide diversity across the mouse genome

or: The myth of sweep-syphus

Tom (Royal T.) Booker

Submitted for the degree of Doctor of Philosophy

University of Edinburgh
2018

Declaration

This place holder text is standing in for the University of Edinburgh's declaration page.

Dedication

I dedicate this thesis to my good friend Arya

Acknowledgements

First and foremost, thanks to Peter Keightley. He has been an excellent supervisor and mentor for the last four years. Peter's help and guidance has helped me develop in many ways and has made my PhD an incredibly enjoyable and rewarding experience. Peter's thorough and rigorous approach to science is something that I aspire to. Thanks for introducing me to disc golf too!

I would also thank Brian Charlesworth for helping me understand the thornier aspects of population genetics that I have come across. Brian has been very patient with me and willing to give help and advice whenever I have asked for it, which has been hugely helpful throughout my PhD.

Thanks to Deborah Charlesworth for being very kind and generous in helping me understand many aspects of evolutionary biology, not limited to my own research. Thanks to participants of the evolutionary genetics lab group and genetics journal club in IEB for great discussions and feedback about my research, particularly Susie Johnston and Konrad Lohse.

Thanks to members of the Keightley lab past and present for listening to me go on and on about selective sweeps or other things for the past four years. In no particular order: thanks Ben Jackson, Rory Craig, Susanne Kraemer, Eva Deinum, Thanasis Kousathanas and Matty Hartfield. Rob Ness taught me to code and helped me out in numerous ways, it's just a shame he left Edinburgh when he did. Without the help of Dan Halligan, who tolerated me bugging him even after he left the lab, I would have been very lost.

I would also extend massive thanks to Sally Otto and members of her lab group at UBC for giving me such a welcoming place to work and write-up my thesis in Vancouver.

On the personal side of things, thanks to my friends in IEB and the whole Berwickshire gang for palling around with me in Edinburgh. Thanks to Jaz and Michael for all the great hangs in Vancouver.

My family has been very supportive throughout my whole PhD, particularly my amazing Mum and Dad. My brothers and their SAPs are great too, but Stella, you really nailed it.

Finally, Arya has looked after me and kept my head straight for the past four years. She is the best and even married me during this PhD, can you believe that?

Publications

The following publications have arisen from this thesis:

- Booker, T. R., Ness, R. W., & Keightley, P. D. (2017). The recombination landscape in wild house mice inferred using population genomic data. *Genetics*, 207(1), 297-309.
- Booker, T. R., Jackson, B. C., & Keightley, P. D. (2017). Detecting positive selection in the genome. *BMC Biology*, 15(1), 98.

The following has been prepared as a research paper is currently under review at *MolecularBiologyandEvolution*:

- Booker, T. R., & Keightley, P. D. (*Submitted*). Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome. *bioRxiv*, 275610.

I contributed to the following papers during my PhD, but these do not form part of this thesis:

- Booker, T., Ness, R. W., & Charlesworth, D. (2015). Molecular evolution: breakthroughs and mysteries in Batesian mimicry. *Current Biology*, 25(12), R506-R508.
- Keightley, P. D., Campos, J. L., Booker, T. R., & 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), 975-984.

Contents

ThesisAbstract

It is well understood that nucleotide diversity varies across the genomes of many eukaryotic species in ways consistent with the effects of natural selection. However, the contribution of selection on advantageous and deleterious mutations to the observed variation is less well understood. In this thesis, I aim to disentangle the contribution of background selection and selective sweeps to patterns of genetic diversity in the mouse genome, thus furthering our understanding of natural selection in mammals. In chapter 1, I introduce core concepts in evolutionary genetics and describe how recombination and selection interact to shape patterns of genetic diversity. I will then describe three projects in which I examine aspects of molecular evolution in house mice. In the first of these, I estimate the landscape of recombination rate variation in wild mice using population genomic data. In the second, I estimate the distribution of fitness effects for new mutations, based on the site frequency spectrum, I then analyze population genomic simulations parametrized using my estimates. In the third, I use a model of selective sweeps to estimate and compare the strength of selection occurring in protein-coding and regulatory regions of the mouse genome. This thesis demonstrates that selective sweeps, are responsible for a large amount of the variation in genetic diversity across the mouse genome.

Chapter 1

Introduction

Parts of this introduction have been published as a review article in BMC Biology. Sections marked with an (*) have been reproduced, with minor modifications to the text. The published article is reproduced in full in the Appendices.

1.1 Understanding the causes of variation in genetic diversity

"If we take the Darwinian view that evolution is the conversion of variation between individuals into variation between populations and species in time and space, then an essential ingredient in the study of evolution is a study of the origin and dynamics of genetic variation within populations" (Lewontin 1974)

This quote, from Lewontin's 1974 book *The Genetic Basis of Evolutionary Change*, eloquently demonstrates that understanding the causes of variation between and within populations is one of, if not the central goal of evolutionary genetics and has been for a long time.

In the latter half of the 20^{th} century, one of the foundational theories in population genetics was developed, the neutral theory of molecular evolution. The neutral theory contended that molecular changes between populations were predominantly the result of random changes in allele frequencies through time (Kimura, 1983). In the past 30 years of population genetic research, the increasing availability of DNA sequence data has led to an understanding that neutral evolution cannot readily explain patterns of variability within species. While it is clear that pure neutrality cannot explain observed patterns, an understanding of the factors that shape molecular variability in natural populations is far from complete.

In this thesis, I describe three projects in which I have analysed different aspects of population genomic data from wild mice, which aim to increase our understanding of the factors shaping molecular variation in the genome. Wild mice are an excellent model system for studying molecular evolution in mammals for several reasons. Firstly, being one of the most well-studied organisms in all of science, the genomic resources developed for *Mus musculus* are among the best available for any animal. Secondly, the size of wild mouse populations are very large, which results in high levels of genetic diversity (*see below*), providing power to statistical analyses.

1.2 Core concepts in evolutionary genetics

Throughout this thesis I will refer to a number of fundamental concepts in population genetics. I give a brief description of several of these here.

1.2.1 Genetic drift

In a finite population, the random sampling of individuals contributing to the next generation causes stochastic changes in allele frequencies. In a randomly mating, diploid population of size N, the probability that a new mutation, free from the effects of selection, goes to eventual fixation is simply its starting frequency, $\frac{1}{2N}$ (assuming autosomal inheritance and non-overlapping generations). This implies that for any new, neutral mutation that arises, there is much larger probability, $1 - \frac{1}{2N}$, that it will be lost from the population. The change in allele frequency caused by random sampling is termed genetic drift. Genetic drift is very effective in small populations and less effective in large populations.

The idealised population stated above is referred to as a Wright-Fisher population. Obviously natural populations violate the stated assumptions, for example humans have overlapping generations. In the above statement of the probability of fixation in the Wright-Fisher model, the census number of individuals (N) appeared. In order to model populations that violate the assumptions of the Wright-Fisher model, the effective population size (N_e) is used. N_e can be thought of as a property of a populations, the number of individuals the number of individuals for a population describes the number of individuals in a population that

As populations tend to infinite size, allele frequencies are not, on average, expected to change from generation to generation (Hardy-Weinberg). Genetic drift does not operate solely on neutral alleles, indeed selected alleles may be lost through drift, but the probability of this depends on the selection coefficient (*see below*).

1.2.2 Neutrality and coalescence

The neutral theory of molecular evolution, proposed by Motoo Kimura and later refined by Tamoka Ohta (REFS), posited that the vast majority of molecular evolution could be attributed to genetic drift with only a minority attributable to adaptation. In the strict sense, the neutral theory deals with mutations that are completely free from selective effects. A more relaxed definition, referred to as the *nearly* neutral theory includes mutations that are weakly deleterious, such that their fates are predominantly decided by drift (*see below*). Although tenets of the neutral theory have largely been shown to not hold in natural populations (KErn and Hahn, Charlesworth, Kreitman), the use of neutrality as a null hypothesis in molecular evolution has led to the development of a large number of statistical tests and even the development of the coalescent.

The coalescent is a framework for modelling molecular evolution that considers the evolutionary history of a sample of alleles drawn from a population. Working backwards in time, two alleles are said to *coalesce* when they share a common ancestor at a particular time in the past. The probability of coalescence t generations ago for a pair of neutrally evolving alleles is,

$$Pr(t) = \left(1 - \frac{1}{2N}\right)^{t-1} \frac{1}{2N}. \tag{1.1}$$

Equation 1.1 is a geometric distribution, which, as $N \to \infty$ can be approximated by the continuous time exponential distribution. The rate parameter of this exponential distribution would be $\lambda = \frac{1}{2N}$ and since the mean of the exponential distribution is the inverse of the rate parameter, the mean time to coalescence for a pair of randomly selected neutral alleles when N is large is simply $2N$.

1.2.3 Population structure

The simplest population genetic models assume that the gametes that are sampled to produce progeny for the next generation behave like beans in a bag, where each bean has an equal probability of being sampled. Natural population do not necessarily behave like bags of beans, however. Consider the case of an animal species that populates a long narrow range that stretches from North to South. Individuals at the top of the range, when seeking a mate, will be more likely to reproduce with an individual in their vicinity, rather than one from the South. Over time, this could lead to differences in allele frequencies along the North/South gradient. If one were to sequence individuals from across the range and perform a population genetic analysis on the resulting data, the differences in allele frequency due to the population structure could generate spurious results. The hypothetical example given above is relatively crude, true populations may be structured in very subtle (or in some cases not so subtle) ways that influence data analysis. Throughout this thesis, population structure is not analysed explicitly, but it is discussed as a potential confounding factor.

1.2.4 Mutation and genetic diversity

Mutation is the ultimate source of all biodiversity. They can be defined as a heritable change in an organisms' genetic sequence. Mutations may occur during DNA repair, be affected by physical or chemical agents (e.g. UV radiation) or caused by errors in replication meiosis or mitosis. There are numerous categories of mutations that have been described (e.g. translocations, inversions, insertions, deletions and point mutations). The most common of type of mutations, and the one most pertinent to this thesis is single nucleotide substitution. The rates at which point mutations arise are far estimated to be an order of magnitude higher than for small insertoin deletions and other structural mutations (REFS). Point mutations involve the change of a single nucleotide from one of the four bases to another (e.g. adenine to cytosine). The second most frequent class are insertion/deletion mutations. Throughout this thesis I only with point mutations, unless otherwise stated.

Depending on the population size, new mutations may be heavily influenced by genetic drift. The rate at which new mutations enter into the population is obviously dependant upon the population size, the more individuals there are, the more chances there are for mutations to occur. The probability that two randomly chosen individuals differ at a particular nucleotide in the genome is dependant on the time since they shared a common ancestor t. Working down both branches leading to the common ancestor, there is a total of $2t$ generations separating the two alleles. If the per site per generation mutation rate is μ, then in the time separating two alleles, there is a probability of $2Nt$ that a mutation arose. As we saw above, the mean time to coalescence for pair of neutral alleles is $2N$, so the probability that two randomly chosen alleles differ is $\theta = 4N\mu$. When dealing with sequence data, the average number of pair wise differences between (π), gives an estimate of the probability that two randomly chosen alleles differ in state, so π is can be used as an estimator of θ.

1.2.5 Selection

The vast majority of the content in mammalian genomes is not evolutionarily conserved, and thus inferred to be non-functional (Graur?). Because of this, a large proportion of the mutations that arise will have little effect on their carriers' fitness. However, new mutations occurring in functional regions may be subject to selection. Throughout this thesis, I will refer to the selection coefficient (s), which is defined in terms of the relative fitnesses of the three genotypes possible at a biallelic locus, where A is the wild-type allele and a is the allele subject to selection:

Genotype	AA	Aa	aa
Fitnesses	1	$1+sh$	$1+s$

Here, h is the dominance coefficient of new mutations; under additivity $h = 0.5$, complete dominance $h = 1$, complete recessivity $h = 0$ and heterozygote advantage $h > 1$. Unless where stated, h selected mutations are expected to behave additively.

Even when selection is acting on a new mutation, genetic drift may still operate. I As derived by both Kimura and Fisher, the fixation probability for a mutation with selective effect s, segregating at a frequency of (q) in a population of size N_e is,

$$u(s, N_e, q) = \frac{1 - e^{-2N_e sq}}{1 - e^{-2N_e s}} \tag{1.2}$$

In a Wright-Fisher population, the frequency of a new mutation is $\frac{1}{2N_e}$, so the exponent in the left-hand side of the numerator in Equation 1.2 is $\approx s$. If N_e were large and the mutation were advantageous, the denominator $(1 - e^{-2N_e s}) \to 1$, so $u(s, N_e, q) \approx s$. From Equation 1.2 it can be observed that the fixation probabilities of new mutations are dependant on the relative magnitudes of N_e and s. If the product $N_e s \leq 1$, then the fate of the mutation is similar to that of a neutral allele and genetic drift dominates. If the product $N_e s \gg 1$, then the mutation behaves deterministically,

Note that throughout this thesis I frequently use the neutral to describe molecular evolution dominated by genetic drift, including cases where $s << \frac{1}{2N}$

1.2.6 The McDonald-Kreitman test

One of the most widely used methods for the detection of positive selection is the McDonald-Kreitman (MK) test (McDonald and Kreitman, 1991). The MK-test contrasts molecular variation at putatively neutral sites with variation at potentially selected sites. Typically, synonymous and nonsynonymous sites in protein-coding genes are used as the neutral and selected site categories, respectively. Under the neutral theory, differences between species accumulate by the random fixation of neutral alleles due to drift, with a negligible contribution from positive selection. As such, the ratio of nucleotide diversity at nonsynonymous/synonymous sites (π_N/π_S) should be statistically indistinguishable from the ratio of nonsynonymous/synonymous divergence (D_N/D_S). An elevation of D_N/D_S above the value expected given π_N/π_S is taken as evidence for positive selection. The MK-framework has been expanded to allow the estimation of the proportion of nucleotide substitutions at a particular class of sites, driven by positive selection (α)

$$\alpha = 1 - \frac{D_N \pi_N}{D_S \pi_S}. \tag{1.3}$$

An extension which is based on Charlesworth (1994). However, since weakly deleterious mutations at, for example, nonsynonymous sites may segregate in a population, estimates of α obtained in this way are likely underestimates. Further developments of the MK-test have been made that estimate the distribution of fitness effects for new mutations and use this to correct for the contribution deleterious mutations make to MK-type analyses.

1.2.7 Recombination

In this thesis, the term recombination is used to describe the exchange of genetic information that occurs between sister chromatids during XXphase of meiosis. Here I briefly outline the main processes involved in meiosis in mammals. in particular in *M. musculus.* During anaphase,

synaptonemal complex

the formation of double-strand breaks

the repair of DSBs through end joining

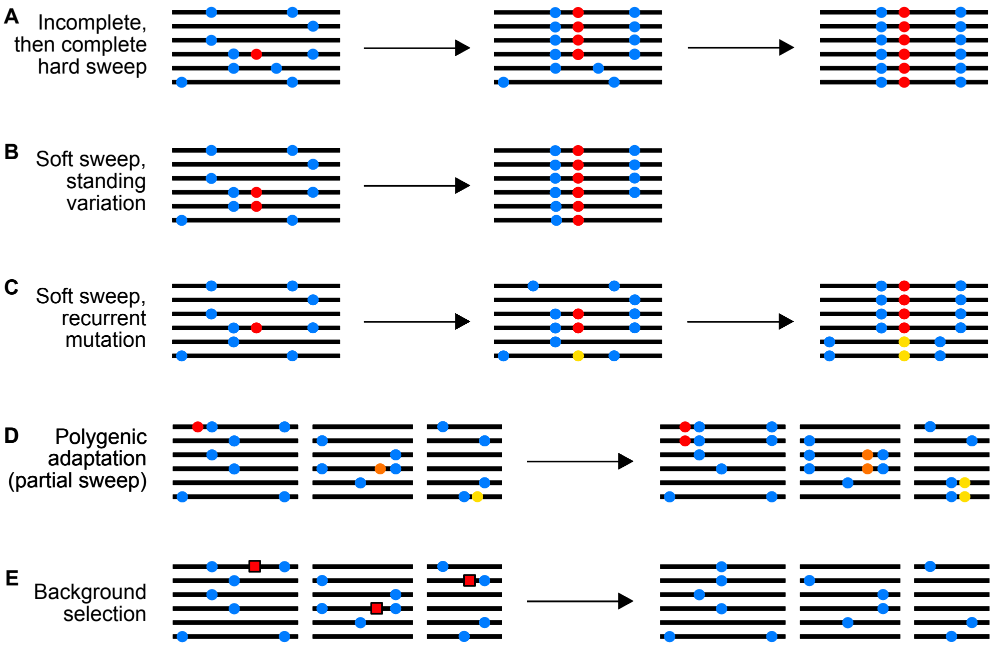
gene conversion or crossing over

the localisation of DSBs - PRDM9 - hotspots

Much focus in the evolutionary literatue has focussed on the

1.3 Models of selective sweeps *

Maynard Smith and Haigh (1974) showed that an advantageous mutation drags with it linked neutral polymorphisms as it rises in frequency. With increasing genetic distance from the selected site, the effect is reduced, resulting in troughs in genetic diversity in surrounding regions.


A Incomplete, then complete hard sweep
B Soft sweep, standing variation
C Soft sweep, recurrent mutation
D Polygenic adaptation (partial sweep)
E Background selection

.

Figure 1.1: Selective sweeps and background selection. Blue circles are neutral varants, Reproduced from Booker et al. (2017a), thanks to Ben Jackson.

Hard/classic sweeps

The most well-studied model of sweeps. A new advantageous mutation rapidly increases in frequency to eventual fixation (shown in Figure 1.1-A). As it sweeps, the adaptive allele carries with it a portion of the haplotype on which it arose, reducing levels of neutral diversity in the surrounding area (Maynard Smith and Haigh, 1974; Barton, 2000).

Soft sweeps

A neutral allele segregating in a population may become favoured (due, for example, to a change in the environment). The segregating allele may be associated with multiple haplotypes, and as it rises in frequency, so do the multiple haplotypes (shown in 1.1-B). A similar process, also termed a soft sweep, can occur if an advantageous mutation arises by multiple, distinct mutation events (shown in 1.1-C).

Incomplete/partial sweeps

If an advantageous allele increases in frequency, but does not reach fixation, there will still be some loss of linked neutral diversity. In this review we use the term incomplete sweeps to describe sweeps that are polymorphic at the time of sampling, but may (or may not) eventually reach fixation (shown in 1.1-A). The term partial sweep describes the situation wherein a sweeping allele becomes effectively neutral at a certain frequency in its trajectory (shown in 1.1-D). The magnitude of both processes on linked neutral diversity depend on the frequency reached by the sweeping allele when selection is turned off or on the time of sampling (Coop and Ralph, 2012). Partial sweeps may be common in cases of adaptation involving selection on quantitative traits (Pritchard et al., 2010).

1.4 Using Models of Selective Sweeps to Estimate Positive Selection Parameters *

Population geneticists have long sought to understand the contribution of natural selection to molecular evolution. A variety of approaches have been proposed that use population genetic theory to quantify the rate and strength of positive selection acting in a species genome. In the following section, I discuss methods that use patterns of between-species nucleotide divergence and within-species diversity to estimate positive selection parameters from population genomic data. We also discuss recently proposed methods to detect positive selection from a populations haplotype structure. The application of these tests has resulted in the detection of pervasive adaptive molecular evolution in multiple species.

If adaptive substitutions are common, selection is expected to leave footprints in genetic diversity at linked sites. In particular, as a positively selected mutation increases in frequency, it tends to reduce diversity at linked neutral loci. Theoretical analysis of this process, termed a selective sweep (*see above*), has shown that the reduction in diversity at a linked neutral locus depends on the ratio of the strength of positive selection to the recombination rate. Thus, comparing diversity at multiple neutral loci linked to selected regions, in principle, should provide an indirect means for estimating parameters of positive selection.

If a population experiences recurrent selective sweeps, there are several patterns predicted by theory. Under recurrent hard selective sweeps, levels of genetic diversity are expected to be lower i) in regions of the genome with restricted recombination, ii) in regions experiencing many sweeps and iii) in the genomic regions surrounding the targets of selection themselves. Each of these of these predictions have been met in empirical studies, and each has been used to estimate parameters of positive selection.

1.4.1 The Correlation Between Diversity and the Rate of Recombination *

In the late 1980s, evidence began to emerge suggesting that genetic polymorphism are less frequent in genomic regions experiencing restricted crossing-over (Aguade et al., 1989; Stephan and Langley, 1989). Soon after, Begun and Aquadro (1992) showed that there is a positive correlation between nucleotide diversity and the rate of crossing-over in *D. melanogaster*, a pattern subsequently observed in other eukaryotic species (Cutter and Payseur, 2013). Begun and Aquadro pointed out that the correlation is qualitatively consistent with the action of recurrent selective sweeps. Wiehe and Stephan (1993) formulated expressions, based on the correlation between nucleotide diversity and the rate of recombination, to estimate the compound parameter $\lambda 2N_e s$, where λ is the rate of sweeps per base pair per generation, N_e is the effective population size and s is the selection coefficient. They applied their method to the data of Begun and Aquadro (1992), estimating $\lambda 2N_e s = 5.37$ x 10^{-8}, but their method could not disentangle the individual parameters. More recently, Coop and Ralph (2012) performed a similar analysis in *D. melanogaster* to explore the effects of partial sweeps on parameter estimates. They showed that when partial sweeps are common, the rate of adaptive evolution is underestimated if the hard sweep model is assumed.

The correlation between diversity recombination observed by Begun and Aquadro (1992) can also be explained by background selection, the reduction in neutral diversity caused by the removal of linked deleterious mutations (Charlesworth et al., 1993). The process of background selection is qualitatively similar to recurrent selective sweeps, since both processes reduce local genetic diversity (Charlesworth, 2009) and skew the SFS towards rare variants (Braverman et al., 1995; Charlesworth et al., 1995a). Models of background selection envisage a neutral site linked to many functional sites at different distances, such that the effects of selection accumulate to reduce diversity (Hudson and Kaplan, 1995; Nordborg et al., 1996). The correlation between neutral diversity and the recombination rate predicted by background selection is quantitatively similar to that observed in *D. melanogaster* (Charlesworth, 1996). Indeed, recent studies suggest that background selection is a major determinant of nucleotide diversity variation at broad scales (>100Kbp) in humans McVicker et al. (2009) and *D. melanogaster* (Charlesworth, 2012; Comeron, 2014). It is clear, then, that background selection is a key confounding factor when attempting to make inferences about positive selection.

1.4.2 Correlation Between Neutral Diversity and Non-Neutral Divergence *

If there is a constant fraction of adaptive substitutions, α, across the genome for a given class of sites, regions that evolve at higher rates should experience a greater number of selective sweeps. Under a model of recurrent sweeps, it follows that there should be a negative correlation between nucleotide divergence at selected sites and diversity at linked neutral sites. This was first described in *Drosophila melanogaster* by Andolfatto (2007), and has been subsequently reported in other *Drosophila* species (Haddrill et al., 2011). Assuming a single rate of sweeps (λ) and a constant scaled strength of positive selection ($2N_es$) for a given class of sites, Andolfatto (2007) generalised formulae of Wiehe and Stephan (1993) based on the correlation between synonymous site diversity and non-synonymous site divergence to estimate $\lambda 2N_es = 3 \times 10^{-8}$ for the X-chromosome in *D. melanogaster*. Note that this $\lambda 2N_es$ estimate is similar to that obtained based on the correlation of synonymous site diversity and recombination rate (Wiehe and Stephan (1993); see above). Using an estimate of $\alpha = 0.50$ obtained from a MK-based analysis, Andolfatto (2007) decomposed the $\lambda 2N_es$ compound parameter, and inferred that $s \approx 0.001\%$ and $\lambda = 3.6x10^{-11}$ /bp/generation, suggesting that adaptation of protein-coding genes in *D. melanogaster* is driven by moderately weak selection (i.e., assuming *D. melanogaster* $N_e = 10^6$, $2N_es \approx 40$). In a related study, Macpherson et al. (2007) estimated $\lambda 2N_es \approx 10^{-7}$ in *D. simulans*, also by examining the correlation between mean neutral diversity and selected (nonsynonymous) divergence. However, their model also included the heterogeneity in levels of diversity, which is related to the rate and strength of sweeps in a different way to the mean, and allowed the individual parameters to be fitted by regression. The estimates of the compound parameter $\lambda 2N_es$ are similar between the two studies, though Macpherson et al. (2007) estimated that $s \approx 1\%$ (compared to Andolfattos estimate of $s \approx 0.001\%$) and $\lambda = 3.6 \times 10^{-12}$ /bp/generation. The discrepancies between the studies may be due to differences in biology between the species, or may reflect methodological differences: For example, if the majority of adaptive substitutions are driven by weakly selected sweeps, which will leave a relatively small signal in levels of polymorphism, the MK-based method may more sensitively detect them, perhaps explaining the higher rate of sweeps inferred by Andolfatto (2007). On the other hand, strongly selected sweeps will leave a larger footprint in levels of diversity, so will be more readily detected using the approach of Macpherson et al. (2007), perhaps explaining why they inferred a lower overall rate of sweeps, with higher selection coefficients (for a full description, see Sella et al. 2009). In both cases, inferences based on variation in polymorphism may reflect processes other than the fixation of adaptive alleles that have gone to fixation, such as partial sweeps and background selection, as these will affect patterns of diversity but not necessarily divergence. Related to this, the approach employed by Andolfatto (2007) has recently been extended by **?**, by estimating the correlation between synonymous site diversity and non-synonymous divergence in the presence of both background selection and gene conversion in *D. melanogaster*. They found that ignoring background selection tends to increase and decrease estimates of selection strength and rate, respectively. The parameter values estimated in their study suggest that 0.02% of new mutations at nonsynonymous sites are strongly selected ($s \approx 0.03\%$, assuming $N_e = 10^6$ for *D. melanogaster*).

1.4.3 Patterns of Diversity Around the Targets of Selection [*]

An individual hard selective sweep is expected to leave a trough in genetic diversity around the selected site. If a large proportion of amino acid substitutions are adaptive, as suggested by MK-type analyses (see Galtier 2016), collating patterns of diversity around all substitutions of a given type should reveal a trough in diversity. Such a pattern is not expected around a control class of sites, such as synonymous sites. This test, proposed by Sattath et al. (2011), was first applied it to *D. simulans*, and the above pattern was found. By fitting a hard sweeps model to the shape of the diversity trough, they estimated values of 5% and 13%, depending on whether one or two classes of beneficial mutational effects were fitted. Note that their estimates of are substantially lower than those obtained using MK-based methods for *D. melanogaster* Andolfatto 2007. Sattath et al. (2011) suggested that modes of selection other than hard sweeps may help explain to this discrepancy. However, even when modelling two classes of beneficial mutations, they found that amino acid substitutions are driven by strongly adaptive mutations ($s\ sim0.5\%$ and $s \sim 0.01\%$). Their estimates of selection strength are therefore in broad agreement with the estimate of s $sim1\%$ obtained by Macpherson et al. (2007), based on the correlation between synonymous diversity and non-synonymous divergence in *D. simulans*. The Sattath et al. (2011) test, then, suggests that adaptation in protein-coding genes is fairly frequent and driven by strong, hard sweeps.

The Sattath test has been applied in a variety of organisms, including humans (Hernandez et al., 2011), wild mice (Halligan et al., 2013), *Capsella grandiflora* (Williamson et al., 2014) and maize (Beissinger et al., 2016). In all but *C. grandiflora*, researchers have found no difference in patterns of diversity around selected and neutral substitutions. These results have been interpreted as evidence that hard sweeps were rare in the recent history of both humans (Hernandez et al., 2011) and maize (Beissinger et al., 2016). However, Enard et al. (2014) pointed out that the Sattath test will be underpowered if there is large variation in levels of functional constraint in the genome. Indeed, through their analyses Enard et al. (2014) found evidence for frequent adaptive substitutions in humans, particularly in regulatory sequence. To address the issues raised by Enard et al. (2014), Beissinger et al. (2016) applied the Sattath test to substitutions in maize genes with the highest and lowest levels of functional constraint separately, but still found no difference in diversity pattern, suggesting either that hard sweeps have been rare in that species or that there is another confounding factor.

One possible explanation is that the species in which the Sattath test did/did not detect hard sweeps have distinct patterns of linkage disequilibrium (LD). LD decays to background levels within hundreds of base-pairs in *D. simulans* (Andolfatto, 2007) and *C. grandiflora* (Josephs et al., 2015), whereas in humans, maize and wild mice it decays over distances closer to 10,000bp (Chia et al., 2012; Deinum et al., 2015; The 1000 Genomes Project Consortium et al., 2015). It may be, then, that the Sattath test is only applicable when there is relatively short-range LD, such that the patterns of diversity around selected substitutions do not substantially overlap with the analysis windows around neutral ones. If this were the case, interpreting the similarity in troughs of diversity around selected and neutral substitutions as evidence for a paucity of hard selective sweeps may not be justified in organisms where LD decays over distances of a similar order of magnitude as the width of the diversity troughs themselves.

1.5 Fitting genome wide patterns *

Methods to estimate the rate and strength of positive selection in the genome employ various combinations of nucleotide diversity, divergence, recombination rates and estimates of background selection effects as summary statistics, averaged over many regions of the genome. Recently, Elyashiv et al. (2016) developed a method that fits a model of hard sweeps and background selection to genome-wide variation in nucleotide diversity and divergence (at both selected and neutral sites). In *D. melanogaster*, they showed that hard sweeps can explain a large amount of genome-wide variation genetic diversity. For nonsynonymous sites, they found that $\alpha = 4.1\%$ for strongly selected mutations ($s \geq 0.03\%$) and $\alpha = 36.3\%$ for weakly selected mutations ($s \approx 0.0003\%$), summing to $\alpha = 40.4\%$, which is similar to the estimate obtained using the MK-test (Andolfatto, 2007). Their results suggest that accounting for weakly selected mutations may help reconcile the discrepancy between MK-based estimates of the rate and strength of selection and parameters estimated from sweep model predictions, described above.

Elyashiv et al. (2016) showed that a map of the effects of hard sweeps and background selection is capable of explaining a large amount of the variation in diversity across the genome, further demonstrating that the action of natural selection is pervasive, at least in *D. melanogaster*. However, their method overestimated the rate of deleterious mutations, which the authors attribute to the presence of modes of adaptation other than hard sweeps in *D. melanogaster*.

1.6 Halligan et al. 2013

The research I have performed throughout my PhD carries on from the work of Halligan et al. (2013), it is fitting, therefore, to give a brief description of the key findings from their study.

Halligan et al. (2013) sequenced the genomes of 10 wild-caught *Mus musculus castaneus* individuals, using high-throughput sequencing methods. They sequenced individuals to high coverage, using multiple libraries of Illumina paired-end reads. Based on an analysis of population structure, the individuals sequenced were thought to represent a single admixed group. Halligan et al. (2013) extracted polymorphism data from genomic regions around both protein-coding exons and conserved non-coding elements (CNEs - which were inferred to be involved in gene regulation), and found dips (or troughs) in average diversity surrounding the elements themselves. As discussed above, they then applied the Sattath analysis to their data, but found that there was no difference in the average diversity around nonsynonymous/synonymous substitutions. Halligan et al. (2013) modelled the contribution BGS made to the troughs in diversity around both protein-coding exons and CNEs using a population genetic model, but found that it could not fully explain the patterns of diversity that they had observed. Our understanding of the factors that shape nucleotide diversity across the mouse genome are, thus, somewhat unclear.

1.7 Thesis aims

The aim of this thesis is to further our understanding of the factors that shape variation in genetic diversity across the mammalian genome using the house mouse as a model. Particularly, I focus on the contributions of background selection and selective sweeps to variation in genetic diversity across the mouse genome.

- In Chapter 2, I leverage information from patterns of linkage disequilibrium across the mouse genome and construct recombination rate maps for the mouse genome.
- In Chapter 3, I estimate the distribution of fitness effects both advantageous and deleterious mutations for multiple classes of sites in the mouse genome. I use these estimates to parametrise forward-in-time population genetic simulations. These simulations are then analysed to scrutinise the parameters of selection we inferred.
- In Chapter 4, I fit a model incorporating the effects of both selective sweeps and background selection to troughs in diversity around functional elements in mice. Using the parameters that provide the best fit to the data, I ask whether adaptation in protein-coding or regulatory regions contributes most to fitness in mice.

Chapter 2

The recombination landscape in wild house mice inferred using population genomic data

This chapter has been published as a research paper in Genetics. I present this chapter as published, with minor alterations to the text. I designed the analyses with Rob W. Ness and Peter D. Keightley, I analysed the data and wrote the paper. Peter and Rob provided comments on previous versions of the manuscript

2.1 Abstract

Characterizing variation in the rate of recombination across the genome is important for understanding several evolutionary processes. Previous analysis of the recombination landscape in laboratory mice has revealed that the different subspecies have different suites of recombination hotspots. It is unknown, however, whether hotspots identified in laboratory strains reflect the hotspot diversity of natural populations or whether broad-scale variation in the rate of recombination is conserved between subspecies. In this study, we constructed fine-scale recombination rate maps for a natural population of the Eastern house mouse, *Mus musculus castaneus*. We performed simulations to assess the accuracy of recombination rate inference in the presence of phase errors, and we used a novel approach to quantify phase error. The spatial distribution of recombination events is strongly positively correlated between our *castaneus* map and a map constructed using inbred lines derived predominantly from *M. m. domesticus*. Recombination hotspots in wild *castaneus* show little overlap, however, with the locations of double-strand breaks in wild-derived house mouse strains. Finally, we also find that genetic diversity in *M. m. castaneus* is positively correlated with the rate of recombination, consistent with pervasive natural selection operating in the genome. Our study suggests that recombination rate variation is conserved at broad scales between house mouse subspecies, but it is not strongly conserved at fine scales.

2.2 Introduction

In many species, crossing-over events are not uniformly distributed across chromosomes. Understanding this variation and its causes is important for many aspects of molecular evolution. Experiments in laboratory strains or managed populations that examine the inheritance of markers through pedigrees have produced direct estimates of crossing-over rates in different genomic regions. Studies of this kind are impractical for many wild populations, however, because pedigrees are largely unknown (but see Johnston et al. 2016). In mice, there have been several genetic maps published (e.g., Jensen-Seaman et al. 2004; Paigen et al. 2008; Cox et al. 2009; Liu et al. 2014), typically using the classical inbred laboratory strains, which are predominantly derived from the Western European house mouse subspecies, *Mus musculus domesticus* (Yang et al. 2011). Recombination rate variation in laboratory strains may not, therefore, reflect rates and patterns in wild mice of other subspecies. In addition, recombination rate modifiers may have become fixed in the process of laboratory strain management. On the other hand, directly estimating recombination rates in wild house mice is not feasible without both a populations pedigree and many genotyped individuals (but see Wang et al. 2017).

Patterns of linkage disequilibrium (LD) in a sample of individuals drawn from a population can be used to infer variation in the rate of recombination across the genome. Coalescent-based methods have been developed to indirectly estimate recombination rates at very fine scales (Hudson 2001; McVean et al. 2002; McVean et al. 2004; Auton and McVean 2007; Chan et al. 2012). Recombination rates estimated in this way reflect long term variation in crossing-over in the populations history, and are averages between the sexes. Methods using LD have been applied to explore variation in recombination rates among mammals and other eukaryotes, and have demonstrated that recombination hotspots are associated with specific genomic features (Myers et al. 2010; Paigen and Petkov 2010; Singhal et al. 2015).

The underlying mechanisms explaining the locations of recombination events have been the focus of much research. In house mice and in most other mammals, the PRDM9 zinc-finger protein binds to specific DNA motifs, resulting in an increased probability of double-strand breaks (DSBs), which can then be resolved by reciprocal crossing-over or gene conversion (Grey et al. 2011; Baudat et al. 2013). Accordingly, it has been shown that recombination hotspots are enriched for PRDM9 binding sites (Myers et al. 2010; Brunschwig et al. 2012). PRDM9-knockout mice still exhibit hotspots, but in dramatically different genomic regions (Brick et al. 2012). Variation in PRDM9, specifically in the exon encoding the zinc-finger array, results in different binding motifs (Baudat et al. 2010). Davies et al. (2016) generated a line of mice in which the exon encoding the portion of the PRDM9 protein specifying the DNA binding motif was replaced with the orthologous human sequence. The recombination hotspots they observed in this humanized line of mice were enriched for the human PRDM9 binding motif.

Great ape species each have different PRDM9 alleles (Schwartz et al. 2014) and relatively little hotspot sharing (Winckler et al. 2005; Stevison et al. 2015). The broad-scale recombination landscapes of the great apes are, however, strongly positively correlated (Stevison et al. 2011; Stevison et al. 2015), suggesting that hotspots evolve rapidly, but that the overall genetic map changes more slowly. Indeed, broad-scale recombination rates are positively correlated between closely related species pairs with different hotspot locations (Smukowski and Noor 2011), and between species that share hotspots or lack them altogether (Singhal et al. 2015; Smukowski Heil et al. 2015).

It has been suggested that a population ancestral to the *M. musculus* subspecies complex split into the present-day subspecies around 350,000 years ago (Geraldes et al. 2011). In this time, functionally distinct PRDM9 alleles and distinct suites of hotspots evolved in the different subspecies (Smagulova et al. 2016). In addition, there is variation in the recombination rate at relatively broad scales across several regions of the genome between members of the *M. musculus* subspecies complex (Dumont et al. 2011), and recombination rates vary between recently diverged *M. m. domesticus* populations (Wang et al. 2017). Brunschwig et al. (2012) analysed single nucleotide polymorphism (SNP) data for classical laboratory strains of mice and used an LD-based approach to estimate the sex-averaged recombination landscape for the 19 autosomes. Their genetic map is similar to a genetic map generated using crosses by Cox et al. (2009). However, both studies were conducted using inbred lines whose ancestry is largely *M. m. domesticus* (Yang et al. 2011), so their recombination landscapes may be different from other members of the *M. musculus* subspecies complex.

In this study, we constructed genetic maps for the house mouse subspecies *M. m. castaneus*. We used the genome sequences of 10 wild-caught individuals of *M. m. castaneus* from the species assumed ancestral range, originally reported by Halligan et al. (2013). In our analysis, we first phased SNPs and estimated rates of error in phasing. Secondly, we simulated data to assess the power of estimating recombination rates based on only 10 individuals and the extent by which phase errors lead to biased estimates of the rate of recombination. Finally, using an LD-based approach, we inferred a sex-averaged genetic map and compared this to previously published maps for *M. musculus*. We show that broad-scale variation in recombination rates in *M. m. castaneus* is similar to that seen in the classical inbred strains. However, we show that the locations of potential recombination hotspots in *M. m. castaneus* exhibit little overlap with those reported in wild-derived laboratory strains.

2.3 Materials and Methods

2.3.1 Polymorphism data for *Mus musculus castaneus*

We analysed the genome sequences of 10 wild-caught *M. m. castaneus* individuals (Halligan et al. 2013). Samples were from North-West India, a region that is believed to be within the ancestral range of the house mouse. Mice from this region have the highest genetic diversity among the *M. musculus* subspecies (Baines and Harr 2007). In addition, the individuals sequenced showed little evidence for substantial inbreeding and a population structure analysis suggested that they represent a single population (Halligan et al. 2010). Halligan et al. (2013) sequenced individual genomes to high coverage using multiple libraries of Illumina paired-end reads, and mapped these to the mm9 reference genome using BWA (Li and Durbin 2009). Mean coverage was ¿20x and the proportion of the genome with ¿10x coverage was more than 80% for all individuals sampled (Halligan et al. 2013). Variants were called with the Samtools mpileup function (Li et al. 2009) using an allele frequency spectrum (AFS) prior. The AFS was obtained by iteratively calling variants until the spectrum converged. After the first iteration, all SNPs at frequencies ¿0.5 were swapped into the mm9 genome to construct a reference genome for *M. m. castaneus*, which was used for subsequent variant calling (for further details see Halligan et al. 2013). The variant call format (VCF) files generated by Halligan et al. (2013) were used in this study. In addition, alignments of *Mus famulus* and *Rattus norvegicus* to the mm9 genome, also generated by Halligan et al. (2013), were used as outgroups.

For the purpose of estimating recombination rates, variable sites were filtered on the basis of the following conditions. Insertion/deletion polymorphisms were excluded, because the method used to phase variants cannot process these sites. Sites at which more than two alleles segregated and those that failed the Samtools Hardy-Weinberg equilibrium test (p ¡ 0.002) were also excluded. The hypermutability of CpG sites violates the assumption of a single mutation rate. We defined sites as CpG-prone if they were preceded by a C or followed by a G in *M. m. castaneus*, *M. famulus* or *R. norvegicus.*

2.3.2 Inferring phase and estimating switch error rates

LDhelmet estimates recombination rates from a sample of phased chromosomes or haplotypes drawn from a population. To infer haplotypes, heterozygous SNPs called in *M. m. castaneus* were phased using read-aware phasing in ShapeIt2 (Delaneau et al. 2013), which phases variants at the level of whole chromosomes using sequencing reads that span multiple heterozygous sites (phase-informative reads, PIRs), and LD. Incorrectly phased heterozygous sites, termed switch errors, tend to upwardly bias estimates of the recombination rate, because they appear identical to legitimate crossing-over events. To assess the impact of incorrect phasing on recombination rate inference, we quantified the switch error rate as follows. The sample of *M. m. castaneus* comprised seven females and three males. The X-chromosome variants in males therefore represent perfectly phased haplotypes. We merged the BAM alignments of short reads for the X-chromosomes of the three males (samples H12, H28 and H34 from Halligan et al. (2013)) to make three datasets of pseudo-females where the true haplotypes are known (H12+H28 = H40; H12+H34 = H46; H28 + H34 = H62). We then jointly re-called variants in the seven female samples plus the three pseudo-females using an identical pipeline as Halligan et al. (2013), using the same AFS prior.

Switch error rates in Shapeit2 are sensitive both to coverage and quality (per genotype and per variant) (Delaneau et al. 2013). We explored the effects of different filter parameters on switch error rates using the X-chromosomes of the pseudo-females. We filtered SNPs based on combinations of variant and genotype quality scores (QUAL and GQ, respectively) and on an individuals sequencing depth (DP) (Table S1). For the individual-specific statistics (DP and GQ), if a single individual failed a particular filter, then that SNP was excluded from further analyses. By comparing the known X-chromosome haplotypes and those inferred by ShapeIt2, we calculated switch error rates as the ratio of incorrectly resolved heterozygous SNPs to the total number of heterozygous SNPs for each pseudo-female individual. We used these results to apply filter parameters to the autosomal data that generated a low switch error rate, while maintaining a high number of heterozygous SNPs. We obtained 20 phased haplotypes for each of the 19 mouse autosomes and 14 for the X-chromosome (plus the 3 from the male samples). With these, we estimated the recombination rate landscape for *M. m. castaneus.*

2.3.3 Estimating genetic maps and validation of the approach

LDhelmet (v1.7; Chan et al. 2012) generates a sex-averaged genetic map from a sample of haplotypes assumed to be drawn from a randomly mating population. Briefly, LDhelmet examines patterns of LD in a sample of phased chromosomal regions and uses a composite likelihood approach to infer recombination rates between adjacent SNPs. LDhelmet appears to perform well for species of large effective population size (N_e) and has been shown to be robust to the effects of selective sweeps, which appear to reduce diversity in and around functional elements of the *M. m. castaneus* genome (Halligan et al. 2013). The analyses of Chan et al. (2012), in which the software was tested, were performed with a larger number of haplotypes than we have in our sample. To assess whether our smaller sample size still gives reliable genetic maps, we validated and parameterized LDhelmet using simulated datasets (see below). It should be noted, however, that model underlying LDhelmet assumes recombination-drift equilibrium. Violation of this assumption may therefore result in biased recombination rate estimates.

A key parameter in LDhelmet is the block penalty, which determines the extent by which likelihood is penalized by spatial variation in the recombination rate, such that a high block penalty results in a smoother recombination map. We performed simulations to determine the block penalty that produces the most accurate estimates of the recombination rate in chromosomes that have diversity and base content similar to *M. m. castaneus*. Chromosomes with constant values of ρ ($4N_e r$) ranging from 2 x 10^{-6} to 2 x 10^{1} were simulated in SLiM v1.8 (Messer 2013). For each value of ρ, 0.5Mbp of neutrally evolving sequence was simulated for populations of N = 1,000 diploid individuals. Mutation rates in the simulations were set using the compound parameter $\theta = 4N_e\mu$, where μ is the per-base, per-generation mutation rate. The mutation and recombination rates of the simulations were scaled to $\theta/4N$ and $\rho/4N$, respectively. θ was set to 0.01 in the simulations, because this value is close to the genome-wide average for our data, based on pairwise differences. Simulations were run for 10,000 generations in order to achieve equilibrium diversity, at which time 10 diploid individuals were sampled. Each simulation was repeated 20 times, resulting in 10Mbp of sequence for each value of . The SLiM output files were converted to sequence data suitable for analysis by LDhelmet using a custom Python script that incorporated the mutation rate matrix estimated for non-CpG prone sites in *M. m. castaneus* (see below). Following (Chan et al. 2012), we inferred recombination rates from the simulated data in windows of 4,400 SNPs with a 200 SNP overlap between windows. We analysed the simulated data using LDhelmet with block penalties of 10, 25, 50 and 100. The default parameters of LDhelmet are tuned to analyze Drosophila melanogaster data (Chan et al. 2012). Since the *D. melanogaster* population studied by Chan et al. (2012) has comparable nucleotide diversity to *M. m. castaneus*, we used default values for other parameters, with the exception of the block penalty.

Errors in phase inference, discussed above, may bias our estimates of the recombination rate, since they appear to break apart patterns of LD. We assessed the impact of these errors on recombination rate inference by incorporating them into the simulated data at a rate estimated from the pseudo-female individuals. For each of the 10 individuals drawn from the simulated populations, switch errors were randomly introduced at heterozygous positions at the rate estimated using the SNP filter set chosen on the basis of the pseudo-female analysis (see Results). We then inferred recombination rates for the simulated population using these error-prone data, as above. We assessed the effect of switch errors on recombination rate inference by comparing estimates from the simulated data with and without switch errors. It is worth noting that switch errors may undo crossing-over events and thereby reduce inferred recombination rates if they affect heterozygous SNPs located at recombination breakpoints.

2.3.4 Recombination rate estimation for *M. m. castaneus*

We used LDhelmet (Chan et al. 2012) to estimate recombination rate landscapes for each of the *M. m. castaneus* autosomes and the X-chromosome. A drawback of LD-based approaches is that they estimate sex-averaged recombination rates. This is a limitation of our study as there are known differences in recombination rates between the sexes in *M. musculus* (Cox et al. 2009; Liu et al. 2014).

We used *M. famulus* and *R. norvegicus* as outgroups to assign ancestral states for polymorphic sites. LDhelmet incorporates the mutation matrix and a prior probability on the ancestral allele at each variable position as parameters in the model. We obtained these parameters as follows. For non-CpG prone polymorphic sites, if the two outgroups shared the same allele, we assigned that allele as ancestral, and such sites were then used to populate the mutation matrix (Chan et al. 2012). This approach ignores the possibility of back mutation and homoplasy. To account for this uncertainty, LDhelmet incorporates a prior probability on the ancestral base. Following Singhal et al. (2015), at resolvable sites (i.e., where both outgroups agreed) the ancestral base was given a prior probability of 0.91, with 0.03 assigned to each of the three remaining bases. This was done to provide high confidence in the ancestral allele, but also to include the possibility of ancestral allele misinference. At unresolved sites (i.e., if the outgroups disagreed or there were alignment gaps in either outgroup), we used the stationary distribution of allele frequencies from the mutation rate matrix as the prior (Table S2).

We analysed a total of 44,835,801 SNPs in LDhelmet to construct genetic maps for the *M. m. castaneus* autosomes and the X-chromosome. Following Chan et al. (2012), windows of 4,400 SNPs, overlapping by 200 SNPs on either side were analysed. We ran LDhelmet for a total of 1,000,000 iterations, discarding the first 100,000 as burn-in. A block penalty of 100 was chosen to obtain conservatively estimated broad-scale genetic maps. For the purposes of identifying recombination hotspots, we re-ran the LDhelmet analysis with a block penalty of 10. We analysed all sites that passed the filters chosen using the pseudo-female phasing analysis regardless of CpG status; note that excluding CpG-prone sites removes 50% of the available data and thus would substantially reduce the power to infer recombination rates. We assumed $\theta = 0.01$, the approximate genome-wide level of neutral diversity in *M. m. castaneus*, and included ancestral allele priors and the mutation rate matrix for non-CpG sites as parameters in the model. Following the analyses, we removed overlapping SNPs and concatenated SNP windows to obtain recombination maps for whole chromosomes.

It is worthwhile noting that our genetic maps were constructed with genotype calls made using the mm9 version of the mouse reference genome. This version was released in 2007 and there have been subsequent versions released since then. However, previously published genetic maps for *M. musculus* were constructed using mm9, so we used that reference to make comparisons (see below).

2.3.5 Broad-scale comparison to previously published maps

We compared the *M. m. castaneus* genetic map inferred using a block penalty of 100 with two previously published maps for *M. musculus.* The first map was generated by analyzing the inheritance patterns of markers in crosses between inbred lines (Cox et al. 2009) (downloaded from http://cgd.jax.org/mousemapconverter/). We refer to this map as the Cox map. The second map was generated by Brunschwig et al. (2012) by analyzing SNPs in classical inbred mouse lines using LDhat (Auton and McVean 2007), the software upon which LDhelmet is based (available at http://www.genetics.org/content/early/2012/05/04/genetics.112.141036). We refer to this map as the Brunschwig map. The Cox and Brunschwig maps were constructed using far fewer markers than the present study, i.e., 500,000 and 10,000 SNPs, respectively, compared to the 45,000,000 used to generate ours. Recombination rate variation in the Cox and Brunschwig maps likely reflects that of *M. m. domesticus*, since both were generated using classical strains of laboratory mice, which are predominantly of *M. m. domesticus* origin (Yang et al. 2011). For example, in the classical inbred strains analysed by Cox et al. (2009), the mean genome-wide ancestry attributable to *M. m. domesticus*, *M. m. musculus* and *M. m. castaneus* are 94.8%, 5.0% and 0.2%, respectively (data downloaded from the Mouse Phylogeny Viewer (Wang et al. 2012) http://msub.csbio.unc.edu). The ancestry proportions for all classical strains, 60 of which were analysed by Brunschwig et al. (2012), are similar (Yang et al. 2011).

Recombination rates in the Brunschwig map and our castaneus map were estimated in units of $\rho = 4N_er$. For comparison purposes, we converted these units to cM/Mb using frequency-weighted means, as follows. LDhat and LDhelmet provide estimates of ρ (per Kbp and bp, respectively) between pairs of adjacent SNPs. For each chromosome, we calculated cumulative ρ, while accounting for differences in the physical distance between adjacent SNPs by using the number of bases separating a pair of SNPs to weight that pairs contribution to the total. By setting the total map length for each chromosome to that of Cox et al. (2009), we converted the cumulative at each analysed SNP position to cM values.

At the level of whole chromosomes, we compared mean recombination rate estimates for castaneus with several previously published maps. Frequency-weighted mean recombination rates (in terms of ρ) for each chromosome in the castaneus and Brunschwig maps were compared with cM/Mb values obtained by Cox et al. (2009) and with independent estimates of per chromosome recombination rates (Jensen-Seaman et al. 2004). Pearson correlations were calculated for each comparison

At the Mbp scale, we compared variation in recombination rates across the autosomes in the different maps using windows of varying length. We calculated Pearson correlations between the frequency weighted-mean recombination rates (in cM/Mb) in non-overlapping windows of 1Mbp to 20Mbp for the castaneus, Cox and Brunschwig maps. For visual comparison of the castaneus and Cox maps, we plotted recombination rates in sliding windows of 10Mbp, offset by 1Mb.

2.3.6 Fine-scale recombination rate variation

To assess the distribution of recombination events in *M. m. castaneus* on a fine scale, we used Gini coefficients and Lorenz curves as quantitative measures of the extent of heterogeneity (e.g., Kaur and Rockman 2014). In the context of a genetic map, Gini coefficients close to zero represent more uniform distributions of crossing-over rates, whereas values closer to one indicate that recombination events are restricted to a small number of locations. We analysed genetic maps generated using a block penalty of 10 to construct Lorenz curves and calculated their Gini coefficients for each chromosome separately.

Recombination hotspots can be operationally defined as small windows of the genome that exhibit elevated rates of recombination relative to surrounding regions. To estimate the locations of potential recombination hotspots, we adapted a script used by Singhal et al. (2016). We divided the genome into non-overlapping windows of 2Kbp, and, using the maps generated with a block penalty of 10, classified as putative hotspots all windows where the recombination rate was at least 5x greater than the recombination rate in the surrounding 80Kbp. Recombination hotspots may be wider than 2Kbp, so neighbouring analysis windows that exhibited elevated recombination rates were merged.

We investigated whether fine-scale recombination rate variation in wild-caught *M. m. castaneus* is similar to that reported for wild-derived inbred lines. Smagulova et al. (2016) generated sequencing reads corresponding to the locations of DSBs (hereafter DSB hotspots) in inbred strains of mice derived from each of the principal *M. musculus* subspecies and M. m. molossinus, an inter-sub-specific hybrid of *M. m. castaneus* and *M. m. musculus*. We used the overlap between our putative hotspots and their DSB hotspots for testing similarity. However, the coordinates of DSB hotspots were reported with respect to the mm10 genome (Smagulova et al. 2016). To allow comparisons with our putative hotspots, we converted the coordinates of DSB breaks in the mm10 reference to mm9 coordinates using the UCSC LiftOver tool (https://genome.ucsc.edu/cgi-bin/hgLiftOver), with default parameters. We compared the locations of putative hotspots identified in our castaneus map with the locations of DSB hotspots using BedTools v2.17.0 (Quinlan and Hall 2010) by counting the number that overlapped. To determine the number of overlaps expected to be seen by chance, we used a randomization approach as follows. The locations of our putative hotspots were randomized with respect to chromosome, and these shuffled coordinates were compared to the locations of DSB hotspots. For each of the inbred strains analysed by Smagulova et al. (2016) this procedure was repeated 1,000 times. The maximum number of overlapping DSB and putative castaneus hotspots observed across all 1,000 replicates was taken as an approximate 0.1% significance threshold.

2.3.7 Examining the correlation between recombination rate and properties of protein coding genes

We used our castaneus map to examine the relationship between recombination rates and nucleotide diversity and divergence as follows. We obtained the coordinates of the canonical spliceforms of protein coding genes, orthologous between mouse and rat from Ensembl Biomart (Ensembl Database 67; http://www.ensembl.org/info/website/archives/index.html). For each protein coding gene, we calculated the frequency-weighted mean recombination rate from the broad-scale map. Using the approximate castaneus reference described above, along with the outgroup alignments, we obtained the locations of 4-fold degenerate synonymous sites and current GC content for each gene. If a site was annotated as 4-fold in all three species considered, it was used for further analysis. We removed poor quality alignments between mouse and rat that exhibited spurious excesses of mismatched sites, where ¿80% of sites were missing. We also excluded five genes where there were mismatches with the rat sequence at all non-CpG prone 4-fold sites, since it is likely that these also represent incorrect alignments. After filtering, there were a total of 18,171 protein-coding genes for analysis.

We examined the correlation between the local recombination rate in protein-coding genes and nucleotide diversity, divergence from the rat and GC-content. Variation in the mutation rate across the genome is a potentially important confounding factor. For example, if the recombination rate and mutation rate are positively correlated, we would expect a positive correlation between neutral nucleotide diversity and recombination rate. Because of this, we also examined the correlation between the ratio of nucleotide diversity to divergence from R. norvegicus at putatively neutral sites and the rate of recombination. We calculated correlations for all sites and for non-CpG-prone sites only. We used non-parametric Kendall rank correlations for all comparisons.

Analyses were conducted using Python scripts, except for the correlation analyses, which were conducted using R (R Core Team 2016) and hotspot identification, which was done using a Python script adapted from one provided by Singhal et al. (2016).

2.3.8 Data availability

The authors confirm that all data necessary for performing the analyses described in the article are fully described in the text. Recombination maps are available in a compressed form from $https : //github.com/TBooker/M.m.castaneus_{r}ecombination-maps$.

2.4 Results

2.4.1 SNP phasing and estimating the switch error rate

To infer genetic maps using our sample of individuals, we required phased SNPs. Taking advantage of the high sequencing depth of the sample generated by Halligan et al. (2013), and using a total of 44,835,801 SNPs (Table S3), we phased SNPs using ShapeIt2, an approach that uses LD and sequencing reads to resolve haplotypes.

We quantified the switch error rate incurred when inferring phase by analyzing pseudo-female individuals. After filtering variants, ShapeIt2 returned low switch error rates for all parameter combinations tested (Table S1). We therefore applied a set of filters (GQ ¿ 15, QUAL ¿ 30) to apply to the actual data that predicted a mean switch error rate of 0.46% (Table S1). When applied to the actual data these filters removed

44% of the total number of called SNPs (Table S3). More stringent filtering resulted in slightly lower mean switch error rates, but also removed many more variants (Table S1), reducing our ability to estimate recombination rates at a fine scale.

2.4.2 Simulations to validate the application of LDhelmet

We used simulations to assess the performance of LDhelmet when applied to our dataset. In the absence of switch errors, LDhelmet accurately inferred the average recombination rate down to values of ρbp^{-1} = 2 x 10^{-4}. Below this value, LDhelmet overestimated the scaled recombination rate (Figure 1). With switch errors incorporated into simulated data, LDhelmet accurately estimated ρ/bp in the range 2 x 10^{-3} to 2 x 10^{2}. When the true ρbp^{-1} was ¡ 2 x 106−3, however, LDhelmet overestimated the mean recombination rate for 0.5Mbp regions (Figure 1). This behavior was consistent for all block penalties tested (Figure S1). We found that inferred rates of recombination typically fell within the range accurately estimated by LDhelmet (Table 1; Figure S2).

2.4.3 Recombination rates in the *M. m. castaneus* genome

We constructed two maps of recombination rate variation for *M. m. castaneus* using LDhelmet. The first was a broad-scale map constructed using a block penalty of 100 (hereafter referred to as the broad-scale map). For the second fine-scale map, we used a block penalty of 10 (hereafter referred to as the fine-scale map). A comparison of broad and fine-scale maps for a representative region of the genome is shown in Figure S2. We analysed a total of 44,835,801 phased SNPs across the 19 mouse autosomes and the X-chromosome. From the broad-scale map, the frequency-weighted mean estimate of ρ/bp for the autosomes was 0.0092. This value is higher than the lower detection limit suggested by the simulations with and without switch errors (Figure 1). For the X-chromosome, the frequency-weighted mean ρ/bp was 0.0026, which is still above the lower detection limit (Figure 1). The lower SNP density on the X-chromosome (Table S3) and the smaller number of alleles available (17 compared to 20 used for the autosomes) may reduce precision.

We assessed variation in whole-chromosome recombination rates between our LD-based castaneus map and direct estimates of recombination rates published in earlier studies. Comparing the mean recombination rates of whole chromosomes provides us with a baseline for which we have two a priori expectations. Firstly, we expect that chromosome 19, the shortest in physical length, should have the highest mean recombination rate, since at least one crossing-over event is required per meiosis per chromosome. Secondly, we expect that the X-chromosome, which only undergoes recombination in females, should have the lowest rate. These expectations are borne out in the results (Table 1), and are consistent with previous studies (Jensen-Seaman et al. 2004; Cox et al. 2009). We also found that frequency-weighted chromosomal recombination rates (inferred in terms of $\rho = 4N_er$) were highly correlated with the direct estimates (in cM/Mbp) from Jensen-Seaman et al. (2004) (Pearson correlation coefficient = 0.59, $p = 0.005$) and Cox et al. (2009) (Pearson correlation coefficient = 0.68, $p = 0.001$). Excluding the X-chromosomes does not substantially change these correlations. These results therefore suggest that our analysis captures real variation in the rate of recombination on the scale of whole chromosomes.

2.4.4 Comparison of the *M. m. castaneus* map with maps constructed using inbred lines

We then compared intra-chromosomal variation in recombination rates between our broad-scale castaneus map and previously published maps. Figure 2 shows a comparison of recombination rates inferred from the castaneus and Cox maps for the longest and shortest autosomes, chromosomes 1 and 19, respectively. It is clear that the castaneus and Cox maps are very similar (see also Figure S3). We compared recombination rates in the castaneus and Cox maps in genomic intervals of various sizes and found that correlation coefficients were ¿0.8 for window sizes of 8Mbp and above (Figure 3). The correlations are smaller if chromosomes are considered separately (Figure S4). Although the correlation coefficients are generally high (Figure 3), there are several regions of the genome where the castaneus and Cox maps have substantially different recombination rates, for example in the center of chromosome 9 (Figure S3). The Cox and castaneus maps are more similar to one another than either are to the Brunschwig map (Figure 3). This is presumably because the Brunschwig map was constructed with a relatively low SNP density and by an LD-based approach using a sample of inbred mouse strains, which violates key assumptions of the method. Population structure in the lines analysed by Brunschwig et al. (2012) or the subspecies from which they were derived would elevate LD, resulting in lower chromosome-wide values of ρ. The average scaled recombination rate estimates differ substantially between the castaneus and Brunschwig maps, i.e., the castaneus chromosomal estimates are 500x higher (Table 1). This is also reflected in N_e, estimated on the basis of the frequency-weighted average recombination rates for each chromosome. Independent polymorphism data suggest that effective populations sizes for *M. m. castaneus* and *M. m. domesticus* are approximately 100,000 and 500,000 respectively (Geraldes et al. 2008; Geraldes et al. 2011). Estimates of N_e from the castaneus map are therefore in line with expectation, while those from the Brunschwig map are not (Table 1).

2.4.5 Analysis of fine-scale recombination rates

To locate potential recombination hotspots in wild *M. m. castaneus*, we generated a fine-scale map, from which we identified 39,972 potential recombination hotspots. For each chromosome, there was an average of 15 hotspots per Mbp. The total number of putative hotspots is more than twice the number identified in CAST/EiJ, an inbred strain derived from wild *M. m. castaneus* (Smagulova et al. 2016).

To obtain a measure of the amount of fine-scale recombination rate heterogeneity across the genome, we constructed Lorenz curves and calculated their Gini coefficients (Figure S5). The mean Gini coefficient for all chromosomes was 0.78. This estimate is very similar to that of Kaur and Rockmans (2014) median Gini coefficient of 0.77 for chromosome 1, obtained from a high-density map of crossing-over locations in inbred mice (Paigen et al. 2008). The Gini coefficients calculated from our fine-scale map suggest that the distribution of recombination rates in wild and inbred mice are similarly heterogeneous. However, the Lorenz curve for the X-chromosome is clearly distinct from that of the autosomes (Figure S5), and its Gini coefficient is 0.95.

There was only a small amount of overlap between the locations of putative recombination hotspots we identified in wild castaneus and the locations of DSB hotspots observed in wild-derived inbred strains (Smagulova et al. 2016) (Table S4). As may be expected, DSB hotspots in the inbred strain derived from *M. m. castaneus* (CAST) exhibited the greatest amount of overlap with the locations of recombination hotspots identified in *M. m. castaneus.* Of all DSB hotspots in CAST, 12.2% (or 4.1% after correcting for the null expectation) overlapped with one of the putative hotspots we identified. Such a low proportion strongly suggests that, even within the *M. m. castaneus* subspecies, the locations of recombination hotspots are highly variable. The PWD strain, which was derived from wild *M. m. musculus*, exhibited the second highest amount of overlap. Less than 1% of the DSB hotspots in each of the three strains derived from *M. m. domesticus* overlapped with putative hotspots in *M. m. castaneus*, after correcting for the number of overlaps expected to be seen by chance. Table S4 shows the overlap for each of the strains analysed by Smagulova et al. (2016).

2.4.6 Correlation between recombination rate and properties of protein coding genes

There is evidence of pervasive natural selection acting in protein-coding genes and conserved non-coding elements of the murid genome (Halligan et al. 2010; Halligan et al. 2011; Halligan et al. 2013). This is expected to reduce diversity at linked neutral sites via background selection and/or selective sweeps, and is therefore expected to generate a positive correlation between diversity and recombination rate, as has been observed in multiple species (Cutter and Payseur 2013).

We examined the correlation between genetic diversity and recombination rate to determine whether our map captures variation in N_e across the genome. We found that the rate of recombination at autosomal protein-coding genes is significantly and positively correlated with genetic diversity of putatively neutral sites (Table 2). Furthermore, the correlation between recombination rate and neutral diversity scaled by divergence (from the rat) was both positive and significant, regardless of base context (Table 2; Figure S6). This indicates that natural selection may have a role in reducing diversity via hitchhiking and/or background selection.

Biased gene conversion can influence levels of between-species nucleotide substitution (Duret and Galtier 2009). GC-biased gene conversion (gcBGC), where G/C alleles are preferentially chosen as the repair template following double-strand breaks, can generate a positive correlation between nucleotide divergence and recombination rate (Duret and Arndt 2008). Gene conversion occurs whether or not a DSB is resolved by crossing-over (Duret and Galtier 2009) and models of gcBGC predict an increase in the rates of nucleotide substitution in regions of high crossing-over (Duret and Arndt 2008). Indeed, human-chimp divergence is positively correlated with rates of crossing-over when considering all base contexts. Consistent with this, we found that 4-fold site nucleotide divergence was significantly positively correlated with recombination rate for the case of all sites (Table 2). In the case of non-CpG-prone sites, however, we found only a weak negative correlation (Table 2). A recent study by Phung et al. (2016) found a positive correlation between human-chimp divergence and recombination rate that persisted after removing CpG-prone sites, so further study is required to analyze the effects of gene conversion on patterns of divergence in mice.

2.5 Discussion

Our analyses suggest that the recombination landscapes of wild house mice and their laboratory counterparts are similar at broad-scales, but are dissimilar at fine-scales. Our broad-scale map captures variation in the recombination rate similar to that observed in a more traditional linkage map, both at the level of whole chromosomes and genomic windows of varying sizes. However, we found that a relatively small proportion of double-strand break (DSB) hotspots identified in wild-derived strains (Smagulova et al. 2016) overlapped with putative recombination hotspots in *M. m. castaneus.* This suggests that recombination rates are highly variable within and between the subspecies at the kilobase scale. We discuss potential reasons for this below.

Recombination landscapes inferred using coalescent approaches, as in this study, reflect ancestral variation in recombination rates. In *M. m. castaneus*, we have shown that this ancestral variation is highly correlated with contemporary recombination rate variation in inbred mice derived from *M. m. domesticus*, suggesting that the broad-scale genetic map has not evolved substantially since the subspecies shared a common ancestor, around 350,000 years ago (Geraldes et al. 2011). At a finer scale however, there is considerable variation in the locations of recombination hotspots between the *M. musculus* subspecies. This was also observed in studies of the great-apes, which suggested that the locations of recombination hotspots have strongly diverged between species, but that broad-scale patterns are relatively conserved (Lesecque et al. 2014; Stevison et al. 2015). There are, however, several relatively large regions of the genome showing substantially different recombination rates between our *M. m. castaneus* map and the Cox map. For example, there are recombination rate peaks in *M. m. castaneus* on chromosomes 4, 5, 14 and 15, which are not present in the Cox map (Figure S3). Directly estimating recombination rates at fine scales in *M. m. castaneus* individuals could potentially reveal whether the broad-scale differences in recombination rate, mentioned above, are present in modern day populations.

The positive correlation between the castaneus map and the Cox map (constructed using a pedigree-based approach) is weaker for the X-chromosome than for autosomes of similar physical length (e.g., chromosomes 2 and 3)(Figure S4). However, SNP density on the *M. m. castaneus* X-chromosome is substantially lower than the autosomes (Table S3). Greater physical distance between adjacent SNPs restricts the resolution of recombination rates in the coalescent-based approach. Thus, in our study, recombination rates are resolved at finer scales on the autosomes than on the X-chromosome. Additionally, we inferred recombination rates on the X-chromosome using 17 gene copies rather than the 20 used for the autosomes. Our findings are consistent, however, with the results of Dumont et al. (2011), who constructed linkage maps in *M. m. castaneus* and *M. m. musculus* (both by crossing with *M. m. domesticus*) using a small number of markers. In that study, the authors found multiple genomic intervals that significantly differed in genetic map distance between the two subspecies, and a disproportionate number of differences were on the X-chromosome. Thus, their results and ours suggest that the recombination landscape of the X-chromosome has evolved faster than that of the autosomes.

A recent study by Stevison et al. (2015) examined pairs of great ape species, and found that correlations between recombination maps (at the 1Mbp scale) declined with genetic divergence. For example, between humans and gorillas, genetic divergence is 1.4%, while the Spearman-rank correlation of their respective recombination rate maps is 0.5. Genetic divergence between *M. m. castaneus* and *M. m. domesticus* is reported to be 0.5% (Geraldes et al. 2008), and we find a Spearman-rank correlation of 0.47 between the castaneus map and the Cox map, also at the 1Mbp scale. Although this is only a single data point, it suggests that recombination rate differences may have accumulated faster relative to divergence between *M. m. castaneus* and *M. m. domesticus* than they have between great ape species. The recombination maps constructed for the great apes by Stevison et al. (2015) were all generated using the same methodology, which is not the case for the comparison we make between our map and that of Cox et al. (2009), so quantitative comparisons between the studies should be treated with caution. Performing a comparative analysis of recombination rates in the different subspecies of house mice and related mouse species (for example, Mus caroli and Mus spretus) using LD-based methods may help us understand whether the rate of evolution of the recombination landscape in wild mice is more rapid than in the great apes.

The locations of the vast majority of recombination hotspots in mice are directed by the binding of the PRDM9 protein (Brick et al. 2012), and there are unique landscapes of DSB hotspots associated with the different PRDM9 alleles present in different wild-derived inbred strains (Smagulova et al. 2016). However, in natural populations there is a great diversity of PRDM9 alleles in each of the *M. musculus* subspecies (Kono et al. 2014), therefore the binding motif will vary, causing different suites of hotspot locations. Thus, the DSB hotspot maps obtained by Smagulova et al. (2016) likely represent a fraction of the diversity of hotspot locations in wild *M. musculus* populations. Indeed, we found that only 12% of the DSB hotspots reported for CAST/EiJ by Smagulova et al. (2016) overlapped with hotspots we inferred for *M. m. castaneus* (Table S4). However, the mean Gini coefficient we estimated for *M. m. castaneus* was almost identical to the value obtained by Kaur and Rockman (2014) from crossing-over data of *M. musculus.* This similarity suggests that while the locations of hotspots may differ, the distribution of recombination rates is similarly heterogeneous in wild and inbred mice.

The *castaneus* map constructed in this study appears to be more similar to the

Cox map than the Brunschwig map (Figure 3). There are number of potential reasons for this. Firstly, we used a much larger number of markers to resolve recombination rates than Brunschwig et al. (2012). Secondly, it seems probable that population structure within and between the inbred and wild-derived lines studied by Brunschwig et al. (2012) could have resulted in biased estimates of the recombination rate. The Brunschwig map does, however, capture true variation in the recombination rate, since their map is also highly correlated with the Cox map (Pearson correlation ¿0.4) for all genomic windows wider than 8Mbp (Figure 3). Indeed, Brunschwig et al. (2012) showed by simulation that hotspots are detectable by analysis of inbred lines, and validated their hotspots against the locations of those observed in crosses among classical strains of *M. m. domesticus* (Smagulova et al. 2011). This suggests that while estimates of the recombination rate in the Brunschwig et al. (2012) map may have been downwardly biased by population structure (see above), variation in the rate and locations of hotspots were still accurately detected.

By simulating the effect of switch errors on estimates of the recombination rate, we inferred the range over which ρ/bp is accurately estimated. Switch errors appear identical to legitimate crossing-over events and, if they are randomly distributed along chromosomes, a specific rate of error will resemble a constant rate of crossing-over. The rate of switch error will then determine a detection threshold below which recombination cannot be accurately inferred. We investigated this detection threshold by introducing switch errors, at random, into simulated data at the rate we estimated using the X-chromosome. We found that, in the presence of switch errors, LDhelmet consistently overestimates the recombination rate when the true value is below 2 x 10^{-3} ρ/bp (Figure 1; Figure S1). This highlights a possible source of bias affecting LD-based recombination mapping studies that use inferred haplotypes, and suggests that error in phase inference needs to be carefully considered.

We obtained an estimate of the switch error rate, using a novel approach that took advantage of the hemizygous sex chromosomes of males. This allowed us to assess the extent by which switch errors affected our ability to infer recombination rates. Our inferred switch error rate may not fully represent that of the autosomes, however, because multiple factors influence the ability to phase variants (i.e., LD, SNP density, sample size, depth of coverage and read length), and some of these factors differ between the X-chromosome and the autosomes. The sex-averaged recombination rate for the X-chromosome is expected to be 3/4 that of the autosomes, so it will likely have elevated LD, and thus there will be higher power to infer phase. In contrast, X-linked nucleotide diversity in *M. m. castaneus* is approximately one-half that of the autosomes (Kousathanas et al. 2014), so there would be a higher number of phase informative reads on the autosomes. While it is difficult to assess whether the switch error rates we estimated from the X-chromosome will be similar to those on the autosomes, the analysis allowed us to explore the effects of different SNP filters on the error rate.

Consistent with studies in a variety of organisms (Cutter and Payseur 2013), we found a positive correlation between genetic diversity at putatively neutral sites and the rate of recombination. Both unscaled nucleotide diversity and diversity divided by divergence between mouse and rat, a proxy for the mutation rate, are positively correlated with the recombination rate (Table 2). Cai et al. (2009) found evidence suggesting that recombination may be mutagenic, although insufficient to account for the correlations they observed. The Kendall correlation between /drat and recombination rate is 0.20 for all 4-fold sites (Table 2), which is similar in magnitude to the corresponding value of 0.09 reported by Cai et al. (2009) in humans. The correlations we report may be downwardly biased, however, because switch errors may result in inflated recombination rates for genomic regions where the recombination rate is low (see above). Genes that have recombination rates lower than the detection limit set by the switch error rate may be reported as having inflated ρ/bp (Figure 1; Figure S1), and this would have the effect of reducing correlation statistics. It is difficult to assess the extent of this bias, however, and in any case the correlations we observed between diversity and recombination suggest that our recombination map does indeed capture real variation in N_e across the genome. This indicates that a recombination mediated process influences levels of genetic diversity. Previously, Halligan et al. (2013) showed that there are reductions in nucleotide diversity surrounding protein coding exons in *M. m. castaneus*, characteristic of natural selection acting within exons reducing diversity at linked sites. Their results and ours suggest pervasive natural selection in the *M. m. castaneus* genome. In contrast, a previous study in wild mice found that, while *M. m. musculus* exhibited a significant correlation between diversity and recombination, the relationship was non-significant for both *M. m. castaneus* and *M. m. domesticus* (Geraldes et al. 2011). This study analysed only 27 loci, so was perhaps underpowered to detect a relatively weak correlation. It should be noted, however, that the measure of recombination rate we used (ρ/bp) and neutral genetic diversity are both functions of the effective population size, so the positive correlation we detected could be partly driven by random fluctuations of N_e across the genome.

Furthering our understanding of the evolution of the recombination landscape in house mice would be helped by comparing fine-scale rates in the different subspecies. In this study we have assumed that inbred lines derived from *M. m. domesticus* reflect natural variation in recombination rates in that subspecies, though this is not necessarily the case. Directly comparing natural population samples of the different subspecies may help reconcile several potentially conflicting results. For example, the hotspots we detected in our study show more overlap with *M. m. musculus* than with *M. m. domesticus*, based on the DSB hotspots reported by Smagulova et al. (2016). However, overall rates of crossing-over in male *M. m. musculus* are higher than in either *M. m. castaneus* or *M. m. domesticus* (Dumont and Payseur 2011). Additionally, there is evidence of recombination rate modifiers of large effect segregating within *M. m. musculus* populations (Dumont et al. 2011). So, although overall rates of crossing-over in *M. m. musculus* are higher than in the other species, its recombination landscape may be more similar to *M. m. castaneus* than to *M. m. domesticus*. A broad survey comparing recombination rate landscapes in the different subspecies of mice would most efficiently be performed using LD-based approaches.

In conclusion, we find that sex-averaged estimates of the ancestral recombination landscape for *M. m. castaneus* are highly correlated with contemporary estimates of the recombination rate observed in crosses of inbred lines that predominantly reflect *M. m. domesticus* (Cox et al. 2009). It has previously been demonstrated that the turnover of hotspots has led to rapid evolution of fine-scale rates of recombination in the *M. musculus* subspecies complex (Smagulova et al. 2016) and our results suggest that even within *M. m. castaneus* hotspot locations are variable. On a broad scale, however, our results suggest that the recombination landscape is very strongly conserved between *M. m. castaneus* and *M. m. domesticus* at least. In addition, our estimate of the switch-error rate implies that phasing errors lead to upwardly biased estimates of the recombination rate when the true rate is low. This is a source of bias that should be assessed in future studies. Finally, we showed that the variation in recombination rate is positively correlated with genetic diversity, suggesting that natural selection reduces diversity at linked sites across the *M. m. castaneus* genome, consistent with the findings of Halligan et al. (2013).

Chapter 3

Understanding the factors that shape patterns of nucleotide diversity in the house mouse genome

This chapter has been prepared as a research paper and is currently under review at Molecular Biology and Evolution. The submitted manuscript was deposited on BioRXiv and is reproduced with minor alternations and re-ordering here. I designed the analyses with Peter D. Keightley. I performed the analyses and wrote the paper. Peter gave comments on previous versions of the manuscript.

3.1 Abstract

A major goal of population genetics has been to determine the extent by which selection at linked sites influences patterns of neutral nucleotide diversity in the genome. Multiple lines of evidence suggest that diversity is influenced by both positive and negative selection. For example, in many species there are troughs in diversity surrounding functional genomic elements, consistent with the action of either background selection (BGS) or selective sweeps. In this study, we investigated the causes of the diversity troughs that are observed in the wild house mouse genome. Using the unfolded site frequency spectrum (uSFS), we estimated the strength and frequencies of deleterious and advantageous mutations occurring in different functional elements in the genome. We then used these estimates to parameterize forward-in-time simulations of chromosomes, using realistic distributions of functional elements and recombination rate variation in order to determine if selection at linked sites can explain the observed patterns of nucleotide diversity. The simulations suggest that BGS alone cannot explain the dips in diversity around either exons or conserved non-coding elements (CNEs). A combination of BGS and selective sweeps produces deeper dips in diversity than BGS alone, but the inferred parameters of selection cannot fully explain the patterns observed in the genome. Our results provide evidence of sweeps shaping patterns of nucleotide diversity across the mouse genome, and also suggest that infrequent, strongly advantageous mutations play an important role in this. The limitations of using the uSFS for inferring the frequency and effects of advantageous mutations are discussed.

3.2 Introduction

Starting with the discovery of a positive correlation between nucleotide polymorphism and the recombination rate in *Drosophila* in the late 1980s and early 1990s (Aguade et al., 1989; Begun and Aquadro, 1992), it has become clear that natural selection affects genetic diversity across the genomes of many species (Cutter and Payseur, 2013; Corbett-Detig et al., 2015). More recently, models incorporating selection at sites linked to those under observation have been shown to explain a large amount of the variation in diversity across the genome (McVicker et al., 2009; Uchimura et al., 2015; Comeron, 2014; Elyashiv et al., 2016). However, a persistent challenge has been to tease apart the contributions of positive and negative selection to the observed patterns.

Because the fates of linked alleles are non-independent, selection acting at one site may have consequences for variation and evolution at another. In broad terms, there are two models describing the effects of directional selection on neutral genetic diversity at linked sites, selective sweeps (SSWs) and background selection (BGS). SSWs occur when positively selected alleles spread through a population, dragging with them the haplotype on which they arose (Maynard Smith and Haigh, 1974; Barton, 2000). There are a number of different types of SSW (reviewed in Booker et al. 2017a), but in the present study, when not made explicit, we use the term selective sweep to refer to the effects of a single de novo advantageous mutation being driven to fixation by selection. BGS, on the other hand, occurs because the removal of deleterious mutations results in a loss of genetic diversity at linked neutral sites (Charlesworth et al., 1993; Charlesworth, 2013). The magnitudes of the effects of SSWs and BGS depend on the strength of selection, the rate of recombination and the mutation rate (Hudson and Kaplan, 1995; Nordborg et al., 1996; Barton, 2000). SSWs and BGS have qualitatively similar effects on genetic diversity, however, and many polymorphism summary statistics have little power to distinguish between them Stephan (2010); Charlesworth (2013).

Several studies have attempted to differentiate between BGS and SSWs. For example, Sattath et al. (2011) examined patterns of nucleotide diversity around recent nucleotide substitutions in *Drosophila simulans.* Averaging across the entire genome, they observed a trough in diversity around nonsynonymous substitutions, whereas diversity was relatively constant around synonymous ones. This difference is expected under a model of recurrent SSWs, but not under BGS. Their results provide evidence that SSWs have been frequent in *D. simulans* since the species shared a common ancestor with *Drosophila melanogaster* (the outgroup used in that study). Similar results have been reported for *Capsella grandiflora* (Williamson et al., 2014). In humans

(Hernandez et al., 2011), house mice (Halligan et al., 2013) and maize (Beissinger et al., 2016), however, there is very little difference between the patterns of diversity around putatively neutral and potentially adaptive substitutions. These results have been interpreted as evidence that hard SSWs are infrequent in those species. However, Enard, et al. (2014) argued that the proportion of neutral amino acid substitutions in regions of the genome with low functional constraint (and thus weak BGS effects) will be higher than the proportion occurring in regions with high functional constraint (and thus stronger BGS effects), so the Sattath test will be difficult to interpret in species with genomes exhibiting highly variable levels of functional constraint, such as humans and mice (but see Beissinger et al. 2016). Indeed, Enard et al. (2014) found evidence that adaptive substitutions are fairly frequent in both protein-coding and non-coding portions of the human genome, suggesting that SSWs are common. Furthermore, Nam et al. (2017) analysed reductions in nucleotide diversity in genomic regions close to genes in the great apes and concluded that strong SSWs, rather than BGS, are required to explain the observed patterns. In that study, although the authors examined a wide range of selection parameters, they did not attempt to identify parameters of selection specific to any ape species.

There are a number of methods for estimating the frequency and strength of advantageous mutations using models of selection at linked sites (Booker et al., 2017a). Recently, Elyashiv et al. (2016) produced a map of the expected nucleotide diversity in *D. melanogaster* by fitting a model incorporating both BGS and hard SSWs to the genome-wide patterns of genetic diversity and the divergence between *D. melanogaster* and *D. simulans*. They concluded that sweeps are required to explain much of the genome-wide variation in diversity. Their analysis conditioned the effects of sweeps on the locations of recent substitutions, which is reasonable in *D. melanogaster* where there is a reduction in mean diversity around nonsynonymous substitutions, but not around synonymous ones (Elyashiv et al., 2016). As described above, this is not the case for wild mice. Indeed, even randomly selected synonymous and nonsynonymous sites in *Mus musculus*, regardless of whether they have experienced a recent substitution, exhibit almost identical reductions in diversity in surrounding regions Halligan et al. (2013). Conditioning a sweep model on the locations of recent substitutions in mice may, therefore, produce spurious parameter estimates. Furthermore, the selection parameters estimated by Halligan et al. (2013) were inferred solely from variation in nucleotide diversity. There is information in the distribution of allele frequencies, the site frequency spectrum (SFS), that can be used to estimate the distribution of fitness effects (DFE) for both deleterious and advantageous mutations (Keightley and Eyre-Walker, 2007; Boyko et al., 2008; Schneider et al., 2011; Tataru et al., 2017). In the present study, we estimate the DFE using such methods, and then use our estimates to parameterise simulations modelling BGS and SSWs.

In this study, we attempt to understand the influence of natural selection on variation at linked sites in the house mouse, *Mus musculus.* Specifically, we analyse *M. m. castaneus*, a sub-species which has been estimated to have a long-term effective population size (N_e) of around 500,000 (Baines and Harr, 2007; Halligan et al., 2010), making it a powerful system in which to study molecular evolution in mammals. Both protein-coding genes and phylogenetically conserved non-coding elements (CNEs, which have roles in the regulation of gene expression (Lowe et al., 2011)) exhibit signatures of natural selection in *M. m. castaneus* (Halligan et al., 2013). In particular, Halligan et al. (2013) showed that there are substantial reductions in diversity surrounding protein-coding exons and CNEs, consistent with selection reducing diversity at linked sites. The trough in diversity surrounding exons was found to be 10x wider than the trough surrounding CNEs, suggesting that selection is typically stronger on protein sequences than regulatory sequences. These results, therefore, suggest that selection at linked sites affects nucleotide diversity across large portions of the genome. However, our understanding of the forces that have shaped patterns of diversity is incomplete.

We analyse data on wild-caught *M. m. castaneus* individuals to obtain estimates of the distribution of fitness effects (DFE) for several classes of functional elements in the mouse genome and then use these to parametrise forward-in-time simulations. We analyse several aspects of our simulation data: 1) the patterns of genetic diversity and the distribution of allele frequencies around both protein-coding exons and conserved non-coding elements; 2) the rates of substitution in different functional elements; and 3) the patterns of diversity around nonsynonymous and synonymous substitutions.

3.3 Materials and Methods

3.3.1 Samples and polymorphism data

We analysed the genome sequences of 10 wild-caught *M. m. castaneus* individuals sequenced by Halligan et al. (2013). The individuals were sampled from an area that is thought to include the ancestral range of the species(Baines and Harr, 2007). A population structure analysis suggested that the individuals chosen for sequencing came from a single randomly mating population (Halligan et al., 2010). Sampled individuals were sequenced to an average depth of 30x using Illumina technology. Reads were mapped to version mm9 of the mouse genome and variants called as described in Halligan et al. (2013). Only single nucleotide polymorphisms were considered, and insertion/deletion polymorphisms were excluded from downstream analyses. We used the genome sequences of *Mus famulus* and *Rattus norvegicus* as outgroups in this study. For *M. famulus*, a single individual was sequenced to high coverage and mapped to the mm9 genome (Halligan et al., 2013). For *R. norvegicus*, we used the whole genome alignment of the mouse (mm9) and rat (rn4) reference genomes from UCSC.

For the DFE-alpha analysis (see below), the underlying model assumes a single, constant mutation rate. Hypermutable CpG sites strongly violate this assumption, so CpG-prone sites were excluded as a conservative way to remove CpG sites from our analyses. A site was labelled as CpG-prone if it is preceded by a C or followed by a G in the 5 to 3 direction in either *M. m. castaneus*, *M. famulus* or *R. norvegicus.*

Additionally, sites that failed a Hardy-Weinberg equilibrium test (p ¡ 0.002) were excluded from further analysis, because they may represent alignment errors.

3.3.2 Functional elements in the murid genome

In this study, we considered three different classes of functional elements in the genome: the exons and untranslated regions (UTRs) of protein-coding genes and conserved non-coding elements (CNEs).

Coordinates for canonical splice-forms of protein-coding gene orthologs between *Mus musculus* and *Rattus norvegicus* were obtained from version 67 of the Ensembl database. We used these to identify untranslated regions (UTRs) as well as 4-fold and 0-fold degenerate sites in the coding regions. We made no distinction between 3 and 5 UTRs in the analysis. Genes containing alignment gaps affecting ¿80% of sites in either outgroup and genes containing overlapping reading frames were excluded. This left a total of 18,171 autosomal protein-coding genes.

The locations of conserved non-coding elements (CNEs) in the house mouse genome were identified as described by Halligan et al. (2013).

Estimating the parameters of the distribution of fitness effects (DFE) for a particular class of sites using DFE-alpha (see below) requires neutrally evolving sequences for comparison. When analysing 0-fold degenerate sites and UTRs, we used 4-fold degenerate sites as the comparator. For CNEs, we used non-conserved sequence in the flanks of CNEs. (Halligan et al., 2013) found that, compared to the genome-wide average, nucleotide divergence between mouse and rat in the 500bp on either side of CNEs is 20% lower than that of intergenic DNA distant from CNEs, suggesting functional constraint in these regions. For the purpose of obtaining a quasi-neutrally evolving reference class of sequence and to avoid these potentially functional sequences, we therefore used sequence flanking the edges of each CNE, offset by 500bps. For each CNE, the total amount of flanking sequence used in the analysis was equal to the length of the focal CNE, split evenly between the upstream and downstream regions. CNE-flanking sequences overlapping with another annotated feature (i.e. exon, UTR or CNE) or the flanking sequence of another CNE were excluded.

3.3.3 The site frequency spectrum around functional elements

For distances of up to 100Kbp on either side of exons and 5Kbp on either side of CNEs, the non-CpG-prone sites in non-overlapping windows of 1Kbp and 100bp, respectively, were extracted. For each analysis window, we calculated the genetic distance to the focal element in terms of the population-scaled recombination rate ($\rho = 4N_er$) using the *M. m. castaneus* recombination map we constructed in an earlier study (Booker et al., 2017b). Sites within analysis windows that overlapped with any of the annotated features described above, or that contained missing data in *M. m. castaneus* or either outgroup were excluded. The data for analysis windows were collated based on either physical or genetic distances distance from the nearest CNE or exon, from which we calculated nucleotide diversity and Tajimas D.

3.3.4 Overview of DFE-alpha analysis

The distribution of allele frequencies in a sample, referred to as the site frequency spectrum (SFS), provides information on evolutionary processes. Under neutrality the SFS reflects past demographic processes, such as population expansions and bottlenecks, and potentially the effects of selection at linked sites. The allele frequency distribution will also be distorted if focal sites are subject to functional constraints. The SFS therefore contains information on the strengths and frequencies of mutations with different selective effects, known as the distribution of fitness effects (hereafter the DFE). Note that balancing selection may maintain alleles at intermediate frequencies

(Charlesworth, 2006), but we assume that the contribution of this form of selection to overall genomic diversity is negligible.

DFE-alpha estimates selection parameters using information contained in the SFS by a two-step procedure (Keightley and Eyre-Walker, 2007). First, a demographic model is fitted to data for a class of putatively neutral sites. Conditional on the demographic parameter estimates, the DFE is then estimated for the selected sites. In the absence of knowledge of ancestral or derived alleles, the folded SFS can be used to estimate the demographic model and the DFE for harmful variants (hereafter referred to as the dDFE) (Keightley and Eyre-Walker, 2007). If information from one or more outgroup species is available, and the ancestral state for a segregating site can be inferred, one can construct the unfolded SFS (uSFS). In the presence of positive selection, such that advantageous alleles segregate at an appreciable frequency, the parameters of the distribution of fitness effects for advantageous mutations can be estimated from the uSFS (Schneider et al., 2011; Keightley et al., 2016; Tataru et al., 2017). In this study, we estimate the proportion of new mutations occurring at a site that are advantageous (p_a) and the strength of selection acting on them ($N_e s_a$).

3.3.5 Inference of the uSFS and the DFE

We inferred the distributions of derived allele frequencies in our sample for 0-fold and 4-fold sites, UTRs, CNEs and CNE-flanks using *M. famulus* and *R. norvegicus* as outgroups, using the two-outgroup method implemented in ml-est-sfs v1.1 (Keightley et al., 2016). This method employs a two-step procedure conceived to address the biases inherent in parsimony methods. The first step estimates the rate parameters for the tree under the Jukes-Cantor model by maximum likelihood assuming a single mutation rate. Conditional on the rate parameters, the individual elements of the uSFS are then estimated.

DFE-alpha fits discrete population size models, allowing up to two changes in population size through time. For each class of putatively neutral sites, one-, two- and three-epoch models were fitted by maximum likelihood and the models with the best fit (as judged by likelihood ratio tests) were used in further analyses. When fitting the three-epoch model, we ran DFE-alpha (v2.16) 10 times with a range of different search algorithm starting values, in order to check convergence. We explored the effect of fitting the demographic model to a smaller number of individuals by down-sampling the 4-fold sites dataset to 5 and 8 individuals, with respect to frequency.

In the cases of 4-fold sites and CNE-flanks, the inferred uSFSs exhibited a higher proportion of high frequency derived alleles than expected under the best-fitting demographic model (Figure S1) (hereafter referred to as an uptick). Such an increase is not possible under the single population, single locus demographic models assumed. There are several possible explanations for the uptick: 1) mis-inference of the uSFS due to an inadequacy of the model assumed in ml-est-sfs; 2) failure to capture the demographic history of *M. m. castaneus* by the models implemented in DFE-alpha; 3) sequencing errors in *M. m. castaneus* or either outgroup generating spurious signals of divergence; 4) SSWs, since they can drag linked alleles to high frequencies (Braverman et al., 1995; Kim, 2006); 5) cryptic population sub-division in our sample of mouse individuals; and 6) positive selection, acting on the putatively neutral sites themselves. We think this latter explanation is unlikely, however, since there is little evidence for selection on synonymous codon usage in *Mus musculus* (dos Reis and Wernisch, 2009). With the exception of direct selection affecting the putatively neutral class of sites, the above sources of bias should also affect the selected class of sites (Eyre-Walker et al., 2006; Glemin et al., 2015; Keightley et al., 2016). We therefore corrected the selected sites uSFS prior to inferring selection parameters by subtracting the proportional deviation between the neutral uSFS expected under the best-fitting demographic model and the observed neutral uSFS (following Keightley et al. 2016; see Supplementary Methods).

Simultaneous inference of the DFE for harmful mutations (dDFE) and adaptive mutation parameters was performed using DFE-alpha (v.2.16) (Schneider et al., 2011). A gamma distribution has previously been used to model the dDFE, since it can take a variety of shapes and has only two parameters (Eyre-Walker and Keightley, 2007). However, more parameter-rich discrete point mass distributions provide a better fit to nonsynonymous polymorphism site data in wild house mice (Kousathanas and Keightley, 2013). We therefore compared the fit of one, two and three discrete class dDFEs and the gamma distribution, and also included one or more classes of advantageous mutations. Nested DFE models were compared using likelihood ratio tests, and non-nested models were compared using Akaikes Information Criteria (AIC). Goodness of fit was also assessed by comparing observed and expected uSFSs using the χ^2 statistic, but the numbers of sites in the i^{th} and $n - i^{th}$ classes are non-independent, so formal hypothesis tests were not performed.

We constructed profile likelihoods to obtain confidence intervals. Two unit reductions in $logL$, on either side of the maximum likelihood estimates (MLEs) were taken as approximate 95% confidence limits.

3.3.6 Two methods for inferring the rates and effects of advantageous mutations based on the uSFS

It has been suggested that estimates of the DFE obtained based on the uSFS may be biased if sites fixed for the derived allele are included in calculations Tataru et al. (2017). Sites fixed for the derived allele are typically a frequent class in the uSFS, and therefore strongly influence parameter estimates. Bias can arise, for example, if the selection strength has changed since the split with the outgroup, such that the number of sites fixed for the derived allele do not reflect the selection regime that generated current levels of polymorphism. If nucleotide divergence and polymorphism are decoupled in this way, selection parameter estimated from only polymorphism data (and sites fixed for ancestral alleles) may therefore be less biased than those obtained when using the full uSFS. To investigate this possibility, we estimated selection parameters either utilising the full uSFS (we refer to this method as Model A) or by analysing the uSFS while fitting an additional parameter (Supplementary Methods), such that sites fixed for the derived allele do not contribute to estimates of the selection parameters (we refer to this method as Model B).

Certain alleles present in a sample of individuals drawn from a population may appear to be fixed that are, in fact, polymorphic. Attributing such polymorphisms to between-species divergence may then influence estimates of the DFE by increasing the number of sites fixed for the derived allele (note that this would only affect estimates obtained under Model A). We corrected the effect of polymorphism attributed to divergence using an iterative approach as follows. When fitting selection or demographic models, DFE-alpha produces a vector of expected allele frequencies. Using this vector, we inferred the expected proportion of polymorphic sites that appear to be fixed for the derived allele. This proportion was then subtracted from the fixed derived class and distributed among the polymorphism bins according to the allele frequency vector. We then refitted the model using this corrected uSFS, and this procedure was applied iteratively until convergence (See Supplementary Methods). For each site class, convergence was achieved within five iterations and the selection parameters for each class did not substantially change between iterations.

3.3.7 Forward-in-time simulations modelling background selection and selective sweeps

We performed forward-in-time simulations in SLiM v1.8 (Messer, 2013) to assess whether the observed patterns of diversity around functional elements (Halligan et al., 2013) can be explained by SSWs or BGS caused by mutations originating in the elements themselves. These simulations focussed on either protein-coding exons or CNEs. We also ran SLiM simulations to model the accumulation of between-species divergence under our estimates of the DFE. In all our simulations, we either assumed the estimates of selection parameters obtained from the full uSFS (Model A) or those obtained when sites fixed for the derived allele do not contribute to parameter estimates (Model B).

Models of BGS and recurrent SSWs predict that the magnitudes of their effects are sensitive to the rate of recombination and mutation rate and the strength of selection (Wiehe and Stephan, 1993; Nordborg et al., 1996; Coop and Ralph, 2012). To parameterise our simulations, we used estimates of compound parameters scaled by Ne. For example, estimates of selection parameters obtained from DFE-alpha are expressed in terms of $N_e s$ (where s is the difference in fitness between homozygotes for ancestral and derived alleles, assuming semi-dominance). For a population where N_e = 1,000 and s = 0.05, for example, the strength of selection is therefore approximately equivalent to that of a population where N_e = 10,000 and s = 0.005. By scaling parameter values according to the population size of the simulations (N_{sim}), we modelled the much larger *M. m. castaneus* population (N_e 500,000; Geraldes et al. 2011) in a computationally tractable way. However, this linear scaling can be problematic, particularly for strong positive selection (Uricchio and Hernandez, 2014). We compared patterns of diversity in simulations with population sizes of N_{sim} = 100, 500, 750, 1,000 and 2,000 diploid individuals to assess the effect of the linear scaling given the selection, recombination and mutation parameters we assumed.

1. Annotating simulated chromosomes

Functional elements are non-randomly distributed across the house mouse genome. For example, protein-coding exons are clustered into genes and CNEs are often found close to other CNEs (Halligan et al., 2013). Incorporating this distribution into simulations is important when modelling BGS and recurrent SSWs, because their effects on neutral diversity depend on the density of functional sequence (Nordborg et al., 1996; Campos et al., 2017). We incorporated the distribution as follows. For each simulation replicate, we chose a random position on an autosome, which was itself randomly selected (with respect to length). The coordinates of the functional elements (exons, UTRs and CNEs) in the 500Kbp downstream of that position were used to annotate a simulated chromosome of the same length. For simulations focussing on exons (CNEs), we only used chromosomal regions that had at least one exon (CNE).

2. Mutation, recombination and selection in simulations

We used an estimate of the population scaled mutation rate, $\theta = 4N_e\mu$, to set the mutation rate (μ) in simulations, such that levels of neutral polymorphism approximately matched those of *M. m. castaneus.* Diversity at putatively neutral sites located close to functional elements (for example, 4-fold synonymous sites) may be affected by BGS and SSWs. To correct for this, we used an estimate of $\theta = 0.0083$, based on the average nucleotide diversity at non-CpG-prone sites at distances ¿75Kbp from protein-coding exons. This distance was used, because it the approximate distance beyond which nucleotide diversity remains flat. The mutation rate in simulations was thus set to $0.0083/4N_{sim}$.

Variations in the effectiveness of selection at linked sites, due to variation in the rate of recombination across the genome, may not be captured by simulations that assume a single rate of crossing over. Recently, we generated a map of variation in the rate of crossing-over for *M. m. castaneus* using a coalescent approach (Booker, et al. 2017b), quantified in terms of the population scaled recombination rate $\rho = 4N_er$. Recombination rate variation in the 500Kbp region used to obtain functional annotation was used to specify the genetic map for individual simulations.

We modelled natural selection at sites within protein-coding exons, UTRs and CNEs in the simulations using the estimates of selection parameters obtained from the

DFE-alpha analysis. In the case of protein-coding exons, 25% of sites were set to evolve neutrally (i.e. synonymous sites), and the fitness effects of the remaining 75% were drawn from the DFE inferred for 0-fold sites (hereafter termed nonsynonymous sites in the simulations). For mutations in UTRs and CNEs, 100% were drawn from the DFEs inferred for those elements. Population scaled selection coefficients were divided by $2N_[sim]$ to obtain values of s for use in simulations. All selected mutations were assigned a dominance coefficient of 0.5, as assumed by DFE-alpha.

3. Patterns of diversity around functional elements in simulations

We examined the contributions of BGS and recurrent SSWs to the troughs in diversity observed around protein-coding exons and CNEs using forward-in-time simulations. Focussing on either protein-coding exons or CNEs, we performed three sets of simulations. The first incorporated only harmful mutations (causing BGS), the second only advantageous mutations (causing SSWs), and the third set incorporated both (causing both processes). Thus, under a given set of DFE estimates, we performed six sets of simulations (three sets focussing on exons and three sets focussing on CNEs). For each simulation set, 2,000 SLiM runs were performed, each using a randomly sampled 500Kbp region of the genome. In each SLiM run, populations of $N_[sim] =$ 1,000 diploid individuals were allowed to evolve for 10,000 generations ($10N_[sim]$) in order to approach mutation-selection-drift balance. At this point, 200 randomly chosen haploid chromosomes were sampled from the population and used to construct SFSs.

For each set of simulations, segregating sites in windows surrounding functional elements were analysed in the same way as for the *M. m. castaneus* data (see above). The SFSs for all windows at the same distance from an element were collated. Analysis windows around protein-coding exons were oriented with respect to the strand orientation of the actual gene. Neutral sites near the tips of simulated chromosomes only experience selection at linked sites from one direction, so analysis windows located within 60Kbp of either end of a simulated chromosome were discarded. For a given distance to a functional element, we obtained confidence intervals around individual statistics by bootstrapping analysis window 1,000 times.

Mutation rate variation is expected to contribute to variation in nucleotide diversity. Nucleotide divergence between mouse and rat is relatively constant in the intergenic regions surrounding protein-coding exons (Halligan, et al. 2013), suggesting that mutation rate variation is not responsible for the troughs in diversity around exons. Around CNEs, however, there is a pronounced dip in nucleotide divergence between *M. m. castaneus* and the rat. A likely explanation for this is that alignment-based approaches to identify CNEs fail to identify the edges of some elements, resulting in the inclusion of functionally constrained sequence in the analysis windows close to CNEs. This factor was not incorporated in our simulations, so in order to correct for this constraint, allowing us to compare diversity around CNEs in *M. m. castaneus* with our simulation data, we scaled values as follows. We divided nucleotide diversity by between-species divergence, in this case mouse-rat divergence, giving a statistic ($\pi/d_[rat]$) that reflects diversity corrected for mutation rate variation. We then multiplied the /drat values by the mean mouse-rat divergence in regions further than 3Kbp from the edges of CNEs to obtain values on the same scale as our simulation data.

When comparing the patterns of diversity around functional elements in our simulations with the observations from *M. m. castaneus*, we used the root mean square (RMS) as a measure of goodness-of-fit.

$$RMS = \sqrt{\frac{1}{n_w}\sum_{i=1}^{n_w}(\pi_{sim,i} - \pi_{obs,i})^2} \tag{3.1}$$

where $\pi_{sim,i}$ and $\pi_{obs,i}$ are the diversity values from simulations and *M. m. castaneus*, respectively, in window i around a particular class of functional element and n_w is the total number of analysis windows. Approximate confidence intervals for RMS values were obtained using the bootstrap replicates described above.

4. Re-inferring the DFE based on simulated population data

We performed two additional sets of simulations to model the accumulation of between-species nucleotide divergence under the DFE estimates obtained by analysis of the full uSFS (i.e. Model A) and those obtained when sites fixed for the derived allele did not contribute to selection parameters (i.e. Model B). These simulations were the same as those described above, except that we ran them for additional generations to approximate the mouse-rat divergence. We ran 4,000 replicates of these simulations. Using polymorphic sites and sites fixed for the derived allele, we constructed the uSFS for each class of functional sites.

In order to model the mouse-rat divergence, we required a time frame to approximate the neutral divergence between those two species. Neutral divergence between *M. m. castaneus* and *R. norvegicus* (K_{rat}) is 15% at non-CpG-prone sites far from protein-coding exons. Under neutrality, divergence is expected to be equal to $2T\mu$, where T is the time in generations since the two-species shared a common ancestor and μ is the mutation rate per base pair per generation. In the simulations, the mutation rate was 2.075 x 10^{-6} bp^{-1} (recall that we scaled mutations rates using an estimate of $4N_e\mu$) and since $K_[rat] = 0.15$, T = 36,145 generations. We thus ran simulations incorporating both deleterious and advantageous mutations, focussing on exons, for 46,145 generations, discarding the first 10,000 as burn-in. At the final generation, we constructed the uSFS for synonymous and nonsynonymous sites from 20 randomly sampled haploid chromosomes. To obtain a proxy for mouse-rat divergence, we counted all substitutions that occurred after the $10N_{sim}$ burn-in phase plus any derived alleles present in all 20 haploid chromosomes.

Using the uSFSs for synonymous and nonsynonymous sites obtained from the simulations, we estimated selection parameters using the methods described above. We first fitted one-, two- and three- epoch demographic models to simulated synonymous site data. For the simulations assuming Model A or Model B, we found that the three-epoch demographic model gave the best fit to the simulated synonymous site uSFS in both cases. Using the expected uSFS under the three-epoch model, we performed the demographic correction (Supplementary Methods) before estimating selection parameters. When estimating selection parameters based on simulation data, we used the same methods as used for the analysis of the *M. m. castaneus* data, i.e. the DFE for Model A simulations was estimated using Model A etc.

5. Patterns of diversity around recent nonsynonymous and synonymous substitutions

Comparisons of the average level of nucleotide diversity around recent synonymous and nonsynonymous substitutions have been used to test for positive selection (Hernandez et al., 2011; Sattath et al., 2011; Halligan et al., 2013; Williamson et al., 2014; Beissinger et al., 2016). In *M. m. castaneus* there is essentially no difference in diversity around recent substitutions at 0-fold and 4-fold sites (Halligan et al., 2013). This could reflect a paucity of SSWs, or alternatively, this particular test may be unable to discriminate between BGS and SSWs in mice. Using our simulation data, in which SSWs are relatively frequent, we tested whether patterns of diversity around selected and neutral substitutions reveals the action of positive selection. In their study, Halligan et al. (2013) used *M. famulus* as an outgroup to locate recent substitutions, because it is much more closely related to *M. musculus* than the rat. We obtained the locations of nucleotide substitutions in our simulations as follows. Neutral divergence between *M. m. castaneus* and *M. famulus* (K_{fam}) is 3.4%. In the simulations, given that the mutation rate was 2.075 x 10^6, 8,193 generations are sufficient to approximate the *M. m. castaneus* lineage since its split with *M. famulus* K_{fam}. Thus, all substitutions that occurred in 8,193 generations were analysed. Neutral diversity around synonymous and nonsynonymous substitutions in non-overlapping windows of 1,000bp up to 100Kbp from substituted sites were then extracted from the simulations. Sites in analysis windows that overlapped with functional elements were excluded. If two substitutions of the same type were located less than 100Kbp apart, analysis windows extended only to the midpoint of the two sites.

3.4 Results

We investigated the causes of variation in genetic diversity around functional elements of house mice by analysing the genomes of 10 wild-caught individuals sequenced to high coverage (Halligan et al., 2013). We compared nucleotide polymorphism and between-species divergence in three classes of functional sites (0-fold nonsynonymous sites, UTRs and CNEs) with polymorphism and divergence at linked, putatively neutral sequences (4-fold synonymous sites and CNE-flanks). The three classes of functional sites had lower levels of within-species polymorphism and between-species divergence than their neutral comparators (Table 3.1). This is expected if natural selection keeps deleterious alleles at low frequencies, preventing them from reaching fixation. Tajimas D is more negative for 0-fold sites, UTRs and CNEs than for their neutral comparators (Table 3.1), further indicating the action of purifying selection. It is notable that the two neutral site types exhibited negative Tajimas D, indicating that rare variants are more frequent than expected in a Wright-Fisher population (Table 3.1). This is consistent either with a recent population expansion or the widespread effects of selection on linked sites, both of which may be relevant for this population (Halligan et al., 2013; Booker et al., 2017b).

Table 3.1: Summary statistics for five classes of sites in *M. m. castaneus*. All values refer to non-CpG prone sites. Nucleotide divergences between *M. m. castaneus* and M. famulus (d_{fam}) and between *M. m. castaneus* and *R. norvegicus* (d_{rat}) were estimated by maximum likelihood using the method described in Keightley et al. (2016).

	π (%)	Tajima's D	d_{fam}	d_{rat}	Sites (Mb)
0-fold	0.134	-0.763	0.239	2.93	10.20
4-fold	0.628	-0.627	1.060	12.70	1.49
CNE	0.274	-1.030	0.418	3.67	24.60
CNE flank	0.670	-0.602	1.030	13.80	17.80
UTR	0.438	-0.702	0.802	10.00	11.30

3.4.1 Inferring the unfolded site frequency spectrum

The distribution of derived allele frequencies in a class of sites (the unfolded site frequency spectrum - uSFS) potentially contains information on the frequency and strength of selected mutations. We estimated the uSFSs for 0-fold sites, UTRs and CNEs using a probabilistic method incorporating information from two outgroups (Keightley et al., 2016). This method attempts to correct for biases inherent in parsimony methods.

A population's demographic history is expected to affect the shape of the SFS. DFE-alpha attempts to correct this by fitting a population size change model to the neutral site class, and, conditional on the estimated demographic parameters, estimates the DFE for linked, selected sites. In the case of 4-fold sites and CNE flanks, a 3-epoch model provided the best fit to the data, based on likelihood ratio tests (Table S1) The trajectories of the inferred population size changes were similar in each case, i.e. a population bottleneck followed by an expansion (Table S2). However, the magnitude of the changes and the duration of each epoch differed somewhat (Table S2). A possible explanation is that the demographic parameter estimates are affected by selection at linked sites, which differs between site classes (Messer and Petrov, 2013; Ewing and Jensen, 2016; Schrider et al., 2016).

To investigate if the inferred parameters of the demographic model are sensitive to the number of sampled individuals, we fitted the 3-epoch model to the 4-fold site data, down-sampled to either 5 or 8 individuals. We found that the magnitude of the population expansion inferred increased with sample size (Table S3). This is presumably caused by an excess of rare variants (for example singletons), as expected under population expansion. Increasing the number of sampled individuals will lead to an increasingly precise estimate of the proportion of rare variants in the population.

We found that the 4-fold site and CNE-flank uSFSs exhibited an excess of high frequency derived alleles relative to expectations under the best-fitting neutral demographic models (Figure **??**). For example, χ^2-statistics for the difference between the observed and fitted number of sites for the last uSFS element (i.e. 19 derived alleles) were 245.9 and 505.6 for 4-fold sites and CNE-flanks, respectively. It is reasonable to assume that the differences between fitted and observed values are caused by processes that similarly affect the linked selected site class. We therefore corrected the 0-fold, UTR and CNE uSFSs by subtracting the proportional deviations between fitted and observed values for neutral site uSFSs prior to estimating selection parameters (see Supplementary Methods). Applying this correction (hereafter referred to as the demographic correction) appreciably reduced the proportion of high frequency derived variants (Figure 3.1).

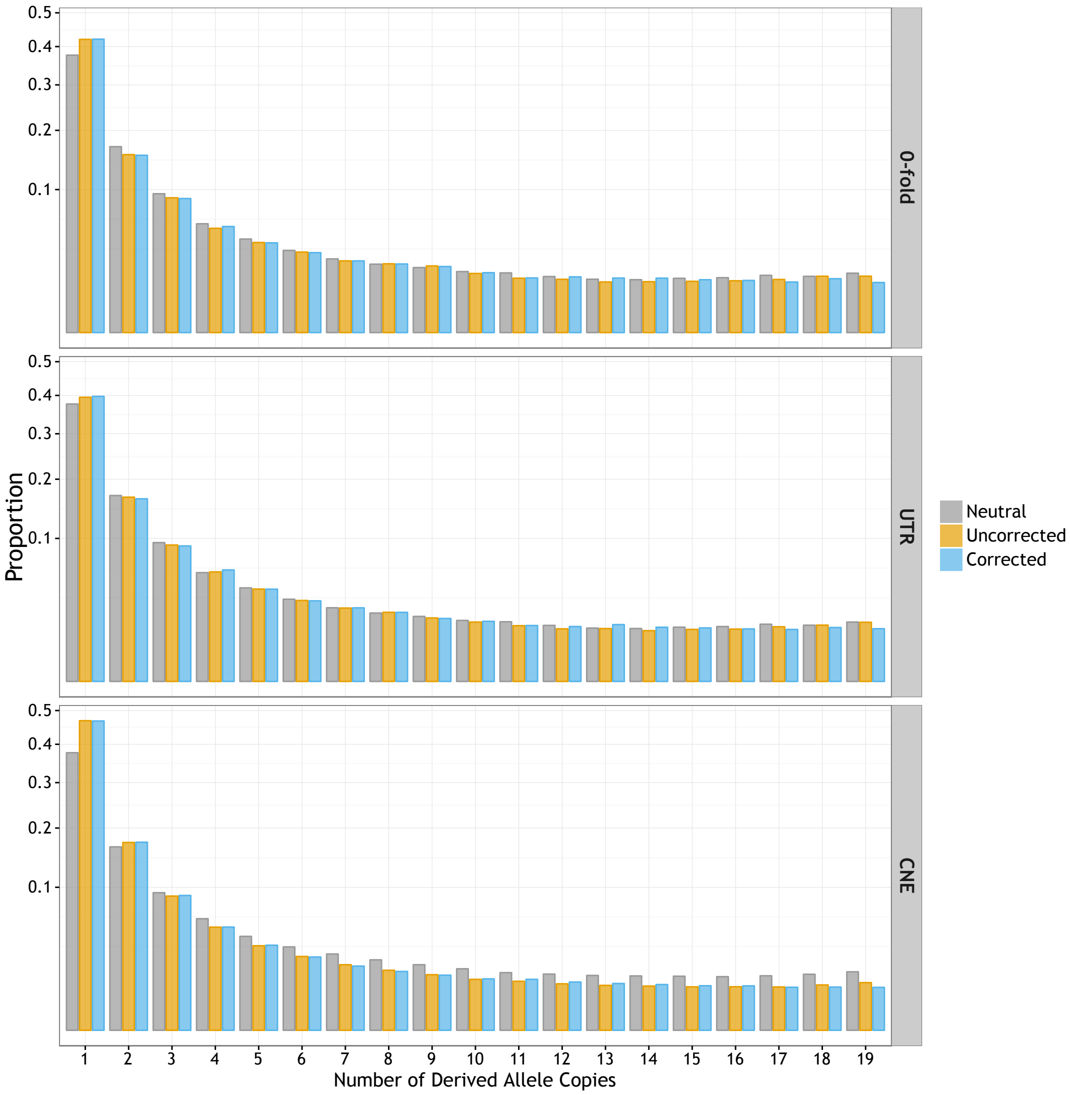


Figure 3.1: The uSFS for three classes of functional sites (yellow and blue bars) compared to a putatively neutral comparator (grey bars). The neutral comparator for 0-fold sites and UTRs was 4-fold degenerate synonymous sites in both cases. For CNEs, the neutral comparator was CNE-flanking sequence. The expected uSFS under a demographic model fitted to a neutral comparator was used to correct the uSFS for the corresponding selected sites (see Methods).

3.4.2 Estimating the frequencies and strengths of deleterious and advantageous mutations

We estimated the DFE for harmful mutations (dDFE) and the rate and strength of advantageous mutations based on the uSFSs for the three different classes of functional sites using DFE-alpha under two different models (Table 3.2). The first, as described by Schneider et al. (2011), used the full uSFS, including sites fixed for the derived allele (hereafter Model A). The second (hereafter Model B), incorporated an additional parameter that absorbs the contribution of sites fixed for the derived allele (see Supplementary Methods). This was motivated by the possibility that between-species divergence may be decoupled from within-species polymorphism (e.g. due to changing selection regimes), and this could lead to spurious estimates of selection parameters Eyre-Walker and Keightley (2009); Tataru et al. (2017). Since Model A is nested within Model B, the two can be compared using likelihood ratio tests. In the remainder of the study, results obtained under Model A are shown in parallel with results obtained under Model B.

Table 3.2: Parameter estimates for the distribution of fitness effects for three classes of sites in *M. m. castaneus* obtained under two models. The first (Model A) estimates of selection parameters based on the full uSFS. Under the second method (Model B), sites fixed for the derived allele were prevented from influencing estimates of selection parameters. The bracketed values are 95% confidence intervals obtained from profile likelihoods. The parameters shown are: $p(i)$ = the proportion of mutations falling into the i^{th} deleterious class; $2N_es(i)$ = the scaled homozygous selection coefficient of the i^{th} deleterious class; p_a = the proportion of advantageous mutations; $2N_es_a$ = the scaled homozygous selection coefficient of the advantageous mutation class.

	Model A: DFE inferred from the full uSFS		
	0-fold	**UTR**	**CNE**
$2N_es(1)$	-0.09	-0.194	-0.646
$p(1)$	0.191	0.701	0.352
$2N_es(2)$	-208	-78.2	-7.96
$p(2)$	0.806	0.286	0.278
$2N_es(3)$	-	-	-155.8
$p(3)$	-	-	0.36
$2N_es_a$	14.54	10.64	18.34
	[9.24 23.4]	[7.82 14.1]	[14.0 41.8]
p_a	0.003	0.013	0.0098
	[0.0019 0.0048]	[0.0097 0.019]	[0.0037 0.0099]
	Model B: Fixed derived alleles do not affect parameter estimates		
	0-fold	**UTR**	**CNE**
$2N_es(1)$	-0.342	-0.32	-0.506
$p(1)$	0.184	0.689	0.342
$2N_es(2)$	-200	-64	-7.68
$p(2)$	0.806	0.281	0.286
$2N_es(3)$	-	-	-152.6
$p(3)$	-	-	0.365
$2N_es_a$	16.6	13.9	17.2
	[12.5 20.2]	[11.1 17.4]	[8.74 25.2]
p_a	0.01	0.0294	0.008
	[0.0030 0.0183]	[0.0181 0.0436]	[0.0004 0.0100]

We performed a comparison of different DFE models, including discrete distributions with one, two or three mutational effect classes and the gamma distribution including or not including advantageous mutations. For each class of functional sites, DFE models with several classes of deleterious mutational effects and a single class of advantageous effects gave the best fit (Table C.4). For each class of functional sites, only a single class of advantageous mutations was supported, since additional classes of advantageous mutations did not significantly increase likelihoods (Table C.5), presumably reflecting a lack of power. These best-fitting models were identified under both Model A or Model B. Parameter estimates pertaining to the dDFE were also similar between Models A and B (Table 3.2).

In this study, we estimated selection parameters based on the uSFS, whereas earlier studies on mice used the distribution of minor allele frequencies, i.e. the folded SFS (Halligan et al., 2010, 2011; Kousathanas et al., 2011; Halligan et al., 2013; Kousathanas et al., 2014). A possible consequence of using the folded SFS is that advantageous mutations segregating at intermediate to high frequencies are allocated to the mildly deleterious class. In the case of 0-fold sites, for example, the best-fitting DFE did not include mutations with scaled effects in the range of 1 ¡ | $2N_es$ | ¡ 200 (Table 3.2). This contrasts with previous studies using the folded SFS, which identified an appreciable proportion of mutations in the 1 ¡ | $2N_es$ | ¡ 200 range (Halligan et al., 2013; Kousathanas and Keightley, 2013). This difference may influence the reductions in diversity caused by background selection, so we performed simulations incorporating either the gamma dDFEs inferred from analysis of the folded SFS by Halligan et al. (2013) or the discrete dDFEs inferred in the present study.

For all classes of functional sites, we inferred that moderately positively selected mutations are fairly frequent under both Models A and B (Table 3.2). In the case of 0-fold sites, for example, the frequency of advantageous mutations was 0.3% (under Model A). Across the three classes of sites, the scaled selection strengths of advantageous mutations were fairly similar (Table 3.2), i.e. $2N_es$ 16, implying that s is on the order of 10^{-5} (assuming N_e = 500,000; Geraldes et al. (2011)). However, we found that estimates of the frequency of advantageous mutations (p_a) obtained under Model B for 0-fold sites and UTRs were 3 times higher than those obtained under Model A. In the cases of both 0-fold sites and UTRs, Model B fitted significantly better than Model A, as judged by likelihood ratio tests (0-fold sites, $\chi^2_{1d.f.} = 4.2$; $p = 0.04$; UTRs,

$\chi^2_{1d.f.} = 9.9$; $p = 0.002$). Interestingly, in the case of CNEs, Models A and B did not differ significantly in fit ($\chi^2_{1d.f.} = 0.26$; $p = 0.60$) and estimates of the advantageous mutation parameters were similar (Table 3.2).

3.4.3 Forward-in-time population genetic simulations

We conducted forward-in-time simulations to examine whether estimates of the DFE obtained by analysis of the uSFS predict patterns of diversity observed around functional elements. In our simulations, we used estimates of selection parameters obtained by DFE-alpha for 0-fold sites, UTRs and CNEs assuming either Model A (i.e. from the full uSFS) or Model B (i.e. by absorbing the contribution of sites fixed for the derived allele with an additional parameter). The selection parameter estimates obtained under Models A and B resulted in major differences in the patterns of diversity around functional elements.

i) Patterns of nucleotide diversity around functional elements in simulated populations

Using the selection parameter estimates obtained from DFE-alpha (Table 3.2), we performed simulations incorporating deleterious mutations, advantageous mutations or both advantageous and deleterious. Our analysis involved computing diversity in windows surrounding functional elements and comparing the diversity patterns with those seen in *M. m. castaneus.* In order to aid visual comparisons, we divided nucleotide diversity (π) at all positions by the mean at physical distances greater than 75Kbp and 4Kbp away from exons and CNEs, respectively. When comparisons were made on the scale of genetic distance, we divided π by its mean at distances greater than $4N_er =$ 1,500 for protein-coding exons and $4N_er = 200$ for CNEs. These distances were chosen because they are the values beyond which π remains approximately constant.

In our simulations, we scaled recombination, mutation and selection parameters by N in a linear fashion. However, linear scaling can become problematic when selection coefficients are strong (Uricchio and Hernandez, 2014). To test whether linear scaling was appropriate for the parameters we estimated, we simulated populations with N = 100, 500, 750, 1,000 and 2,000. We found that patterns of genetic diversity converged in populations with N = 750, 1,000 and 2,000 (Figure C.2). The following simulations results were obtained assuming N = 1,000.

Simulations incorporating only deleterious mutations predicted a chromosome-wide reduction in genetic diversity. Around exons and CNEs, diversity plateaued at 94% of the neutral expectation (Figure C.3-C.4). Simulations involving only BGS did not fully predict the observed troughs in diversity around functional elements. Specifically, the predicted troughs in diversity around both protein-coding exons and CNEs, were not as wide nor as deep as those observed in the real data (Figures 3.2-3.3; C.3-C.4). Similar predictions were obtained for Models A or B (Figures 3.2-3.3; C.3-C.4) and for the gamma dDFEs inferred by Halligan et al. (2013) (Figure C.5). Our simulations incorporating deleterious mutations suggest, then, that while BGS affects overall genetic diversity across much of the genome, positive selection presumably also makes a substantial contribution to the dips in diversity around functional elements.

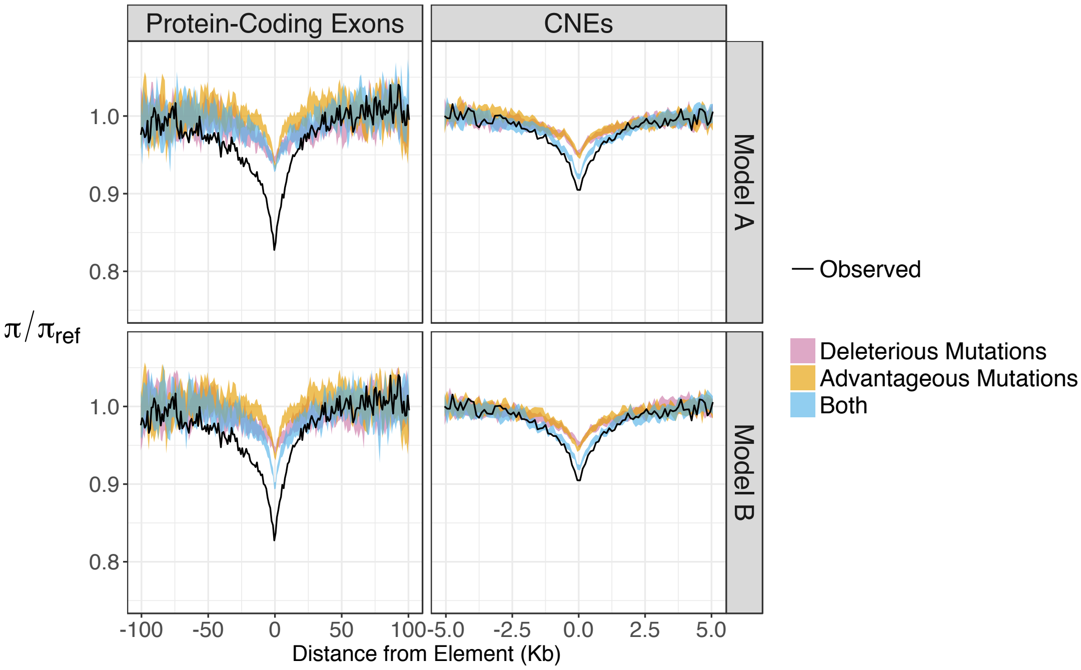


Figure 3.2: Estimates of scaled diversity (π/π_{Ref}) around protein-coding exons and CNEs (black lines) in *M. m. castaneus* compared to results from simulations (colored ribbons). The panels show diversity observed in simulated populations assuming DFE estimates obtained by analysis of the full uSFS (Model A) or when sites fixed for the derived allele do not influence selection parameters (Model B). Colored ribbons represent 95% confidence intervals obtained from 1,000 bootstrap samples.

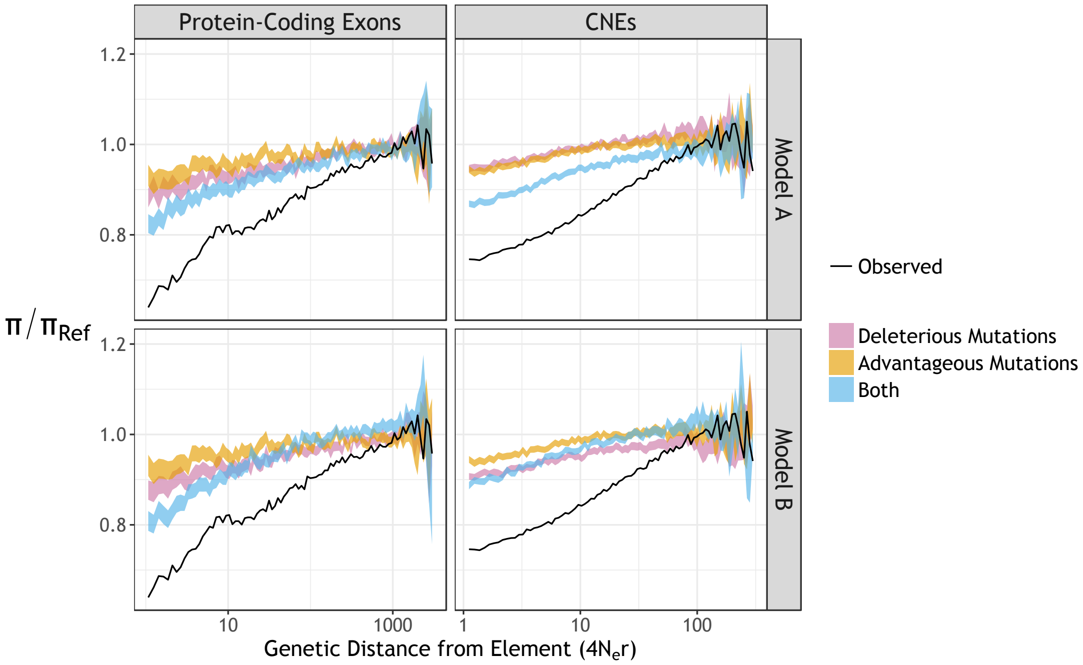


Figure 3.3: Estimates of scaled diversity (π/π_{Ref}) plotted against genetic distance from exons and conserved non-coding elements (CNEs) in *M. m. castaneus* compared to results from simulations (colored ribbons). The panels show diversity observed in simulated populations assuming DFE estimates obtained by analysis of the full uSFS (Model A) or when sites fixed for the derived allele do not influence selection parameters (Model B). Nucleotide diversity (π) is scaled by the mean diversity at population-scaled genetic distances ($4N_er$) more than 1,500 from exons and 200 from CNEs (π_{Ref}). Colored ribbons represent 95% confidence intervals obtained from 1,000 bootstrap samples. Diversity downstream of functional elements is shown.

In our simulations of exons and surrounding regions, recurrent SSWs produced troughs in diversity, but they were both narrower and shallower than those observed in the house mouse. However, the results are sensitive to the model used to estimate selection parameters (Figure 3.2; Table 3.3). Assuming the selection parameters estimated under Model A (i.e. analysing the full uSFS) we found that advantageous mutations produced a small dip in diversity around exons, which was both shallower and narrower than the one generated by deleterious mutations alone (Figure 3.2; Table 3.3). In contrast, the advantageous mutation parameters estimated under Model B (i.e. where sites fixed for the derived allele do not influence selection parameters) resulted in a marked trough in diversity around exons in simulations (Figure 3.2; Table 3.3). In simulations incorporating both advantageous and deleterious mutations, the troughs in diversity around exons were not as large as those observed in*M. m. castaneus* (Figure 3.2; Table 3.3), and even at very large distances from exons, diversity was around 90% of neutral expectation (Figure **??**). Assuming Model B selection parameters resulted in a trough in diversity that was both deeper and wider than the one generated when assuming Model A parameters (Figure 3.2). The differences between Model A simulations and Model B simulations presumably arise because under Model B the frequency of advantageous nonsynonymous mutations was 3 times higher than under Model A (Table 3.2). When analysis windows are binned based on genetic distance rather than physical distance, the differences between observed and simulated diversity patterns are even more striking (Figure 3.3 and Figure **??**).

We also carried out simulations focussing on CNEs and found that the combined effects of BGS and recurrent SSWs, as generated by our estimates of selection parameters, resulted in diversity troughs that were slightly shallower than observed (Figure 3.2; Table 3.3). Selection parameters obtained under Models A and B produced similar results. The troughs in diversity around CNEs in simulations incorporating only advantageous mutations were slightly shallower than the ones generated by deleterious mutations alone (Figure **??**). The troughs in diversity around CNEs in our simulations were also slightly shallower than those observed (Figure 3.2). This could be because we failed to detect infrequent, strongly selected advantageous mutations in CNEs or we underestimated the true frequency of advantageous mutations occurring in those elements. When plotted on the scale of genetic distance, the differences between our simulated data and the *M. m. castaneus* data become strikingly apparent (Figure 3.3).

ii) The site frequency spectrum around functional elements

SSWs and BGS are known to affect the shape of the SFS for linked neutral sites (Braverman et al., 1995; Charlesworth et al., 1995b; Kim, 2006). SSWs and BGS generate troughs in diversity at linked sites (Figures 3.2 and 3.3), but nucleotide diversity on its own does not contain information about the shape of the SFS. Tajimas D is a useful statistic for this purpose, because it is reduced when there is an excess of rare polymorphisms relative to the neutral expectation and increased when intermediate frequency variants are more common (Tajima, 1989). We therefore compared Tajimas D in the regions surrounding functional elements in simulations with values observed in the real data. It is notable that average Tajimas D is far lower in *M. m. castaneus* than in our simulations (Figure 3.4). This likely reflects a genome-wide process, such as population size change, that we have not modelled.

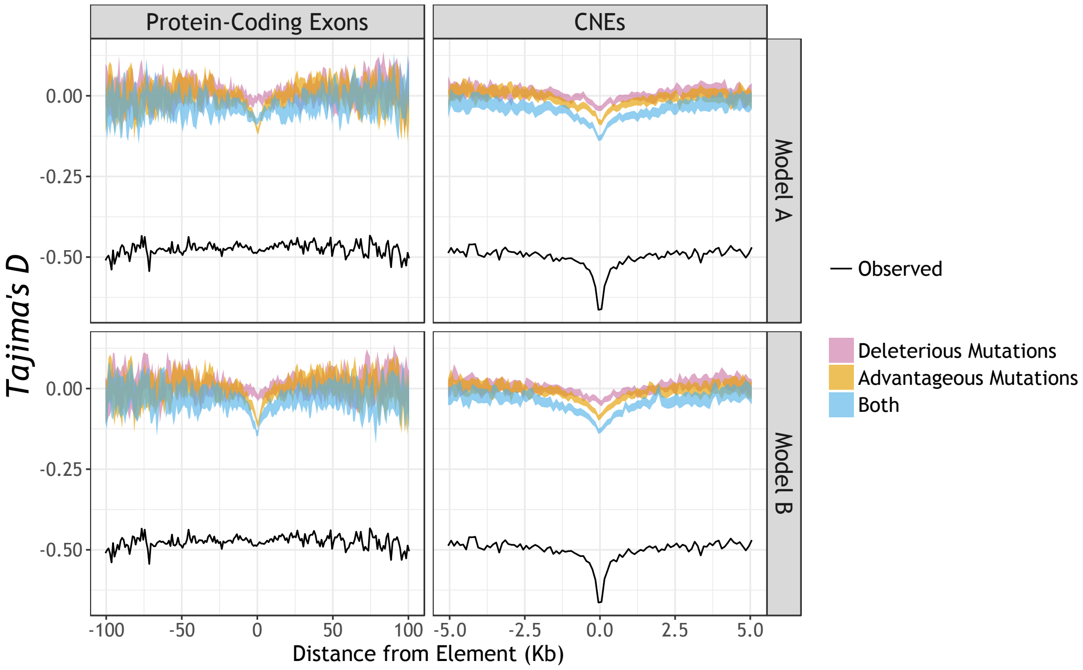


Figure 3.4: Tajimas D around protein-coding exons and CNEs in *M. m. castaneus* compared to simulated data. The black lines show Tajimas D computed from the *M. m. castaneus* genome sequence data around protein-coding exons or CNEs. The coloured ribbons show the 95% bootstrap intervals from simulated data assuming the DFEs estimated under either Model A (i.e. analysing the full uSFS) or Model B (i.e. fixed derived sites do not contribute to the likelihood for selection parameters)

If we assume selection parameters obtained under Model A, Tajimas D around protein-coding exons is relatively invariant, and matches the pattern observed in the real data fairly well (Figure 3.4). However, under Model B, the simulations exhibit a marked dip in Tajimas D, which is not observed in the real data (Figure 3.4).

In the case of CNEs, we observed a trough in Tajimas D in the real data (Figure 3.4), and simulations predict similar troughs under Models A and B (Figure 4). However, the trough in Tajimas D may be caused by the presence of functionally constrained sequences in the immediate flanks of CNEs (See Methods), making a comparison between the simulations and the observed data problematic.

iii) Rates of substitution in functional elements

Incorporating information from sites fixed for the derived allele when estimating the DFE (as in Model A) or disregarding this information (as in Model B) had a striking effect on estimates of the frequency and effects of advantageous mutations (Table 3.2). In the case of 0-fold sites, for example, p_a was 3x higher under Model B than Model A (Table 3.2). We therefore investigated the extent by which such differences affect the divergence at selected sites under the two models. Nucleotide divergence at putatively neutral sites between the mouse and the rat is approximately 15%, so we simulated an expected neutral divergence of 7.5% for one lineage.

We compared the ratio of nucleotide divergence at selected sites to the divergence at neutral sites ($d_{[}sel]/d_{[}neut]$) between the simulated and observed data. In simulations that assumed the selection parameters obtained under Model A, $d_{[}sel]/d_{[}neut]$ values were similar to those observed in *M. m. castaneus* for all classes of selected sites (Table 3.4). Under Model B, however, the simulations predicted substantially more substitutions at nonsynonymous sites and UTRs than were seen in the real data (Table 3.4). This suggests that, under Model B, the frequency of advantageous mutations for 0-fold sites and UTRs may be overestimated.

iv) Examining the uSFS and estimating the DFE from simulated data

BGS and SSWs both perturb allele frequencies at linked neutral sites, distorting site frequency spectra, which can lead to the inference of spurious demographic histories (Messer and Petrov, 2013; Ewing and Jensen, 2016; Schrider et al., 2016). By fitting a model incorporating three epochs of population size to the putatively neutral site data, we inferred that *M. m. castaneus* has experienced a population bottleneck followed by an expansion (Table C.2). To investigate the possibility that the inferred demographic histories could be an artefact of selection at linked sites, we fitted demographic models to the uSFS obtained from simulated synonymous sites. Visual comparison of the uSFS from simulated populations with the uSFS obtained from *M. m. castaneus* reveals that our simulations do not fully capture the excess of high frequency variants observed in the mouse population (Figure C.6). However, for simulations assuming the selection parameters obtained under either Model A or B, a 3-epoch population size model gave the best fit to the data. The estimated demographic histories were somewhat different between simulations assuming Model A or Model B, but in each case a population bottleneck followed by an expansion was inferred (Table C.6). This is an interesting observation: our simulations assumed a constant population size, but selection at linked sites appears to distort the neutral uSFS, such that a demographic history similar to the one inferred from the real data is estimated (Table C.6).

Our simulations also indicate that selection parameters are difficult to accurately infer using the uSFS alone. In the case of Model A simulations, the selection strength and frequency of deleterious mutations was accurately estimated, as was the combined frequency of all effectively neutral mutations (Table C.6). However, in Model A simulations, DFE-alpha did not accurately estimate the strength and frequency of advantageous mutations. Estimates of selection parameters in Model B simulations were similar to the input parameters, but a notable exception was that the frequency of advantageous mutations (p_a) was overestimated (Table C.6). These results suggest that some features of the uSFS inferred for *M. m. castaneus* have been captured by our simulated data. However, the demographic correction which we applied to the uSFS before estimating selection parameters (see Supplementary Methods), had a substantially greater impact on the *M. m. castaneus* data than for the simulated data, particularly in the case of high frequency derived alleles (Figure C.6). A possible explanation is that strong SSWs, which cause a greater increase in the proportions of high frequency derived alleles for linked sites (Kim, 2006), have left a signal in the neutral site uSFS, but these cannot be accurately inferred from the selected site uSFS itself. If this were the case, then there may be information in the uSFS for linked neutrally evolving sites that could be used when estimating selection parameters. This would require the expected uSFS arising under the joint effects of SSWs and BGS.

Table 3.3: The root mean square difference between values of π around functional elements predicted in simulations and π observed in *M. m. castaneus.* Confidence intervals were obtained from 1,000 bootstrap samples (see Methods). Values shown are for the patterns of diversity when measured on the scale of physical distance.

		Exons		CNEs	
		Median	95% range	Median	95% range
Model A	Deleterious Mutations	0.0327	0.0311 - 0.0343	0.0164	0.0151 - 0.0179
	Advantageous Mutations	0.0422	0.0403 - 0.0442	0.0177	0.0161 - 0.0195
	Both	0.0312	0.0297 - 0.0340	0.01	0.0088 - 0.0113
Model B	Deleterious Mutations	0.0331	0.0314 - 0.0351	0.0157	0.0144 - 0.0171
	Advantageous Mutations	0.038	0.0355 - 0.0406	0.0162	0.0147 - 0.0179
	Both	0.0274	0.0253 - 0.0294	0.0088	0.0078 - 0.0101

v) Patterns of diversity around sites that have recently experienced a substitution

In general, it has been difficult to discriminate between BGS and SSWs, because their effects on genetic diversity and the site frequency spectrum are qualitatively similar. It has been suggested that the two processes can be teased apart by taking advantage of the fact that hard SSWs should be centred on a nucleotide substitution, whereas this is not the case for BGS. Comparing the average genetic diversity in regions surrounding recent putatively selected and putatively neutral substitutions (e.g. 0-fold and 4-fold sites, respectively) may therefore reveal the action of SSWs (Hernandez et al., 2011; Sattath et al., 2011). Halligan et al. (2013) performed such an analysis in *M. m. castaneus* using the closely related *M. famulus* as an outgroup, and found that the profiles of neutral diversity around 0-fold and 4-fold substitutions were virtually identical. Similar findings have been reported in other species (Hernandez et al., 2011; Beissinger et al., 2016). One interpretation of these results is that hard SSWs are rare. To investigate this, we measured the average neutral diversity around nonsynonymous and synonymous substitutions in simulations for the case of frequent hard SSWs.

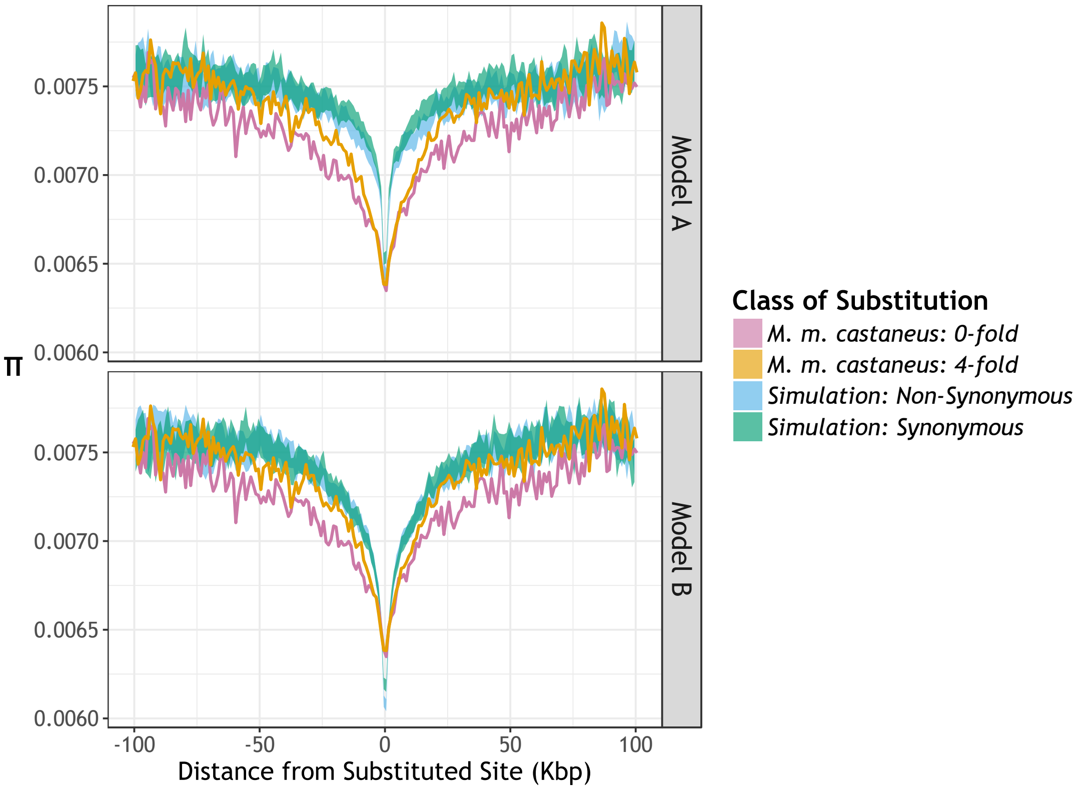


Figure 3.5: Nucleotide diversity (π) around substituted sites in *M. m. castaneus* compared to the same pattern obtained from simulation data. Nucleotide diversity in *M. m. castaneus* was scaled by divergence between mouse and rat to correct for variation in local mutation rates. The *M. m. castaneus* data are from Halligan et al. (2013).

In our simulations, we measured diversity around substitutions occurring on a time-scale equivalent to the divergence time between *M. m. castaneus* and *M. famulus.* The average diversities around nonsynonymous and synonymous substitutions in the simulated data were very similar, regardless of whether simulations assumed Model A or Model B selection parameters (Figure 3.5). However, the troughs in diversity around substitutions were deeper in the simulations assuming Model B (Figure 3.5), reflecting the higher frequency of advantageous mutations (Table 3.2). In the immediate vicinity of nonsynonymous substitutions, diversity was lower than the corresponding value for synonymous substitutions (Figure 3.5). However, the differences are slight, so it would be difficult to draw firm conclusions about the action of either SSWs or BGS. Taken together, these results suggest that analysing patterns of diversity around recent substitutions does not provide enough information to convincingly discriminate between SSWs and BGS in *M. m. castaneus*, even when hard sweeps are fairly frequent. Further analysis is required to assess whether this is also the case for other organisms.

Table 3.4: Comparison of the accumulation of nucleotide divergence in simulated populations between different functional site types. In the cases of 0-fold sites and UTRs, d_{neu} refers to 4-fold sites. For CNEs, d_{neu} refers to CNE flanking sites. In all simulations, d_{neu} was set to 7.5%.

		Simulation DFE			
	M. m. castaneus	Model A		Model B	
Site Class	d_{sel}/d_{neu}	d (%)	d_{sel}/d_{neu}	d (%)	d_{sel}/d_{neu}
0-fold	0.225	1.66	0.221	2.26	0.301
UTR	0.757	5.76	0.767	6.85	0.914
CNE	0.406	3.31	0.44	3.07	0.409

3.5 Discussion

There are a number of observations suggesting that natural selection is pervasive in the murid genome. First, there is a positive correlation between synonymous site diversity and the rate of recombination (Booker, et al. 2017b). Secondly, there is reduced diversity on the X-chromosome compared to the autosomes, which cannot readily be explained by neutral or demographic processes (Baines and Harr 2007).

Thirdly, there are troughs in genetic diversity surrounding functional elements, such as protein-coding exons and CNEs, which are consistent with the action of BGS and/or SSWs (Halligan, et al. 2013). In this paper, we analysed the sequences of 10 *M. m. castaneus* individuals sampled from the ancestral range of the species (Halligan, et al. 2013). We estimated the DFEs for several classes of functional sites (0-fold nonsynonymous sites, UTRs and CNEs), and used these estimates to parametrise forward-in-time simulations. We investigated whether the simulations predict the observed troughs in diversity around functional elements along with the between-species divergence observed between mice and rats.

3.5.1 Estimating selection parameters based on the uSFS

Relative to putatively neutral comparators, 0-fold sites, UTRs and CNEs all exhibited reduced nucleotide diversity, reduced nucleotide divergence and an excess of low frequency variants (Table 3.1; Figure 3.1), consistent with the action of natural selection (Halligan et al., 2010, 2013). The estimates of the DFEs included substantial proportions of strongly deleterious mutations (Table 3.2). In addition, the best-fitting models also included a single class of advantageous mutations. Although additional classes were not statistically supported, in reality, there is almost certainly a distribution of advantageous selection coefficients (Bank et al., 2014; McDonald et al., 2016). A visual examination of the fitted and observed uSFSs, however, shows that the estimated DFEs fitted the data very well (Figure **??**), suggesting that there is limited information in the uSFS to estimate a range of positive selection coefficients.

When estimating the DFE for a particular class of sites, we analysed either the full uSFS including sites fixed for the derived allele (Model A) or we ignored sites fixed for the derived allele (i.e. Model B). Recently,Tataru et al. (2017) showed that selection parameters can be accurately estimated from the uSFS in simulations that ignored between-species divergence, if the frequency of advantageous mutations (p_a) is sufficiently high. In our analysis of 0-fold sites and UTRs, Model B gave a significantly better fit and higher estimates pa than Model A (Table 3.2). For CNEs, however, Models A and B did not significantly differ in fit, and the selection parameter estimates were very similar (Table 3.2). The goodness-of-fit and parameter estimates obtained under Models A and B may differ if the processes that generated between species-divergence are decoupled from the processes that produce within species diversity. There are several factors that could potentially cause such a decoupling. 1) Past demographic processes may have distorted the uSFS in ways not captured by the corrections we applied; 2) there may be error in assigning alleles as ancestral or derived; 3) the nature of the DFE may have changed since the accumulation of between-species differences began; and 4) there could be rare, strongly advantageous mutations that contribute to divergence, but contribute negligibly to polymorphism. It is difficult to know which of these factors affected the outcome of our analyses. However, we found that Model B gave a better fit to the uSFS than Model A for 0-fold sites and UTRs, but not CNEs. There is an a priori expectation that the strength of selection on protein-coding sequences is greater than regulatory sequences Halligan et al. (2013), so we think the latter factor is likely to have been important.

3.5.2 Patterns of diversity and Tajimas D around functional elements

We performed simulations incorporating our estimates of deleterious and advantageous mutation parameters to dissect the contribution of BGS and selective sweeps to diversity dips around functional elements. Our simulations suggest that BGS and SSWs both produce genome-wide reductions in neutral diversity (Figures C.3-C.4), but neither process on its own fully explains the troughs in diversity around protein-coding exons and CNEs, regardless of which model (A or B) is used to estimate selection parameters (Figures 3.2-3.3). Around protein-coding exons, the combined effects of advantageous and deleterious mutations generated a shallower trough in diversity than the one observed (Figure 3.2). These patterns are qualitatively similar when analysing physical or genetic distances, but the differences between observed and simulated patterns are more apparent when analysing genetic distances (Figure 3.3). A possible explanation for this is that rare, strongly selected advantageous mutations are undetectable by analyses based on the uSFS (discussed below). In contrast, the combined effects of BGS and SSWs predicted troughs in diversity surrounding CNEs that closely match those observed, when measured on the physical scale (Figure 3.2). Troughs of diversity around CNEs on the scale of genetic distance are not nearly so similar in simulated populations to those observed in *M. m. castaneus* as they are on the physical scale. Specifically, the observed trough in diversity around CNEs on the genetic distance scale is both deeper and wider than in simulated populations. Analysing patterns of diversity on the physical scale is analogous to assuming that there is a uniform recombination rate across the genome. Our results highlight the importance of incorporating recombination rate variation when performing such analyses, particularly in species that exhibit highly variable recombination rates.

There is an excess of rare variants in *M. m. castaneus* relative to the neutral expectation, as indicated by a strongly negative Tajimas Ds for putatively neutral sites (Table 1) and for regions surrounding exons and CNEs (Figure 3.4). Our simulations incorporating both advantageous and deleterious mutations also exhibited negative Tajimas D, but not nearly so negative as in the real data (Figure 3.4). This difference between the observed data and the simulations indicates that there may be processes generating an excess of rare variants, such as a recent population expansion, which were not incorporated in the simulations.

3.5.3 Rates of nucleotide substitutions in simulations

Our simulations suggest that the frequency of advantageous mutations (pa) estimated for 0-fold sites and UTRs under Model B may be unrealistically high. This is because several aspects of the results were incompatible with the observed data. Firstly, we found that the substitution rates for simulated nonsynonymous and UTR sites were higher than those observed between mouse and rat (Table 3.4). Secondly, we observed a pronounced dip in Tajimas D around simulated exons, which is not present in the real data (Figure 3.4), suggesting that under Model B, either the strength or frequency of positive selection at 0-fold sites is overestimated.

3.5.4 Do our results provide evidence for strongly selected advantageous mutations?

Estimation of the strength and frequency of advantageous mutations based on the uSFS relies on the presence of positively selected variants segregating within the population (Boyko et al., 2008; Schneider et al., 2011; Tataru et al., 2017). The frequency of advantageous mutations may impose a limit on the parameters of positive selection that can be accurately estimated. Indeed, Tataru et al. (2017) recently showed that p_a may be overestimated when analysing the uSFS, if the true value of p_a is low.

If advantageous mutations are infrequent, those with larger effects on fitness will be less likely to observed segregating than those with milder effects, as strongly selected mutations have shorter sojourn times than weakly selected ones (Fisher 1930; Kimura and Ohta 1969). This could explain why the selection parameters we estimated fail to predict the troughs in diversity observed in the real data (Figure 3.2). Furthermore, the fact that Model B gave a better fit than Model A for 0-fold sites and UTRs suggests that polymorphism and divergence have become decoupled for those sites. This is also consistent with the presence of infrequent, strongly selected mutations that become fixed rapidly and are thus not commonly observed as polymorphisms.

Relevant to this point, an interesting comparison can be made between two recent studies to estimate the frequency and strength of positive selection using the same *D.*

melanogaster dataset. The first, by Keightley et al. (2016), used the uSFS analysis methods of Schneider et al. (2011) (i.e. Model A in the present study), and estimated the frequency of advantageous mutations (p_a) = $4.5x10^{-3}$ and the scaled strength of selection ($2N_es_a$) = 23.0 for 0-fold nonsynonymous sites. In the second study, Campos et al. (2017) estimated p_a = 2.2 x 10^{-4} and $2N_es_a$ = 241, based on the correlation between synonymous site diversity and nonsynonymous site divergence. Although the individual parameter estimates differ substantially, the compound parameter $2N_es_ap_a$ (which approximates the rate of SSWs) was similar between the studies (0.055 and 0.104 for Campos et al. (2017) and citeRN321 respectively). It is expected that synonymous site diversity is reduced by SSWs, so the method used by Campos et al. (2017) may be sensitive to the presence of strongly selected mutations, whereas the Keightley et al. (2016) approach may have been more sensitive to weakly selected ones. It seems plausible then, that the two studies capture different aspects of the DFE for advantageous mutations (a similar argument was made by Sella, et al. (2009)). Supporting this view, Elyashiv et al. (2016) recently estimated the DFE in *D. melanogaster*, incorporating both strongly and weakly selected advantageous mutations, by fitting a model incorporating BGS and SSWs to genome-wide variation in genetic diversity. They inferred that weakly selected mutations are far more frequent than strongly selected ones, but that both contribute to variation in genetic diversity across the *D. melanogaster* genome. In the present study, we used similar methods as Keightley et al. (2016) to estimate the frequency and strength of advantageous mutations, so the estimated parameters of positive selection may represent only weakly selected mutations. Indeed, patterns of diversity at microsatellite loci suggest that there are strongly selected, infrequent sweeps in multiple European *M. musculus* populations (Teschke et al., 2008), so infrequent strong sweeps may be a general feature of mouse evolution.

We tested the hypothesis that undetected, strongly selected mutations are chiefly responsible for the reductions in diversity around functional elements by performing additional simulations. In this exercise, we assumed far stronger selection on advantageous mutations for protein-coding regions than we estimated for 0-fold sites (Table C.7). When modelling strong selection, we reduced pa such that rate of sweeps (proportional to the product $2N_e s_a p_a$) was either that estimated from the uSFS under Model A, or double that value (Table C.7). As can be seen in Figure 3.6, increasing the strength of selection acting on advantageous mutations, while simultaneously decreasing the p_a parameter resulted in troughs in diversity around protein-coding exons that were both deeper and wider than those observed in *M. m. castaneus*. By chance, we identified a set of parameters ($2N_e s_a = 400$; $p_a = 0002$) that can provide a relatively close correspondence between simulated and observed patterns of diversity. However, it must be stressed that these parameters were chosen arbitrarily and that there is no statistical support for them.

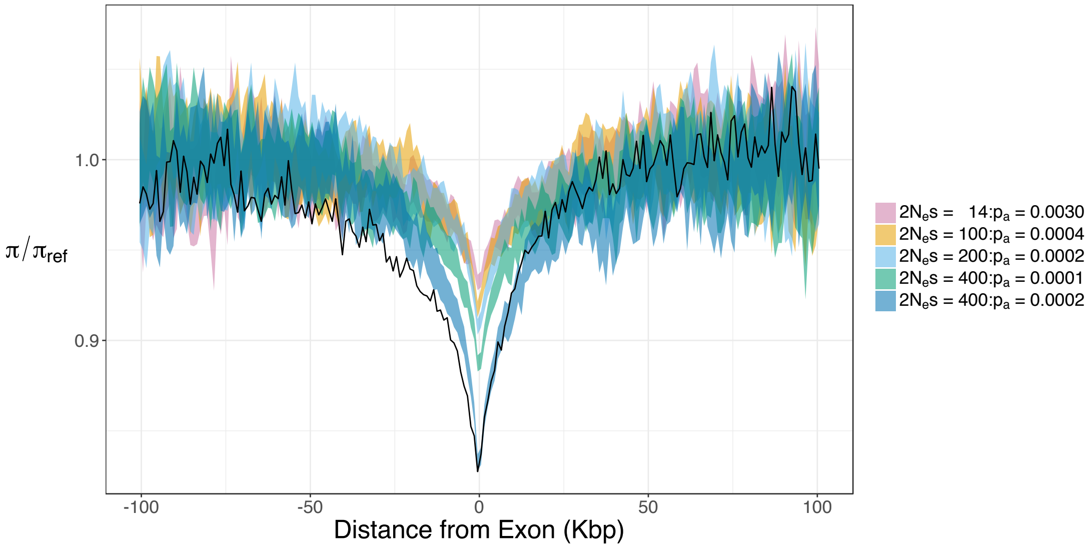


Figure 3.6: Patterns of nucleotide diversity around protein-coding exons in simulations assuming strongly selected advantageous mutations. Nucleotide diversity in *M. m. castaneus* is shown in black. Nucleotide diversity (π) is scaled by the mean diversity at distances more than 75 Kbp from exons (π_{Ref}). Coloured ribbons represent 95% confidence intervals obtained from 1,000 bootstrap samples.

Although strongly selected mutations can generate a similar trough in average diversity around protein-coding exons as observed in the real data, they do not produce the apparent genome-wide reduction in Tajimas D observed in *M. m. castaneus*

(Figure C.8). Indeed, in all simulations modelling strongly advantageous mutations, we observed a trough in Tajimas D that plateaued at values close to 0 in regions surrounding protein-coding exons (Figure C.8). In this exercise, we manipulated the selection parameters for 0-fold sites only, but it is possible that there are strongly advantageous mutations in all of the site classes. A combination of DFE parameters for the different functional elements in mice could therefore explain the reductions in diversity and the genome-wide negative Tajimas D. On the other hand, it could also be that recent demographic processes have swamped the signal of linked positive selection in the site frequency spectrum. Our results highlight the need for methods that can simultaneously estimate selection parameters for multiple functional elements and demographic history.

Understanding the contributions of regulatory and protein change to phenotypic evolution has been an enduring goal in evolutionary biology King and Wilson (1975); Carroll (2005); Franchini and Pollard (2017). If selection is strong relative to drift (i.e. $2N_e s_a$ ¿ 1) then the rate of change of fitness from the fixation of advantageous mutations is expected to be proportional to the square of the selection coefficient (Falconer and Mackay 1996). In this study, we inferred that the strength of selection acting on new advantageous mutations in CNEs and 0-fold sites are roughly equivalent, but that advantageous mutations occur more frequently in CNEs (Table 3.2). Given that there are more CNE nucleotides in the genome than there are 0-fold sites (Table 3.1), this could imply that adaptation at regulatory sites causes the greatest fitness change in mice. However, our analyses suggest that both protein-coding genes and CNEs may experience strongly selected advantageous mutations, which are undetectable by analysis of the uSFS. If this were the case, protein-coding mutations could make a larger contribution to fitness change than mutations in regulatory sites.

3.5.5 Limitations of the study

There is a growing body of evidence suggesting that hard sweeps may not be the primary mode of adaptation in both *D. melanogaster* and humans. Firstly, soft sweeps, where multiple haplotypes reach fixation due to the presence of multiple de novo mutations or selection acted on standing variation, may be common. Garud et al. (2015) developed a suite of haplotype-based statistics that can discriminate between soft and hard SSWs. The application of these statistics to North American and Zambian populations of *D. melanogaster* suggested that soft sweeps are the dominant mode of adaptation in that species, at least in recent evolutionary time (Garud et al., 2015; Garud and Petrov, 2016). Furthermore, Schrider and Kern (2016) recently reported that signatures of soft sweeps are more frequent than those of hard sweeps in humans. However, their method did not explicitly include the effects of partial sweeps and/or BGS. Under a model of stabilising selection acting on a polygenic trait, if the environment changes, adaptation to a new optimum may cause small shifts in allele frequency at numerous loci without necessarily resulting in fixations REF NEEDED(Barton and Keightley 2002; Pritchard, et al. 2010). Genome-wide association study hits in humans exhibit evidence that such partial SSWs may be common (Field et al., 2016). These results all suggest that the landscape of adaptation may be more complex than the model of directional selection acting on a *de novo* mutation assumed in this study. For example, our simulations did not incorporate changing environments or stabilising selection, so we were unable to model adaptive scenarios other than hard sweeps.

Further work should aim to understand the probabilities of the different types of sweeps. Different functional elements have different DFEs for harmful mutations. In particular, regulatory elements seem to experience more mildly selected deleterious mutations than coding sequences (Halligan et al., 2013; Carroll, 2005) (Table 3.2). It has been argued that such differences in constraint between coding and non-coding elements may be due to a lower pleiotropic burden on regulatory sequences (Carroll, 2005). Differences in the DFE among different genomic elements is expected to affect genetic diversity within these elements. This, in turn, may affect the types of sweeps that occur, since the relative probabilities of a hard versus soft sweep depend on the level of standing genetic variation (reviewed in Hermisson and Pennings (2017)).

In our simulations, we treated Ne as constant through time, but this is an oversimplification. We analysed two different classes of putatively neutral sites, and inferred there has been a population size bottleneck followed by an expansion (Table C.2). In our simulations, however, we showed that the inferred demographic history may largely be an artefact of selection at linked sites (Table C.6). There is a strongly negative Tajimas D in genomic regions far from functional elements, which is not explained by selection (or at least the selection parameters we inferred) (Figure 3.4). This reduction is presumably caused by a demographic history or strong selection that was not included in our simulations. Better estimates of the demographic history of*M. m. castaneus* may be obtained, for example, from regions of the genome experiencing high recombination rates, located far from functional elements. Finally, the size of mouse populations may oscillate seasonally REF NEEDED(Pennycuik, et al. 1986) and if this were the case, so would the effective selection strength of new mutations (and thus the probabilities of SSWs) (Otto and Whitlock, 1997).

In house mice, crossing over events predominantly occur in narrow windows of the genome termed recombination hotspots Brick et al. (2012). The locations of recombination hotspots have evolved rapidly between and within *M. musculus* subspecies (Smagulova et al., 2016), but at broad-scales recombination rates are relatively conserved (Booker et al., 2017b). Assuming a single suite of recombination hotspots in simulations may produce misleading results if hotspot locations evolve faster than the rate of neutral coalescence. While we included fine-scale variation in recombination rates in our simulations, we used a recombination map was inferred at a broader scale than the scale of hotspots (Booker et al., 2017b). However, hotspots are an important feature of the recombination landscape in mice and thus potentially influence the patterns of diversity around functional elements, but the appropriate way to model them is unclear.

3.6 Conclusions

Using simulations, we have shown that estimates of the DFE obtained by analysis of the uSFS cannot fully explain the patterns of diversity around both CNEs and protein-coding exons. We argue that, while frequent mutations with moderate advantageous effects occur in different functional elements in the mouse genome (Table 3.2), strongly advantageous mutations that are undetectable by analysis of the uSFS generate the bulk of the reductions in diversity. Estimates of the strength and rate of advantageous mutations could be obtained by directly fitting a sweep model to the troughs in diversity around functional elements. We have shown that BGS makes a substantial contribution to these troughs, and applying models that incorporate both BGS and sweeps (Kim and Stephan 2000; Elyashiv, et al. 2016; Campos, et al. 2017) might allow us to make more robust estimates of selection parameters.

3.7 Chapter 3 Acknowledgements

We owe thanks to Brian Charlesworth, Ben Jackson and two anonymous reviewers for useful comments on the manuscript and to Deborah Charlesworth, Dan Halligan, Rory Craig and the evolutionary genetics lab group at the University of Edinburgh for helpful discussions throughout the course of the project.

Chapter 4

Estimating parameters of selective sweeps from patterns of genetic diversity in house mice

I designed this analysis with Brian Charlesworth. I performed all analyses.

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 nonsynonymous 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* by fitting 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 (**?**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 advantageous mutations ($2N_es \approx$ 5,000) occur in the fruitfly's genome. In the house mouse, there is also a trough in diversity around recent nonsynonymous substitutions, but an almost identical trough is observed around synonymous substitutions, furthermore a similar trough is observed around even randomly selected synonymous and nonsynonymous sites in the genome (Halligan et al., 2013). This all, perhaps, suggests that the reductions in diversity caused by selection at linked sites extend beyond the average distance separating nonsynonymous substitutions, so that the methods employed by Sattath et al. (2011) are not effective in mice (Halligan et al., 2013). 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.

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

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. Using simulations, we show that the selection parameters that explain the troughs in diversity are out of the range detectable by analysis of the uSFS. We find that, as expected *a priori*, the strength of selection acting on protein-coding exons is far greater than that acting in regulatory elements. Finally, using a simple model of the fitness change brought about by adaptive evolution, we show that, despite adaptation occurring more frequently in regulatory regions, adaptation in protein-coding regions may contribute more to phenotypic evolution in mice.

Materials and Methods

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 of neutrally evolving intronic sequence. Nonsynonymous mutations were modelled as 75% of mutations occurring in exons, the remaining 25% were strictly neutral (i.e. synonymous sites). We varied the γ_a and p_a parameters across simulations, but kept the product $\gamma_a p_a$ equal to 0.1. We based this value of $\gamma_a p_a \approx 0.1$ on a recent study in *Drosophila melanogaster* (Keightley et al., 2016). All simulations incorporated the same gamma dDFE (β = 0.2 and $\hat{\gamma_d}$ = -1,000). The advantageous mutation parameters we simulated are listed in Table **??**. The population-scaled mutation and recombination rates (i.e. $\theta = 4N_e\mu$ and $\rho = 4N_er$, respectively) were set to 0.01. Populations of N = 1,000 diploid individuals were simulated for an initial burn-in of

$10N$ generations to establish equilibrium conditions. After the burn-in, 20 haploid chromosomes were sampled every $2N$ generations for a further $100N$ generations. We performed 10 simulation replicates for each set of selection parameters (Table **??**). Across simulation replicates, time-points and loci we extracted the simulated nonsynonymous and synonymous sites, giving uSFS data for 10,000 'genes'. We sampled the set of 10,000 'genes' with replacement 100 times, collating the nonsynonymous and synonymous site uSFSs for each replicate.

Analysis of the uSFS

We estimated the DFEs in our simuations by analysing the uSFS using the methods of Tataru et al. (2017) as implemented in the polyDFE (v1.1) package. PolyDFE fits an expression for the uSFS expected in the presence of both advantageous and deleterious mutations to data from putatively neutral and selected classes of sites, by maximum likelihood. The neutral class uSFS is used to determine distortions to the uSFS caused by processes such as selection at linked sites and a history of population size change. In addition, polyDFE corrects for polymorphism misattributed to divergence, mutation rate variability and error in assigning sites as ancestral/derived. Tataru et al. (2017) performed extensive simulations and showed that accurate estimates of the parameters for both deleterious and advantageous mutations can be obtained using their methods. However, there are a range of parameters that they did not test which may be biologically relevant, specifically when advantageous mutations are strongly selected, but infrequent.

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

Model of Recurrent Sweeps with Background Selection

? gave expressions for the neutral diversity expected under the combined effects of background selection (BGS) and selective sweeps (SSWs). They assumed that the effects of BGS and SSWs act independently so that their effects can simply be summed. However, background selection causes a reduction to the effective population size (N_e) at a neutral locus, j by some fraction B_j. The rate and fixation probability of new advantageous mutations is dependant upon N_e, so we scale the sweep effect by B_j in a modified version of the model used by **?**,

$$\frac{\pi_j}{\pi_0} \approx \frac{1}{B_j^{-1} + B_j 2N_e P_{sc,j}}. \tag{4.1}$$

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

$$P_{sc,j} \approx V_a \tau \gamma_a^{\frac{-4r_{i,j}}{s}} \tag{4.2}$$

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 proportion of new mutations that are advantageous and γ_a is the scaled selection coefficient ($2N_e s_a$) of those mutations(Kimura and Ohta?). τ is the number of selected sites in a functional element. The recombination fraction between a functional element (i) and the focal neutral site is $r_{i,j}$. When assuming that recombination proceeds solely by crossing over $r_{i,j}$ is simply the product of the physical distance ($d_{i,j}$) and the local crossing-over rate (r_c). When incorporating gene conversion, we use Equation 1 from Thornton (2014):

$$r_{i,j} = d_{i,j} r_c + g_c d_g \left(1 - e^{-\frac{d_{i,j}}{d_g}} \right) \tag{4.3}$$

where g_c is the rate of gene conversion and d_g is the mean gene conversion tract length, assuming that the distribution of tract lengths is exponential. When applying Equation 4.3 we use $g_c = \kappa r_c$ where κ is the ratio of the non-crossovers to crossovers.

Theory suggest that the distribution of fitness effects for advantageous mutations is likely exponential Hernandez and Uricchio (2015) Griffiths REFS). We incorporate an exponential distribution of advantageous mutation effects to Equation 3 as follows:

$$P_{sc,j} \approx \int_{0}^{\infty} f_x(\gamma) V_a \tau \gamma_a^{\frac{-4r_{i,j}}{s}} \, d\gamma \tag{4.4}$$

We estimated γ_a and p_a by fitting Equation 4.1 to the relationship between nucleotide diversity and distance to functional elements using non-linear least squares with the *lmfit* (0.9.7) package for Python 2.7. When analysing the mouse data, see below, we compared the fit of Equation 4.1 incorporating either one or two discrete classes of beneficial mutations (Equation 4.2) or the exponential distribution (Equation 4.4) using Aikieke's Information Criterion (AIC).

Analysis of Mouse Data

We analysed patterns of genetic diversity in 10 wild-caught *M. m. castaneus* individuals, first reported by Halligan et al. (2013). Breifly, Halligan et al. (2013) sequenced individual genomes to high coverage ($\approx$ 30x) using Illumina paired-end reads, which were mapped to the mm9 mouse reference genome using BWA. Variants were called using a Samtools pipeline. Note that we only analyse SNP data in this study, insertion/deletion variants are not included. For further details of the sequencing and variant calling methods see Halligan et al. (2013). Protein-coding exons present in the version 67 of the Ensembl annotation database and the locations of conserved non-coding elements identified by Halligan et al. (2013) using an alignment of placental mammals were used in this study. The mean length of a protein-coding exon is 151bp, of which we assume 75% of sites are subject to selection and the mean length of a conserved non-coding exon is 51bp, of which 100% of sites are subject to selection.

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

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

In order to disentangle the sweep parameters we estimate using Equation 4.1, we assume the point mutation rate for mice to be 5.4×10^{-9} (Uchimura et al., 2015).

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 diversity, B, to fit as a covariate when fitting Equation 4.1 to the diversity troughs. There are formulae for calculating B given the DFE as well as mutation and recombination rates (Nordborg et al., 1996; Hudson and Kaplan, 1995), but these over-predict the effects of BGS when purifying selection is weak ($\gamma_d < 1$) (Good and Desai; 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 Chapter 3, we used simulations to estimate the contribution of background selection to patterns of nucleotide diversity around both protein-coding exons and CNEs. These simulations incorporated recombination rate variation, the actual distribution of functional elements in the genome and dDFEs specific to each of the functional elements analysed. By extracting diversity as a function of genetic distance to both protein-coding exons and CNEs from these simulations, we obtained estimates of B that can be used when fitting Equation 4.1.

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

Results

Estimating selection parameters from the uSFS of simulated data

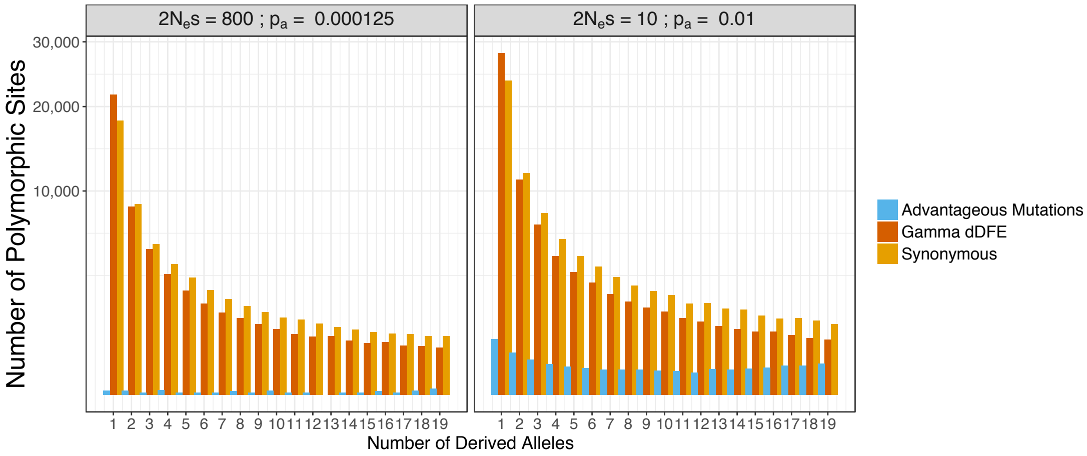


Figure 4.1: An example of the uSFSs for deleterious (Gamma dDFE) and advantageous nonsynonymous sites with neutral synonymous sites from simulated populations. Results shown are from simulations modelling strongly or weakly advantageous mutations. The uSFS model the same rate of synonymous substitutions $\gamma_a p_a = 0.1$. Simulated datasets included 10Mbp of exonic sites (3:1 nonsynonymous:synonymous sites).

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 (Schneider et al., 2011; Tataru et al., 2017), but it has been argued that strongly selected advantageous mutations, which may contribute little to standing variation, will be undetectable by such methods (**?**). In this study, we confirm this verbal argument using simulations, showing that accurate estimation of positive selection parameters does indeed depend on the strength and relative frequencies of advantageous mutations.

We used forward-in-time simulations that incorporated linkage, because selection at linked sites can distort the uSFS in ways that likely affect real data and thus cannot be ignored. For each set of advantageous mutation parameters, we simulated 10Mbp of gene-like sequences giving a total of 7.5Mbp of nonsynonymous sites and 2.5Mbp of synonymous sites which we used to construct the uSFS for 20 haploid individuals. This sample size and quantity of data is fairly typical of population genomic 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, the effects of demography and selection at linked sites.

Consistent with Tataru et al. (2017) we found that polyDFE gave estimates of the dDFE were very accurate. In particular, the shape parameter of the gamma dDFE was estimated with precision. Overall, the estimation performed most poorly when divergence was included, but only a dDFE was inferred. These results replicate the findings of Tataru et al. (2017) and further emphasize the importance of specifying a full DFE model when making inferences of selection from the uSFS.

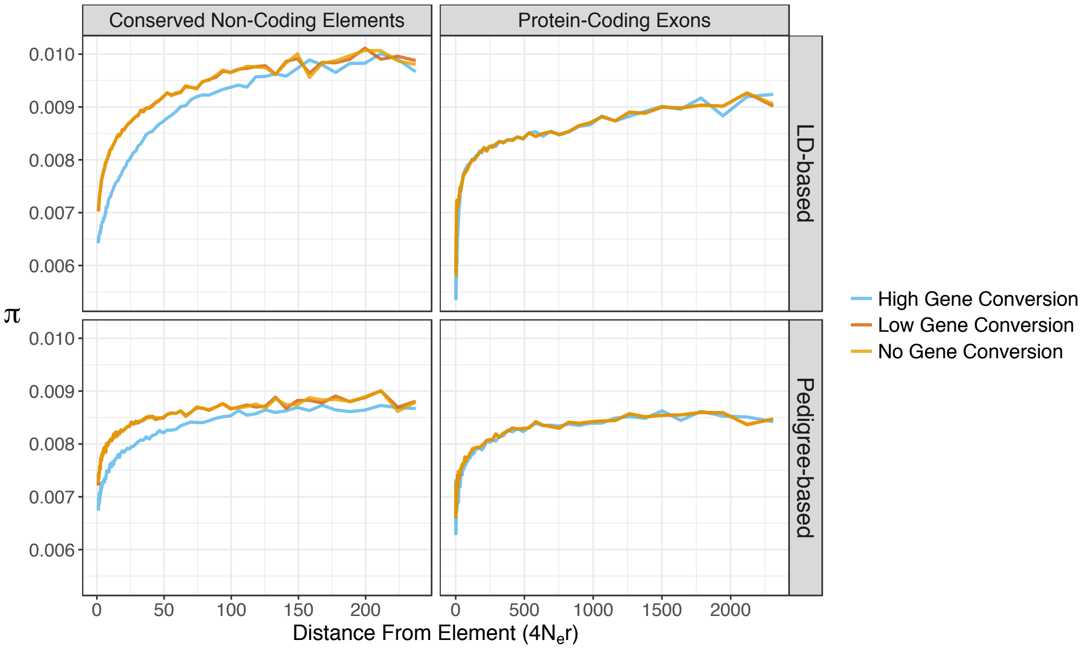
We analysed the uSFS from our simulated populations and found that when advantageous mutations are relatively frequent ($p_a > 0.0005$), but weakly selected ($\gamma_a < 100$), both γ_a and p_a parameters can be estimated with precision (Table REF). However, we found that when advantageous mutations 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 $\gamma_a \geq 100$ and likelihood ratio tests typically failed to detect positive selection (Table REF).

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. All simulations were subject to the same dDFE, so the extent of background selection should be fairly similar. We found that selection at linked sites reduced synonymous site diversity below the expectation value of 0.01 in all simulations (Table **??**), 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 **??**). As expected, the relative fixation rate of nonsynonymous mutations (meausured using dN/dS) did not vary systematically across simulations (Table **??**).

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 **??**). 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

Recombination rates can be estimated in various ways, which have different pros and cons. For instance, the population-scaled recombination rate (ρ) can be inferred from a relatively small sample of unrelated individuals at very fine-scales using patterns of linkage disequilibrium (LD) (REVIEW?). However, selection at linked sites influences local LD and may therefore affect recombination rate estimates obtained in this way (REF?). Alternatively, direct estimates of the recombination rate (r) can be obtained from crossing experiments, but to achieve sufficient power to generate recombination



.

Figure 4.2: Nucleotide diversity in regions surrounding protein-coding exons and conserved non-coding elements in wild mice. Population-scaled genetic distances($4N_er$) were calculated using either an LD-based recombination map constructed for *M. m. castaneus* or the pedigree based *M. musculus* genetic map constructed by Cox et al. (2009). Gene conversion was included assuming either the gene conversion rates

maps a very large number of individuals need to be genotyped, which has typically precluded the use of whole-genome re-sequencing, limiting resolution. In summary, high resolution recombination maps can be generated using patterns of LD, but these may be biased by selection at linked sites, while unbiased recombination maps may be generated using crosses, though these typically have low resolution. When analysing patterns of genetic diversity using a model of selection at linked sites, the way in which recombination rate estimates were obtained may, therefore, affect parameter estimates.

In this study, we analysed patterns of genetic diversity in *M. m. castaneus* and calculated genetic distances assuming either the high resolution recombination map constructed from LD by Booker et al. (2017b)(the *castaneus* map) or the pedigree-based map of Cox et al. (2009) (the Cox map). The choice of recombination map had a substantial effect on patterns of nucleotide diversity. We found that, in the

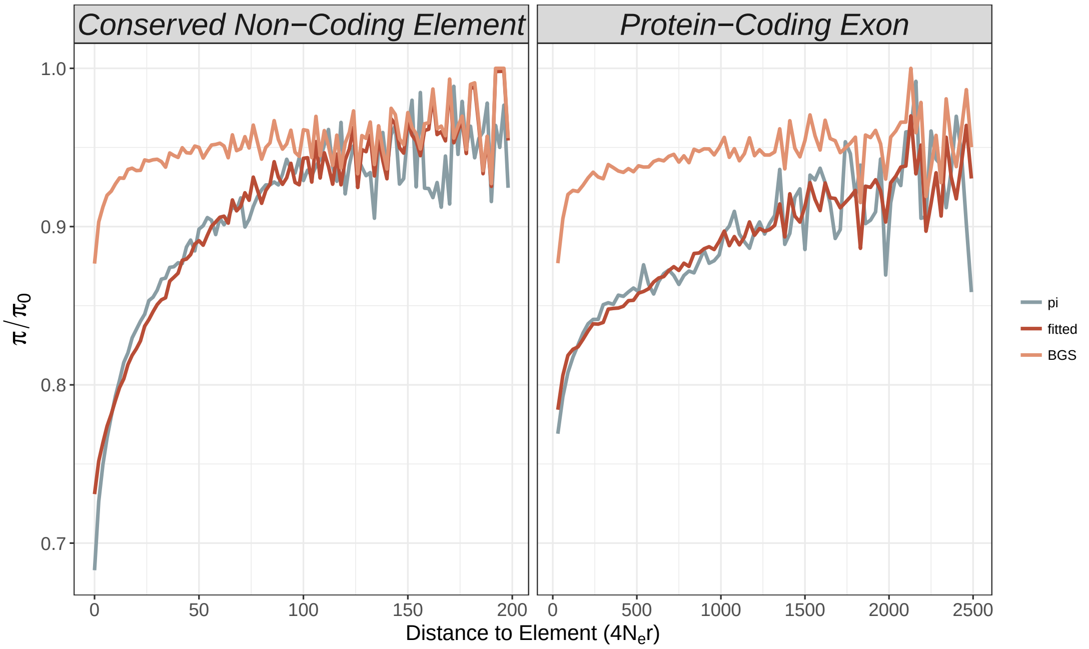


Figure 4.3: The pattern of scaled nucleotide diversity around protein-coding exons and CNEs in *M. m. castaneus*.

immediate flanks of both exons and CNEs, diversity was lower when assuming the LD-based *castaneus* map than when assuming the pedigree-based Cox map (Figure X). This difference is consistent with the idea that regions of the genome close to functional elements, where the effects of BGS and/or SSWs are strongest, and which exhibit reduced diversity, may yield downwardly biased estimates of the recombination rate. An alternative explanation is that the Cox map, which lacks resolution, does not fully capture regions of low recombination rate, so analysis windows that are tightly linked to functional elements may appear less tightly linked. An additional caveat is that in order to scale recombination rate estimates in the Cox map to ρ values, we assumed a single N_e for the entire genome, though N_e may very well vary across the genome. However, genetic diversity plateaus at a higher level when assuming the *castaneus* map, suggesting that the Cox map may not capture some of the highly recombining portions of the genome. The choice of recombination rate estimates will, therefore, have an impact on the parameters of selection inferred from the patterns of diversity. Throughout the rest of the paper, we present, in parallel, the results of analyses based on the *castaneus* map with those based on the Cox map.

Diversity expected in the absence of selection, π_0

A key parameter in Equation 4.1 is π_0, the nucleotide diversity expected in the absence of the effects of selection at linked sites. This parameter is very difficult to estimate and may even prove unobservable in real data given the ubiquity of the effects of selection at linked sites (Kern and Hahn, 2018). 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 **??**. 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

Parameters of selective sweep obtained from patterns of nucleotide diversity

By fitting Equation 4.1 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

What was the effect of including background selection or not?

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, but found that a single class of effects gave the best fit. Using AIC, 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.

Using Equation 4.3 we incorporated gene conversion into the analysis. However, given estimates of gene conversion parameters in mice, that it did not substnatially influence the analysis. We assumed gene conversion parameters estimated by Paigen et al. (2008), but incorporating these did not influence the selection parameters. In that study, the ratio of non-crossover gene conversion to crossing-over (NC/CR) was estimated to be 0.105, while gene conversion tracts measuring 9-279bp were detected. When assuming the center of this range (144bp) as the mean tract length and a NC/CR ratio of 0.105, gene conversion did not affect the selection parameter estimates. However, the gene conversion parameters estiamted by Paigen et al. (2008) were based on a small number of observations and the true parameters may be quite different. I This is not particularly surprising since the physical distances analysed are far greater than the mean tract length assumed.

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

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.

4.0.1 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. Consistent with Tataru et al. (2017), we found that if advantageous mutations are present, but unaccounted for, estimates of the dDFE become inaccurate.

SOMETHING ABOUT SOFT SWEEPS AND OTHER MODES OF ADAPTATION

In collating the patterns of genetic divrsity around either CNEs or protein-coding exons across the entire genome, it is likely that we have lost some valuable information. An alternative approach would be to fit Equation 4.1 to genome-wide variation in nucleotide diversity, conditioning on the locations of functional elements and a genetic map, which is, in effect what the methods of Elyashiv et al. (2016) do. In their study,Elyashiv et al. (2016) fitted genome-wide variation in *D. melanogaster* using a model that combined background selection and selective sweeps conditioning on the locations of recent substitutions to estimate the effects of selective sweeps and functional elements to ascertain the effects of background selection. Impressive though their methods are, their model does not make use of the information present in the SFS which can be used to accurately estimate dDFE parameters, even when positive selection is present. Elyashiv et al. (2016) found that their best fitting models overestimated the deleterious mutation rate which they attributed to the effects.

The model of selective sweeps that we used in this study is of so-called 'hard' (or classic) selective sweeps, whereas studies in both humans and *Drosophila* suggest that 'soft' selective sweeps are common (Garud and Petrov, 2016; Garud et al., 2015; Schrider and Kern, 2017). A 'soft' selective sweep differs from the model outlined in the Methods section of this paper in that multiple haplotypes reach fixation.

As Elyashiv et al. (2016) pointed out, if selective sweeps arising from standing genetic variation were common, then it is likely that we would overestimated the strength of selection.

Another relevant model is that of partial selective sweeps. Under this model advantageous mutations become effectively neutral at some point in their sojourn towards fixation. Partial sweeps may occur when a complex trait, which controlled by many loci of small effect may shift in allele frequency in response to environmental change Elyashiv et al. (2016) also discussed how partial sweeps (where sweeping allele become neutral in their sojourn to fixation)

In this study, gene conversion made little to no difference to parameter estimation, but this depends on the gene conversion parameters assumed. We assumed the estiamtes obtained by Paigen et al. (2008) when performing our analyses, which yielded little diffenrece in the parameter estimates.

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, 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, we can use the parameter estimates we obtained in this study to try and shed light on this question.

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:

$$\Delta W \propto \mu p_a n_a E(s_a)^2, \tag{4.5}$$

We parametrized Equation 4.5 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 4.5.

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), 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 a single discrete class of advnatgeous mutations gave the better fit to the data (TABLE REF), we do not suppose that the DFE for advantageous mutations is, in reality so simple. Secondly, we have assumed that all elements of a particular class share a common set of selection parameters. 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 different selective pressures). Indeed, sub-categorisations of protein-coding genes may also be subject to different selection pressures. For instance, immune related genes have evolved faster in mice than house-keeping genes and may 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). Since broad-scale recombination rates are strongly correlated between mice and rats (Jensen-Seaman et al., 2004), qualitatively similar conclusions regarding the contribution of protein-coding versus regulatory change to adaptive evolution may be reached when analysing patterns of genetic diversity in rats.

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?), giving us a certain confidence in the estimated effects of background selection. We used such estimates when fitting the sweep model to troughs in diversity around protein-coding exons and conserved non-coding elements. The parameter estimates we obtained suggest that positive selection is, on average, stronger in protein-coding regions of the genome than in regulatory regions, but that the influx of advantageous mutations into mouse populations is far larger for regulatory regions. Despite this, a model of the rate of change in fitness due to new advantageous mutations, suggests that protein change may contribute more than regulatory change.

Chapter 5

General Discussion

5.1 Overview

The main aim of this thesis has been to increase our understanding of the factors that have shaped nucleotide diversity across the mammalian genome. To that end, I have focussed on understanding how processes of selection at linked sites contribute to variation in genetic diversity across the genome. Each of the projects described in the preceding three chapters have touched upon a different aspect of this. Here, I briefly summarise the main points, as regards selection at linked sites, that can be gleaned from the three projects I presented.

In Chapter 2, I used a coalescent-based method to infer a recombination rate map from patterns of LD for *M. m. castaneus*. I used this map to analyse the relationship between synonymous site diversity for protein-coding genes and local recombination rate. I found that putatively neutral diversity and recombination rate were positively correlated. Since both background selection and selective sweeps are expected to cause reductions in diversity, which are positively correlated to the rate of recombination, the relationship I found is indicative of the widespread effects of selection at linked sites. However, the positive correlation does not, on its own, carry information on the contributions of background selection and selective sweeps. The recombination map that I generated was important for the analyses performed in Chapters 3 and 4.

In Chapter 3, I estimated the DFE for both harmful and advantageous mutations by analysing the unfolded site frequency spectrum (uSFS). Given the parameters obtained, I found that a combination of BGS and SSWs, could not fully explain the dips in diversity observed around functional elements in *M. m. castaneus.* Using simulations, I found circumstantial evidence that selective sweeps, driven by strongly advantageous mutations, are a major contributor to the dips in putatively neutral diversity around functional elements in mice.

In Chapter 4, I used a model that combined the effects of background selection and selective sweeps to estimate parameters of positively selected mutations that could generate the troughs in diversity observed around protein-coding exons and conserved non-coding elements (CNEs). My analyses suggested that strongly selected advantageous mutations in protein-coding exons are less frequent, but have substantially larger selection coefficients than those that occur in CNEs. Using the parameters, I found that the contribution to fitness change brought about by the substitution of advantageous mutations in protein-coding regions may somewhat outweigh that of advantageous mutations in regulatory elements.

The main finding from thework I have carried out is that strongly advantageous mutations are chiefly responsible for the large troughs in diversity, observed by Halligan et al. (2013), around both protein-coding exons and CNEs. In Chapter 3, I showed that BGS makes a contribution to the observed troughs, but in Chapter 4 I showed that strong SSWs are required to fully explain them. I also presented evidence that, for at least two classes of functional sites in mice, there is a multimodal distribution of advantageous mutational effects.

5.2 Limitations

There are a number of useful insights to be gleaned from the three projects described in this thesis, but the work I have done does not close the book on our understanding of selection in the house mouse genome. There are a number of assumptions that permeate this thesis, and here I give a brief description of these and the limitations they impose on the work carried out.

5.2.1 Neutrally evolving sequences

Probably the most obvious assumption I made is that that neutrality can be ascribed to certain classes of sites and regions of the genome. I made this assumption when analysing recombination rates in Chapter 2, the method I used (LDhelmet) invokes neutrality to model recombination rate variation across a chromosomal region (Chan et al., 2012). I also made this assumption in Chapter 3, where I assumed that the uSFSs for 4-fold and CNE-flanking sites represent neutrally evolving sequences. Although the sites may not themselves be the targets of selection, they may be influenced by selection at linked sites. There is ample evidence for selection acting on several classes of functional elements in the house mouse genome (Kousathanas et al., 2011; Halligan et al., 2011) and for some classes of sites, this affects genetic diversity in surrounding genomic regions (Halligan et al. 2013; Chapter 3). This will have influenced my efforts to estimate recombination rates and to estimate DFEs in Chapters 2 and 3, respectively.

Recombination rate issues

The method used to infer recombination rate variation in Chapter 2, LDhelmet (Chan et al., 2012), is built upon coalescent theory, which makes the assumption that

SNPs in a focal region are evolving neutrally. For a pair of SNPs, the likelihood of an observed genealogy under a particular recombination rate (in terms of $4N_er$) is calculated. By estimating the recombination rate between pair-wise combinations of SNPs, one can then build up a recombination map for a particular genomic region. Both population size change and selection at linked sites can influence linkage disequilibrium (REFS), distort genealogies away from neutral expectation (Barton, 1998) and reduce the total number of segregating variants in a region (Braverman et al., 1995). Both processes could, therefore, potentially bias and affect and resolution of recombination rate maps obtained using methods such as LDhelmet.

Firstly, methods such as LDhelmet rely on the presence of SNPs, so a low number of variants in a region limits the resolution to which recombination rates can be inferred. The effects of population size change have recently been incorporated into the methodology (Kamm et al., 2016), but this relies on accurate estimates of the demographic history of the population, which is not necessarily straightforward (*see below*).

Low recombination rates estimated from LD may be a caused by the effects of selection at linked sites, the true recombination rate being low, or both. If the true recombination rate were low and selection at linked sites were operating, then estimates of the recombination could potentially become severely biased. We showed in Chapter 2 that my LD-based estimates of the recombination rate are similar to pedigree-based estimates at the megabase scale, but without high resolution pedigree-based estimates of the recombination rate for *M. m. castaneus* it is difficult to assess the extent of this bias at finer scales.

Finally, there is a degree of circularity in using LD-based estimates of the recombination rate for analysing selection at linked sites. For instance, in Chapter 3, we simulated positive and negative selection and set recombination rates using the LD-based estimates of ρ obtained in Chapter 2. A recent selective sweep in *M. m. castaneus* could result in elevated LD in a region, and thus downwardly biased of $\rho = 4N_er$ in the recombination map. When simulating such regions the downwardly biased estimate of ρ may have exacerbated the signal of selection at linked sites. This circularity has the effect of making the analysis in Chapters 3 conservative, but may biased the selection parameters obtained in Chapter 4.

Inferring the DFE

Estimates of the DFE for mutations that affect fitness can be obtained by contrasting the distribution of allele frequencies in a class of sites assumed to be subject to selection with that of a putatively neutrally evolving comparator. The uSFS analysis methods used in Chapters 3 and 4, DFE-alpha and polyDFE respectively, assume that the neutrally evolving reference class is interspersed among the selected sites of interest. A good example of this interspersion are synonymous and nonsynyonmous sites of protein-coding genes. However, in Chapter 3, when estimating DFEs For CNEs and UTRs, we used neutral comparators that are tightly linked, but not interspersed among the selected site class. Kousathanas and Keightley (2013) showed that linked putatively neutral sites can be used as a neutral reference for inferring the DFE using simulations. However, in the case of the CNEs identified by Halligan et al. (2013), there is evidence for the presence of functionally constrained sequences in the flanks of inferred elements. One potential consequence of selected mutations segregating in these functionally constrained sequences would be underestimation of the strength of selection by analysis of the SFS.

We assumed that 4-fold sites and CNE flanking sites evolve neutrally in order to estimate DFEs from linked sites. While there is little evidence of codon-usage bias in mice, one potential source of selection on synonymous sites (dos Reis and Wernisch, 2009), there is evidence that synonymous sites within splice enhancers are conserved and are thus potentially subject to purifying selection (Savisaar and Hurst 2016).

5.2.2 Categorisation of functional elements

Throughout Chapters 3 and 4, we have assumed that all CNEs share a single DFE. This is a reasonable starting point, but may be problematic. The CNEs analysed in this thesis were identified by Halligan et al. (2013) using a alignment-based approach called phastCons. phastCons identifies conserved elements using an alignment of genomes, identifying individual elements using a phylogenetic hidden Markov model. The model emits discrete intervals of the genome, which are inferred to be functional on the basis of sequence conservation. In vertebrates, CNEs identified by phastCons appear to have arisen during three main time periods, apparently corresponding to the evolution of biological innovations (Lowe et al., 2011). For example, CNEs associated with genes involved in mouse coat development appear to have arisen at a similar time to the ancestor of amniotes (Lowe et al., 2011).

CNEs identified using phastCons (Lowe et al., 2011) may play various biochemical roles, for example insulation and repression, and it seems reasonable to expect that they may be subject to different selection pressures. If this were the case, we may have incurred bias when estimating selection coefficients by treating all CNEs a homogeneous group. Dividing the set of identified CNEs up by biochemical function would perhaps be difficult. However, with data of the kind generated by the ENCODE project, one could attempt to partition CNEs into different biochemical categories and then determine whether these categories experience different selection regimes. Ascribing biological function to phastCons elements is not necessarily straightforward, however, since the parameters used to tune the model (the expected length and coverage of a conserved elements) can influence the length and number of functional elements identified (Siepel et al., 2005). Of course, any analysis that would sub-categorise the set of CNEs would have to balance the number of categories analysed with statistical power.

5.2.3 Soft selective sweeps in mice

The work in this thesis has relied on an assumption of hard selective sweeps. Both the DFE-alpha and polyDFE analyses in Chapters 3 and 4 assume that the strength of selection acting on advantageous or deleterious mutations does not change through time. However, in a rapidly changing environment, alleles that were once neutral could become advantageous or disadvantageous in a new context. Alleles that become advantageous in this manner and subsequently fix generate soft selective sweeps (Hermisson and Pennings, 2017). If soft sweeps were common, the assumption of hard sweep may have influenced the outcomes of the analyses. Soft sweeps can also occur due to multiple copies of an advantageous allele arising in a population. Throughout the course of my PhD, there has been a lively debate in the literature as to the relative importance of the soft and hard sweep models (Jensen, 2014; Hermisson and Pennings, 2017).

If a soft sweep arises due to selection acting on standing variation, properties of the reductions in diversity are distinct from those of hard sweeps (Hermisson and Pennings, 2017). The reason for this is that as multiple haplotypes go to fixation, more of the neutral polymorphism present before the onset of selection is preserved, causing the trough in diversity around the selected site to be shallower than for a hard sweep. If soft sweeps were the dominant mode of adaptation in mice, then the selection parameters obtained in Chapter 4 would likely be underestimates of the true values.

A crucial parameter for the probability of whether selection from standing variation results in a soft sweep or not is the frequency at which the sweeping allele was present in the population at the onset of selection, x_0. In the case of 0-fold nonsynonymous sites in mice, for example, most standing variation is at low frequencies (Chapter 3). It has been argued that at the onset of selection x_0 will be close to $\frac{1}{2N}$, so when the allele becomes advantageous it will be present on only a very small number of haplotypes, which would likely lead to a hard sweep (Jensen, 2014). In contrast, Hermisson and Pennings (2017) showed that if the distribution of x_0 (i.e. the site frequency spectrum) is taken into account the probability of a soft sweep from standing variation is far higher than when assuming a single fixed value. The reason for this is that the adaptation acting on higher frequency standing variants are far more likely to cause soft sweeps. Since site frequency spectra for different classes of sites differ, the probabilities of soft sweeps occurring in different classes of functional elements may differ.

Schrider and Kern (2017) used a machine learning approach to classify regions of the human genome as either having experienced or having been linked to a hard or soft selective sweep. They used a method, S/HIC, which uses ‘random forest’ classification methods (Schrider and Kern, 2016). The basic protocol is as follows: Data are simulated under a specific model, e.g. neutrality, soft sweeps or hard sweeps, and summary statistics are calculated. Many simulations are performed varying a range of parameters, such as recombination rate and strength of selection. From a randomly sampled subset of these simulations, summary statistics are extracted and used to generate a decision tree that, when presented with a new set of summary statistics, classifies the input dataset as having been generated under a particular model (in this case neutral, hard or soft sweep or linked to a hard or soft sweep). This process is repeated with many randomly generated trees (populating the ‘random forest’). The model that obtains the most classifications (or ‘votes’) from the large set of random trees provides insight into the underlying process that generated the data.

Schrider and Kern (2017) applied S/HIC to human polymorphism data from the 1000 Genomes project (The 1000 Genomes Project Consortium et al., 2015) and found that soft selective sweeps are far more common than hard sweeps. One of the useful properties of the ‘random forest‘ methods employed by Schrider and Kern (2017) is that they allow one to rank the most influential summary statistics. In the paper describing S/HIC, Schrider and Kern (2016) analysed human chromosome 18 and reported that H2/H1, a statistic that summarises the distribution of haplotype frequencies in a sample

(Garud et al., 2015), was highly influential for their classifications. This is of note, because allelic gene conversion, which was not included in the simulations they used to build their machine learning classifiers, can cause the haplotype distribution under hard sweeps to resemble that of soft sweeps (Schrider et al., 2015). It remains to be seen whether the findings of Schrider and Kern (2017) are robust to the effects of gene conversion.

If soft sweeps are the dominant mode of adaptation in humans as the analyses of Schrider and Kern (2017) suggest, then it seems likely that it would also be so for mice. On the basis of within-species polymorphism, humans are estimated to have a $\theta = 4N_e\mu$ of around 0.001 (Yu et al., 2002). The probability of soft selective sweeps occurring from either standing genetic variation or recurrent mutations are proportional to θ (Hermisson and Pennings, 2017), so wild mice, which have an estimated θ of around 1%, should thus be more predisposed to soft sweeps than humans. If there were evidence for frequent soft sweeps in mice, which was robust to the effects of gene conversion, then the analysis methods used throughout this thesis would have to be scrutinised under a model of adaptation from standing variation or from multiple mutations.

5.2.4 Dominance and Haldane's sieve

In 1927 Haldane demonstrated that newly introduced recessive beneficial mutations are far more likely to be lost by chance than dominant mutations with the same selective advantage (Haldane 1927). This effect, which has become to be known as Haldane's sieve, thus predicts that most beneficial mutations that become fixed are dominant (assuming that an equal number of dominant and recessive mutations arise). If adaptation proceeds from standing variation, Haldane's sieve may not be relevant, because recessive and dominant alleles have similar probabilities of fixation when they are at high frequencies (Orr and Betancourt 2001). My analyses of the uSFS and of patterns of genetic diversity in Chapters 3 and 4 relied on models which assume that all new mutations are additive (or semi-dominant) in their effects. In the case of the analyses in Chapter 4, 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* (the dominance coefficient h and a selection coefficient s) are similar (Ewing et al. 2011). Because of this, if new mutations are neither fully recessive nor fully dominant, the selection coefficients estimated from the patterns of diversity they leave behind should be directly proportional to the true values. It is somewhat unclear, however, how varying dominance coefficients would influence SFS-based analyses.

5.2.5 The interaction between natural selection and demographic history

This thesis has focussed on the effects of background selection and selective sweeps, and has assumed, except where explicitly modelled, that the demographic history of *M. m. castaneus* has not influenced the analyses. This may potentially bias the results presented in Chapters 3 and 4, because the effects of BGS can become amplified under population size change (Torres et al. 2018). In Chapter 3, we inferred that *M. m. castaneus* has recently undergone a dramatic population expansion, a result obtained from two quasi-independent classes of putatively neutral sites (4-fold degenerate synonymous sites and CNE-flanking sites). It is tempting to interpret these results in light of recent human history: Mice are commensal to humans so their population numbers have likely exploded in the recent past. However, as we also showed in Chapter 3, selection at linked sites can cause one to infer a population expansion even there is not one. *M. m. castaneus* may have undergone a rapid population expansion in the recent past, but it is likely that the demographic parameters we inferred are highly influenced by selection at linked sites.

Across the *M. m. castaneus* genome, there is a strongly negative Tajima's D of around -0.5, consistent with both widespread selection at linked sites and a recent population expansion. In Chapter 3, we showed that selection at linked sites (as generated by the DFEs we inferred from the mouse population data we analysed) does not result Tajima's D values as negative as those observed. Even when I modelled relatively strong selection ($\gamma_a = 400$), SSWs resulted in a localised trough in Tajima's D around protein-coding exons, but this recovered almost to 0 in surrounding regions (Figure C.8). This suggests that the seemingly genome-wide negative Tajima's D is not solely explained by the effects of selection at linked sites, but of course there may be a combination of selection parameters that generate the observed values. Additionally, it is possible that relatively recent demographic processes have erased the signal of selection at linked sites across the genome. To fully investigate this possibility, however, estimates of the demographic history for mice, unbiased by the effects of selection at linked sites are required. There are number of strategies that could be employed to obtain these.

5.3 Moving Forward

There are many possible directions that could be taken with to further our understanding of the factors that shape patterns of genetic diversity across the mouse genome. As the final part of this thesis, I will describe several possible areas for further study.

5.3.1 Robust demographic models

The majority of demographic models assume neutrally evolving sites, so it is desirable to parametrise them from regions of the genome that are free from the effects of selection at linked sites. One could use regions of the genome far from functional elements (both coding and non-coding), because these are expected to be the most free from the effects of selection at linked sites, especially if they are in highly recombining regions. One could go a step further and fit a model of selection at linked sites to genome-wide polymorphism data (e.g. Elyashiv et al. 2016) and identify regions that only experience small effects of selection. However, such methods rely on a perfect knowledge of the locations of functional elements in the genome. Since the methods used to identify conserved elements may fail to detect rapidly evolving sequences and weakly selected regions, there is the possibility that BGS and SSW influence genetic variation even when there are no annotations present. An alternative strategy would be to use methods such as the machine learning method S/HIC, which uses a machine learning classifier that can discriminate between neutrally evolving sequences and sequences influenced by selection at linked sites by 'learning' the properties of neutrally evolving sequences on the basis of summary statistics. Parametrising demographic models from the neutrally evolving sequences identified using the above described methods may give estimates of the demographic history that are least affected by selection at linked sites.

5.3.2 Making better use of the available data

A substantial hurdle to population genomic research is in making use of all the available data. For example, in Chapters 3 and 4 I have analysed either the site frequency spectrum or nucleotide diversity. These are just two data summaries that can be conveniently analysed in population genetic models, but there are others. As I demonstrated in Chapter 4, the SFS is a useful summary of the data that can be used to estimate the distribution of fitness effects for harmful variants, but the uSFS can be uninformative for estimating the parameters of strongly selected advantageous mutations, particularly if they are rare. In such cases, patterns of genetic diversity are more informative. Ideally, one would make use of information in both the uSFS for potentially selected sites, whilst simultaneously modelling the reductions in neutral diversity caused by selection at those sites. There is information about the effects of selection in both linkage disequilibrium and population haplotype structure, but these may be very difficult to incorporate into an analytical expression along with the SFS and diversity at linked sites. One possibility for using as much of the available data as possible would be to perform approximate Bayesian computation (ABC) or machine learning with forward-in-time population genetic simulations.

The basic idea is as follows: Simulate data under a model, sampling the parameters of interest from plausible ranges, and compare summary statistics from your dataset to those obtained by simulation (Beaumont et al., 2002). The parameter sets that generated summary statistics most resembling those in the data give an estimate of the underlying parameters. This is the basic idea behind ABC and machine learning approaches so far developed in population genomics, although ML methods have the benefit of allowing the user to assess the importance of particular parameters (Schrider and Kern, 2018). In the context of inferring the dDFE and positive selection parameters, one could simulate a chromosome or chromosomal regions with the same structure as the species of interest (like I did in Chapter 3). Many thousands of different combinations of DFE parameters could be simulated, and from these data, one extract summary statistics for the site frequency spectrum, linkage disequilibrium and haplotype structure within and in the regions surrounding, several classes of functional elements. The biggest barrier to applying an analysis such as this is the computational demands of the many simulations required.

The simulations used in this thesis were performed with SLiM v1.8, a program which was, at the time of its release, among the most computationally efficient forward-in-time simulators available (Messer, 2013). Forward simulators have historically been much slower than coalescent simulators because the evolution of whole chromosomes is typically tracked. In the original SLiM publication, Messer (2013) described how by tracking just the mutations, simulations of purifying selection acting on a whole human chromosome (100Mbp long; 10^4 diploid individuals; for 10^5 generations) took just four days. As impressive as that is, it is not feasible to perform ABC or ML using such simulations. In the four years since starting my PhD a number of increasingly efficient forward-in-time simulators have been developed (Hernandez and Uricchio, 2015; Haller and Messer, 2017; Thornton, 2014), but even with these it would be difficult to perform ABC or ML as described. However, very recent advances in computational efficiency of forward-in-time simulators (Kelleher et al., 2018) may bring approaches of the kind outlined above within reach.

5.3.3 Of mice and men (and fruit flies)

The vast majority of the literature cited in this thesis has involved mice, humans or *D. melanogaster*. Being three of the most well studied organisms in biology means that the genomic resources available for these three species are excellent. However, it is not obvious how generalisable findings in these groups are. There are substantial barriers to performing population genomic studies in non-model organisms, for example in obtaining a reference genome and in obtaining functional annotations of the genome (Ellegren, 2014). However, analysing close relatives of model organisms is a way to give evolutionary studies a broader focus. A number of studies in great apes provide a template for comparisons between closely related species. For instance, across great ape species Stevison et al. (2015) showed that there is a negative correlation between recombination rate similarity and nucleotide divergence and Nam et al. (2017) concluded that strong sweeps are largely responsible for reductions in diversity near genes. In both these studies, reference genomes and annotations obtained for the model organisms were used when analysing sister species. Similar comparative studies could be performed in murid rodents, using available datasets; population samples of the three principle *M. musculus* sub-species and *Mus spretus* (Harr et al., 2016) and the brown rat (Deinum et al., 2015) are publicly available.

One of the intriguing findings from Chapter 4 was that adaptation in protein-coding regions may be driven by mutations that are, on average, far stronger than those that occur in regulatory regions. Performing analyses using the rodent datasets described above could further our understanding of whether adaptation in protein-coding regions contributing more to phenotypic change than regulatory regions is a general feature of mammalian evolution or specific to *M. m. castaneus*.

Bibliography

Aguade, M., Miyashita, N., and Langley, C. H. (1989). Reduced variation in the yello-achaete-schute region in natural populations of drosophila melanogaster. *Genetics*, 122:607–615.

Andolfatto, P. (2007). Hitchhiking effects of recurrent beneficial amino acid substitutions in the drosophila melanogaster genome. *Genome Res*, 17(12):1755–62.

Baines, J. F. and Harr, B. (2007). Reduced x-linked diversity in derived populations of house mice. *Genetics*, 175(4):1911–1921.

Bank, C., Hietpas, R. T., Wong, A., Bolon, D. N., and Jensen, J. D. (2014). A bayesian mcmc approach to assess the complete distribution of fitness effects of new mutations: uncovering the potential for adaptive walks in challenging environments. *Genetics*, 196(3):841–52.

Barton, N. H. (1998). The effect of hitch-hiking on neutral genealogies. *Genetical Research*, 72:123–133.

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

Beaumont, M. A., Yang, W., and Balding, D. J. (2002). Approximate bayesian computation in populaiton genetics. *Genetics*, 162:2025–2035.

Begun, D. J. and Aquadro, C. F. (1992). Levels of naturally occuring dna polymorphism correlate with recombination rate in drosophila melanogaster. *Nature*, 356.

Beissinger, T. M., Wang, L., Crosby, K., Durvasula, A., Hufford, M. B., and Ross-Ibarra, J. (2016). Recent demography drives changes in linked selection across the maize genome. *Nature Plants*, 2(7):16084.

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

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

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

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

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., and Petukhova, G. V. (2012). Genetic recombination is directed away from functional genomic elements in mice. *Nature*, 485(7400):642–5.

Brunschwig, H., Liat, L., Ben-David, E., Williams, R. W., Yakir, B., and Shifman, S. (2012). Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics*, 191:757–764.

Campos, J. L., Zhao, L., and Charlesworth, B. (2017). Estimating the parameters of background selection and selective sweeps in drosophila in the presence of gene conversion. *Proc Natl Acad Sci*, Early Online.

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

Chan, A. H., Jenkins, P. A., and Song, Y. S. (2012). Genome-wide fine-scale recombination rate variation in drosophila melanogaster. *PLoS Genet*, 8(12):e1003090.

Charlesworth, B. (1994). The effect of background selection against deleterious mutations on weakly selected, linked variants. *Genetical Research*, 63(03):213.

Charlesworth, B. (1996). Background selection and patterns of genetic diversity in drosophila melanogaster. *Genetical Research*, 68:131–149.

Charlesworth, B. (2009). Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet*, 10(3):195–205.

Charlesworth, B. (2012). The role of background selection in shaping patterns of molecular evolution and variation: evidence from variability on the drosophila x chromosome. *Genetics*, 191(1):233–46.

Charlesworth, B. (2013). Background selection 20 years on: the wilhelmine e. key 2012 invitational lecture. *J Hered*, 104(2):161–71.

Charlesworth, B., Morgan, M. T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134:1289–1303.

Charlesworth, D. (2006). Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet*, 2(4):e64.

Charlesworth, D., Charlesworth, B., and Morgan, M. T. (1995a). The pattern of neutral molecular variation under the background selection model. *Genetics*, 141.

Charlesworth, D., Charlesworth, B., and Morgan, M. T. (1995b). The pattern of neutral molecular variation under the background selection model. *Genetics*, 141:1619–1632.

Chia, J. M., Song, C., Bradbury, P. J., Costich, D., de Leon, N., Doebley, J., Elshire, R. J., Gaut, B., Geller, L., Glaubitz, J. C., Gore, M., Guill, K. E., Holland, J., Hufford, M. B., Lai, J., Li, M., Liu, X., Lu, Y., McCombie, R., Nelson, R., Poland, J., Prasanna, B. M., Pyhajarvi, T., Rong, T., Sekhon, R. S., Sun, Q., Tenaillon, M. I., Tian, F., Wang, J., Xu, X., Zhang, Z., Kaeppler, S. M., Ross-Ibarra, J., McMullen, M. D., Buckler, E. S., Zhang, G., Xu, Y., and Ware, D. (2012). Maize hapmap2 identifies extant variation from a genome in flux. *Nat Genet*, 44(7):803–7.

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.

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

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

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

dos Reis, M. and Wernisch, L. (2009). Estimating translational selection in eukaryotic genomes. *Mol Biol Evol*, 26(2):451–61.

Ellegren, H. (2014). Genome sequencing and population genomics in non-model organisms. *Trends Ecol Evol*, 29(1):51–63.

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.

Enard, D., Messer, P. W., and Petrov, D. A. (2014). Genome-wide signals of positive selection in human evolution. *Genome Res*, 24(6):885–95.

Ewing, G. B. and Jensen, J. D. (2016). The consequences of not accounting for background selection in demographic inference. *Mol Ecol*, 25(1):135–41.

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

Eyre-Walker, A. and Keightley, P. D. (2009). Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *Mol Biol Evol*, 26(9):2097–108.

Eyre-Walker, A., Woolfit, M., and Phelps, T. (2006). The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics*, 173(2):891–900.

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

Franchini, L. F. and Pollard, K. S. (2017). Human evolution: the non-coding revolution. *BMC Biology*, 15(1).

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

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

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

Geraldes, A., Basset, P., Smith, K. L., and Nachman, M. W. (2011). Higher differentiation among subspecies of the house mouse (mus musculus) in genomic regions with low recombination. *Mol Ecol*, 20(22):4722–36.

Glemin, S., Arndt, P. F., Messer, P. W., Petrov, D., Galtier, N., and Duret, L. (2015). Quantification of gc-biased gene conversion in the human genome. *Genome Res*, 25(8):1215–28.

Haddrill, P. R., Zeng, K., and Charlesworth, B. (2011). Determinants of synonymous and nonsynonymous variability in three species of drosophila. *Mol Biol Evol*, 28(5):1731–43.

Haller, B. C. and Messer, P. W. (2017). Slim 2: Flexible, interactive forward genetic simulations. *Mol Biol Evol*, 34(1):230–240.

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.

Halligan, D. L., Oliver, F., Eyre-Walker, A., Harr, B., and Keightley, P. D. (2010). Evidence for pervasive adaptive protein evolution in wild mice. *PLoS Genet*, 6(1):e1000825.

Halligan, D. L., Oliver, F., Guthrie, J., Stemshorn, K. C., Harr, B., and Keightley, P. D. (2011). Positive and negative selection in murine ultraconserved noncoding elements. *Mol Biol Evol*, 28(9):2651–60.

Harr, B., Karakoc, E., Neme, R., Teschke, M., Pfeifle, C., Pezer, Ž., Babiker, H., Linnenbrink, M., Montero, I., Scavetta, R., Abai, M. R., Molins, M. P., Schlegel, M., Ulrich, R. G., Altmüller, J., Franitza, M., Büntge, A., Künzel, S., and Tautz, D. (2016). Genomic resources for wild populations of the house mouse, mus musculus and its close relative mus spretus. *Scientific Data*, 3:160075.

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

Hermisson, J. and Pennings, P. S. (2017). Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. *Methods in Ecology and Evolution*, 8(6):700–716.

Hernandez, R., Kelly, J. L., Elyashiv, E., Melton, S. C., Auton, A., McVean, G., Project, . G., Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. *Science*, 331:920–924.

Hernandez, R. D. and Uricchio, L. H. (2015). Sfscode: More efficient and flexible forward simulations. *bioRxiv*.

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

Jensen, J. D. (2014). On the unfounded enthusiasm for soft selective sweeps. *Nat Commun*, 5:5281.

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.

Josephs, E. B., Lee, Y. W., Stinchcombe, J. R., and Wright, S. I. (2015). Association mapping reveals the role of purifying selection in the maintenance of genomic variation in gene expression. *Proc Natl Acad Sci*, 112(50):15390–5.

Kamm, J. A., Spence, J. P., Chan, J., and Song, Y. S. (2016). Two-locus likelihoods under variable population size and fine-scale recombination rate estimation. *Genetics*, 203(3):1381–1399.

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

Keightley, P. D. and Eyre-Walker, A. (2007). Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics*, 177(4):2251–61.

Kelleher, J., Thornton, K., Ashander, J., and Ralph, P. (2018). Efficient pedigree recording for fast population genetics simulation. *bioRxiv*.

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

Kim, Y. (2006). Allele frequency distribution under recurrent selective sweeps. *Genetics*, 172(3):1967–78.

Kimura, M. (1983). *The Neutral Theory of Molecular Evolution*. Cambridge University Press.

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

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

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

Kousathanas, A., Oliver, F., Halligan, D. L., and Keightley, P. D. (2011). Positive and negative selection on noncoding dna close to protein-coding genes in wild house mice. *Mol Biol Evol*, 28(3):1183–91.

Lowe, C. B., Kellis, M., Siepel, A., Raney, B. J., Clamp, M., Salama, S. R., Kingsley, D. M., Lindblad-Toh, K., and Haussler, D. (2011). Three periods of regularory innovation during vertebrate evolution. *Science*, 333(6045):pp. 1019–1024.

Macpherson, J. M., Sella, G., Davis, J. C., and Petrov, D. A. (2007). Genomewide spatial correspondence between nonsynonymous divergence and neutral polymorphism reveals extensive adaptation in drosophila. *Genetics*, 177(4):2083–99.

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.

McDonald, M. J., Rice, D. P., and Desai, M. M. (2016). Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature*, 531(7593):233–6.

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.

Messer, P. W. and Petrov, D. A. (2013). Frequent adaptation and the mcdonald-kreitman test. *Proc Natl Acad Sci*, 110(21):8615–8620.

Nam, K., Munch, K., Mailund, T., Nater, A., Greminger, M. P., Krutzen, M., Marques-Bonet, T., and Schierup, M. H. (2017). Evidence that the rate of strong selective sweeps increases with population size in the great apes. *Proc Natl Acad Sci U S A*, 114(7):1613–1618.

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

Otto, S. P. and Whitlock, M. C. (1997). The probability of fixation in populations of changing size. *Genetics*, 146(723-733).

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

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

Pritchard, J. K., Pickrell, J. K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol*, 20(4):R208–15.

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

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

Schrider, D. R. and Kern, A. D. (2016). S/hic: Robust identification of soft and hard sweeps using machine learning. *PLoS Genet*, 12(3):e1005928.

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

Schrider, D. R. and Kern, A. D. (2018). Supervised machine learning for population genetics: A new paradigm. *Trends Genet*, 34(4):301–312.

Schrider, D. R., Mendes, F. K., Hahn, M. W., and Kern, A. D. (2015). Soft shoulders ahead: spurious signatures of soft and partial selective sweeps result from linked hard sweeps. *Genetics*, 200(1):267–84.

Schrider, R. D., Shanku, G. A., and Kern, D. A. (2016). Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, (Early Online Access).

Sella, G., Petrov, D. A., Przeworski, M., and Andolfatto, P. (2009). Pervasive natural selection in the drosophila genome? *PLoS Genet*, 19(6).

Siepel, A., Bejerano, G., Pedersen, J. S., Hinrichs, A. S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., Weinstock, G. M., Wilson, R. K., Gibbs, R. A., Kent, W. J., Miller, W., and Haussler, D. (2005). Evolutionarily conserved elements in vertebrate, insect, worm and yeast genomes. *Genome Res*, 15:1034–1050.

Smagulova, F., Brick, K., Yongmei, P., Camerini-Otero, R. D., and Petukhova, G. V. (2016). The evolutionary turnover of recombination hotspots contributes to speciation in mice. *Genes and Development*, 30:277–280.

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.

Stephan, W. and Langley, C. H. (1989). Molecular genetic variation in the centromeric region of the x chromosome in three drosohila ananassae populations. i. constrasts between the vermilion and forked loci. *Genetics*, 121:89–99.

Stevison, L. S., Woerner, A. E., Kidd, J. M., Kelley, J. L., Veeramah, K. R., McManus, K. F., Great Ape Genome, P., Bustamante, C. D., Hammer, M. F., and Wall, J. D. (2015). The time scale of recombination rate evolution in great apes. *Mol Biol Evol*.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by polymorphism. *Genetics*, 123:585–595.

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.

The 1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., and Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, 526(7571):68–74.

Thornton, K. R. (2014). A c++ template library for efficient forward-time population genetic simulation of large populations. *Genetics*, 198(1):157–66.

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.

Uricchio, L. H. and Hernandez, R. D. (2014). Robust forward simulations of recurrent hitchhiking. *Genetics*, 197(1):221–236.

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

Williamson, R. J., Josephs, E. B., Platts, A. E., Hazzouri, K. M., Haudry, A., Blanchette, M., and Wright, S. I. (2014). Evidence for widespread positive and negative selection in coding and conserved noncoding regions of capsella grandiflora. *PLoS Genetics*, 10(9):e1004622.

Appendices

Appendix A

Booker *et al.* 2017 - BMC Biology

Appendix B

Recombination in wild mice

Included here are the supplementary figures and tables for Chapter 2 as well as a reproduction of Booker et al. (2017b).

B.1 Supplementary Material

Table B.1: Summary of sex-averaged recombination rates *M. m castaneus* compared with the rates from Brunschwig et al. (2012) and Cox et al. (2009). Rates for the castaneus and Brunschwig maps are presented in terms of $4N_e r/bp$. Estimates of N_e were obtained by assuming the recombination rates from Cox et al. (2009).

						Switch Errors		
Filter Set	HWE	Min DP	Max DP	Min GQ	H40	H46	H62	
1	-	-	-	-	5,148/409,486	4,819/407,422	5,020/394,778	0.0124
2	¡0.0002	10	-	15	1,690/338,592	1,451/334,111	1,452/324,199	0.0046
3	¡0.0002	10	100	5	2,460 /341,744	2066 /339,508	2536 /328,998	0.0070
4	¡0.0002	-	-	40	523 /288,471	444 /286,636	550 /281,266	0.0018

Table B.2: The normalized mutation rate matrix and stationary distribution of base frequencies estimated with two out-groups, M. famulus and R. norvegicus, using the method described by Chan et al. (2012).

	A	C	G	T
A	0.48	0.09	0.32	0.11
C	0.19	0.00	0.12	0.69
G	0.69	0.12	0.00	0.19
T	0.11	0.32	0.08	0.48
Stationary Distribution	0.34	0.16	0.16	0.34

Table B.3: The total number of SNPs in the dataset, and the number of SNPs after applying filters.

		# SNPs	
Chromosome	Physical Size (Mbp)	Raw	Filtered
1	197.2	6,250,153	3,557581
2	181.7	5,420,000	3,095049
3	159.6	5,207,849	2,961039
4	155.6	4,916,193	2,655529
5	152.5	4,786,546	2,639326
6	149.5	4,831,712	2,658278
7	152.5	4,296,986	2,266748
8	131.7	4,089,400	2,309811
9	124.1	3,871,695	2,221982
10	130.0	4,323,747	2,440209
11	121.8	3,744,895	2,100852
12	121.3	3,674,871	2,036520
13	120.3	3,760,538	2,137776
14	125.2	3,874,312	2,164901
15	103.5	3,333,449	1,877022
16	98.3	3,193,551	1,822476
17	95.3	3,111,409	1,627303
18	90.8	2,926,381	1,692050
19	61.3	1,949,809	1,101783
X	166.7	2,535,365	1,469566
Total		80,098,861	44,835,801

Table B.4: The overlap between the hotspots we identified in M. m. castaneus and the locations of DSB hotspots in wild-derived strains obtained by Smagulova et al. (2016). The corrected overlap is the number of overlapping hotspots, above the null expectation, over the total.

Strain ID	Sub-species	# DSB Hotspots	# Overlaps	% Overlap Uncorrected	Null Expectation	% Overlap Corrected
13R	*domesticus*	14744	1202	8.2	1169	0.2
B6	*domesticus*	19455	1533	7.9	1505	0.1
C3H	*domesticus*	14635	1399	9.6	1308	0.6
CAST	*castaneus*	15061	1831	12.2	1221	4.1
MOL	*molossinus*	15718	1559	9.9	1351	1.3
PWD	*musculus*	14483	1569	10.8	1205	2.5

B.2 Booker *et al.* 2017 - Genetics

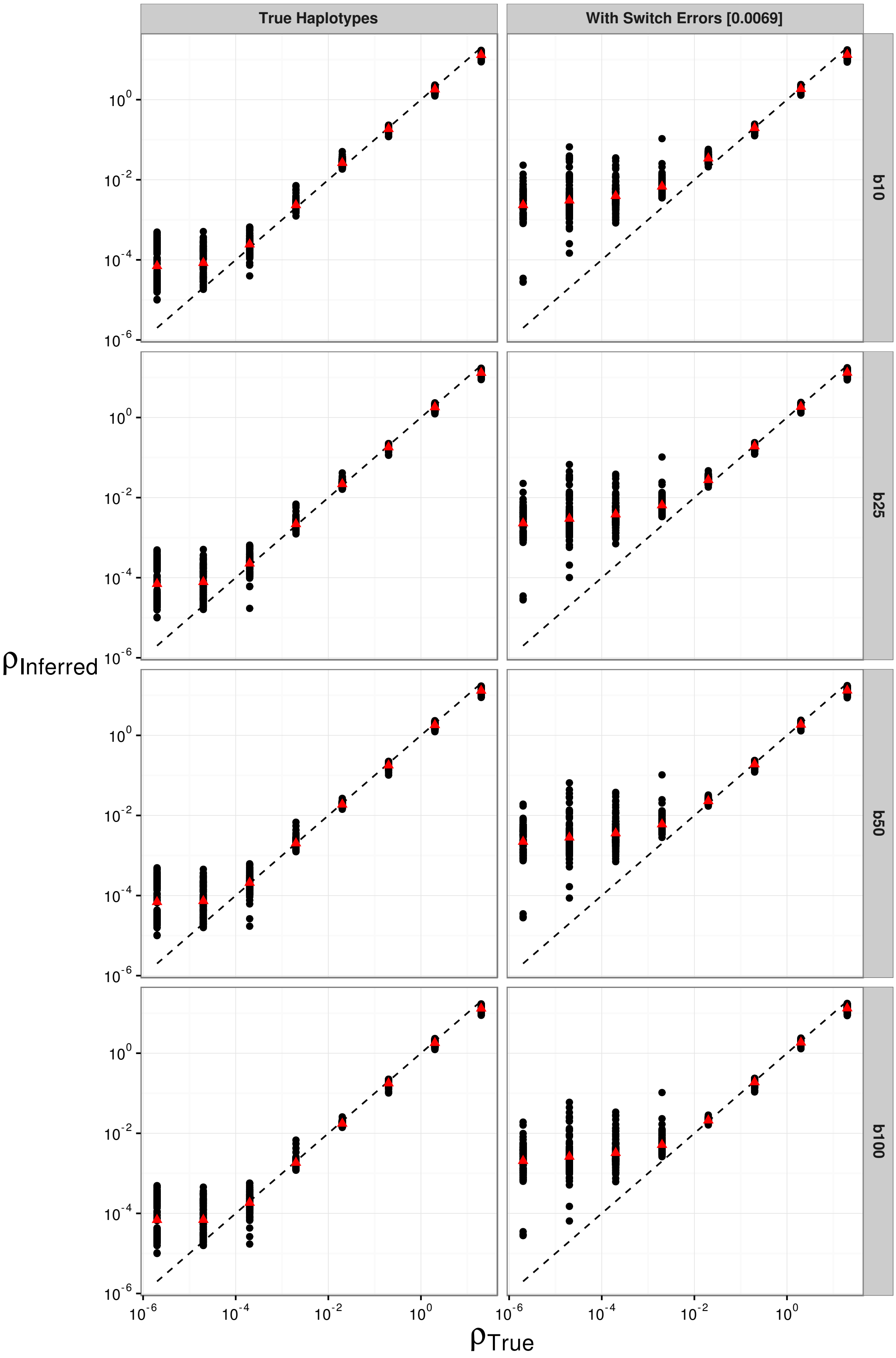


Figure B.1: The effect of switch errors and block penalty on the mean recombination rate inferred using LDhelmet. Block penalties (b) of 10, 25, 50 and 100 were used, shown in the vertically ordered facets from top to bottom.

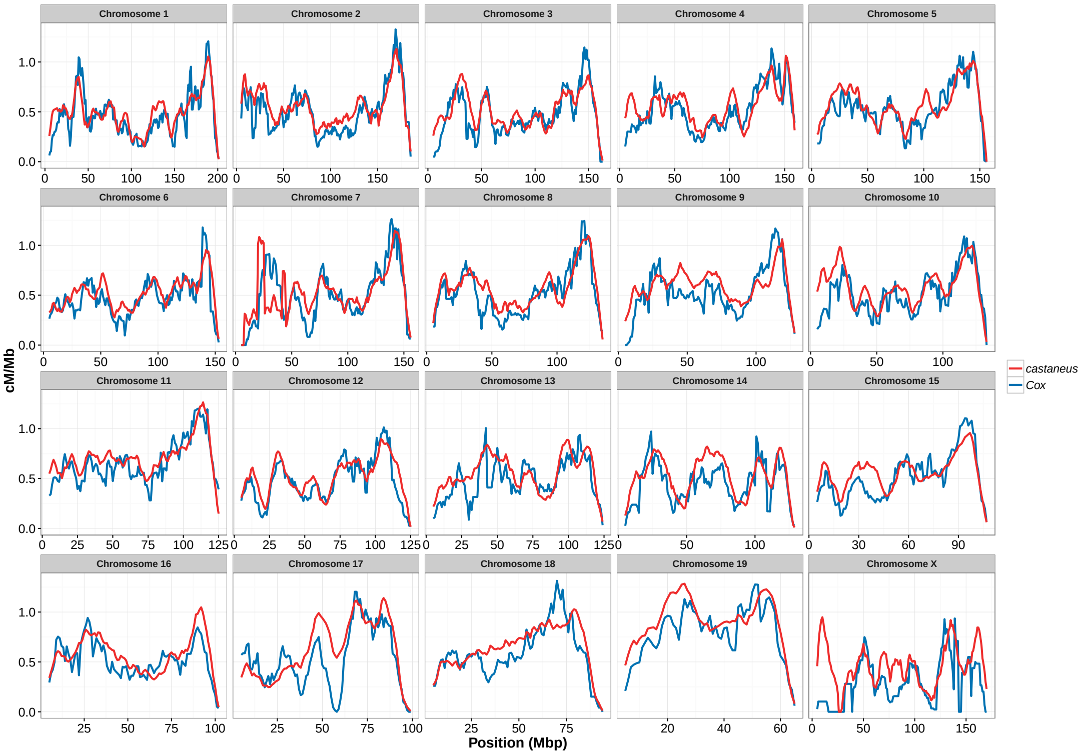


Figure B.2: Comparison of recombination rates inferred for *M. m. castaneus* using LDhelmet and recombination rates reported by Cox et al. (2009). Recombination rates in units of ρ/bp for the *castaneus* map were converted to cM/Mb by scaling using the genetic length of the corresponding chromosome in the Cox map.

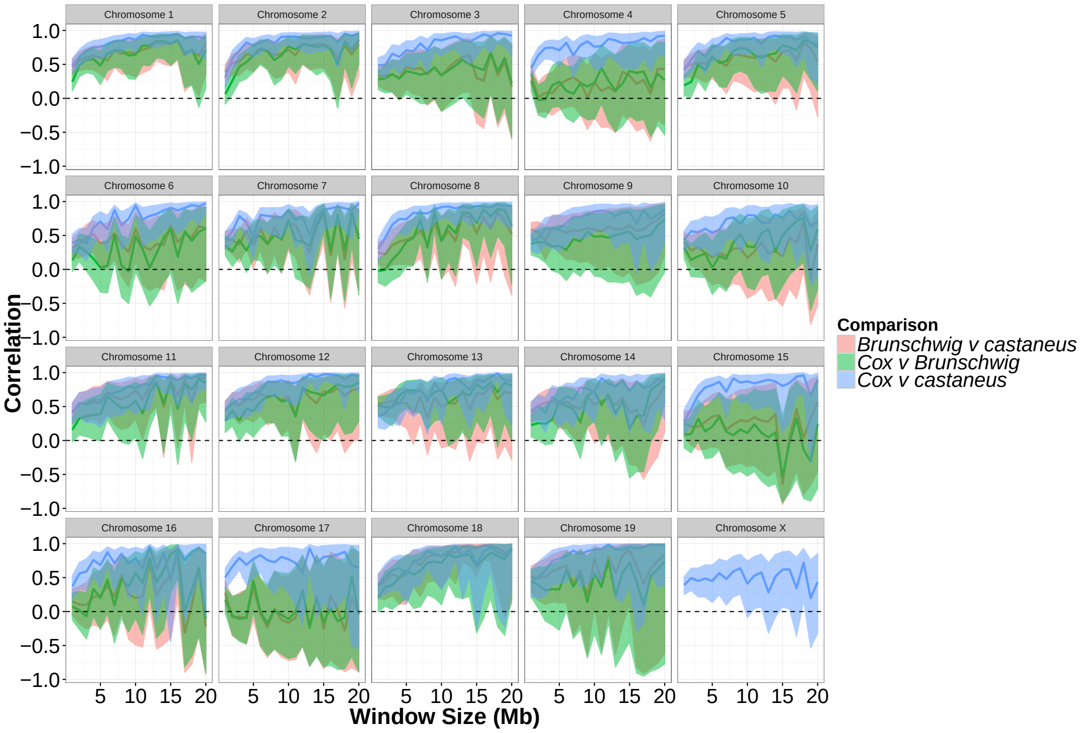


Figure B.3: Pearson correlation coefficients between the recombination map inferred for *M. m. castaneus*, the Brunschwig et al. (2012) map and the Cox et al. (2009) map for each chromosome separately. Comparisons for the X-chromosome were not made with Brunschwig et al. (2012) map, as it was not included in that study.

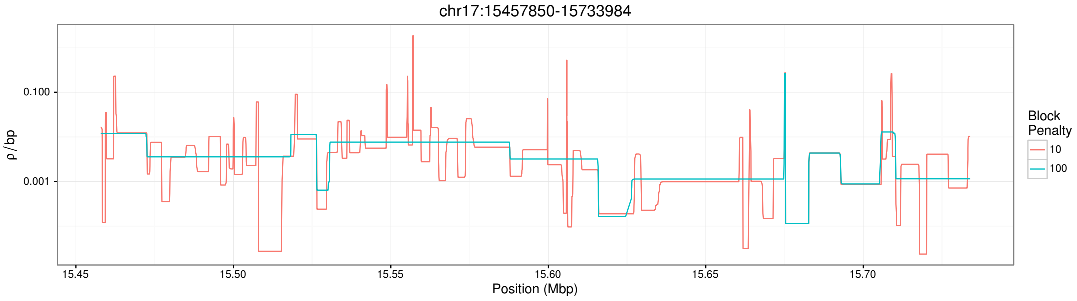


Figure B.4: A snapshot of the recombination rate landscape inferred for *M. m. castaneus* with LDhelmet block penalties. Recombination hotspots were inferred using the map constructed using a block penalty of 100.

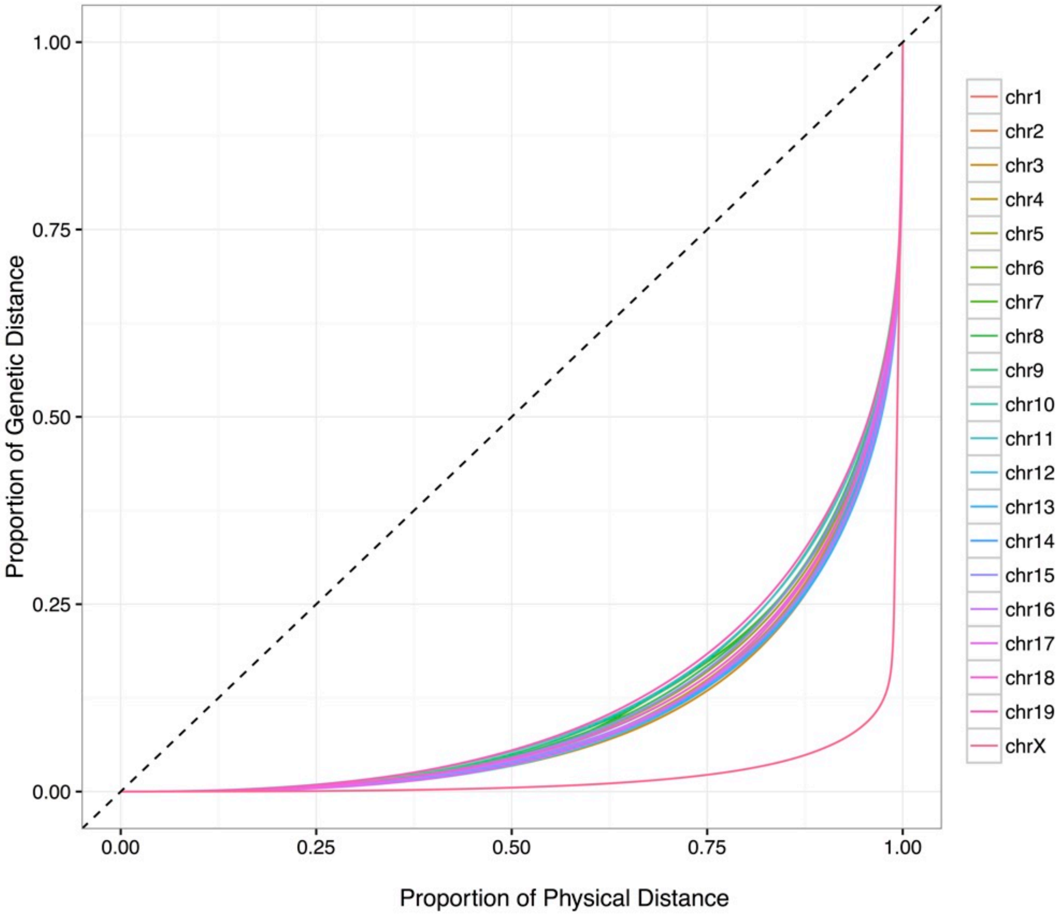


Figure B.5: Lorenz curves for each of the autosomes of *M. m. castaneus* as well as the X-chromosome.

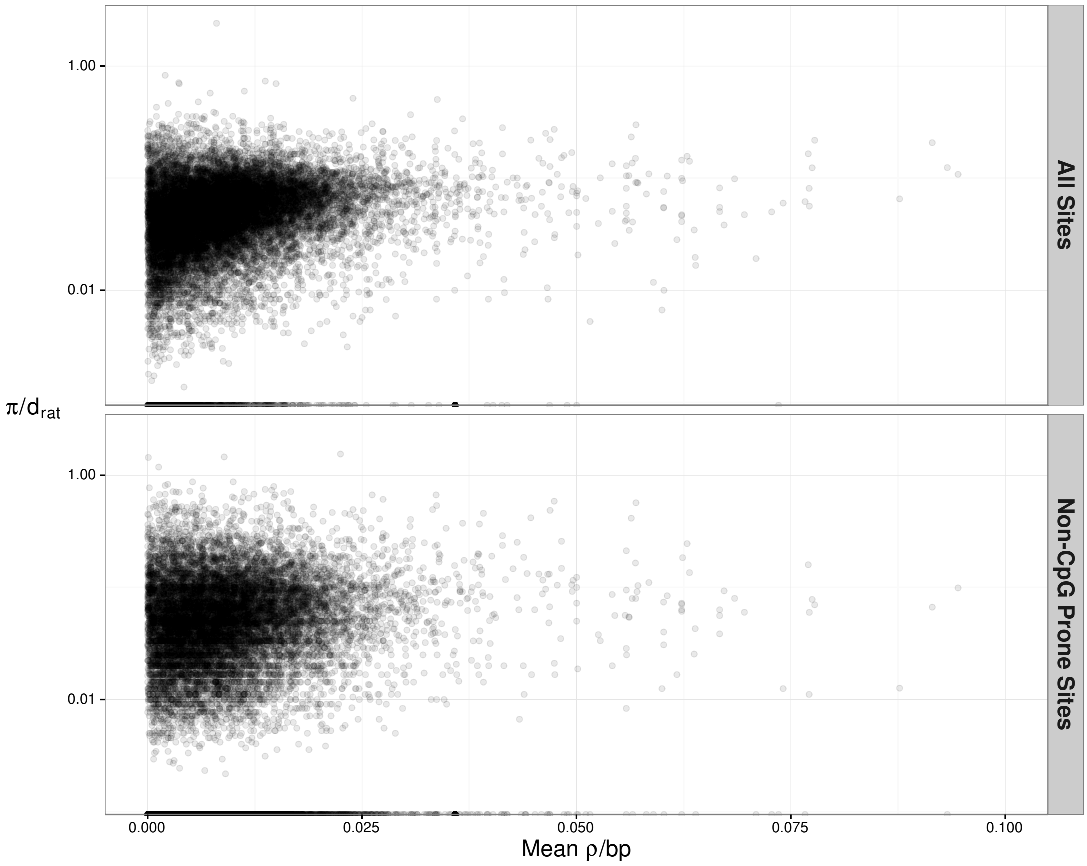


Figure B.6: The effect of switch errors on the mean recombination rate inferred using LDhelmet with a block penalty of 100. Each black point represents results for a window of 4000 SNPs, with 200 SNPs overlapping between adjacent windows, using sequences simulated in SLiM for a constant value of ρ/bp. Red points are mean values. Switch errors were randomly incorporated at heterozygous SNPs with probability 0.0046. The dotted line shows the value when the inferred and true rates are equal

Appendix C

Selection at linked sites in wild mice

C.1 Supplementary Methods

C.1.1 Demographic correction

DFE-alpha corrects for population demographic history when inferring selection parameters. However, there are several processes that can cause distortions of the uSFS that are not captured using the relatively simple demographic models. In particular, we found that the high frequency elements of the uSFS are not accurately captured under the demographic models implemented in DFE-alpha (Figure S1) (see main text). We make the assumption that the processes causing distortions to the uSFS for neutral sites also affect the selected site uSFS. Under this assumption, we correct the individual elements of the uSFS for selected sites using Equation 7 from Keightley et al. (2016):

$$N`_j = \frac{N_j}{1 + \frac{S_j - E_j}{E_j}} for j = 0, 1, 2...n \tag{C.1}$$

Where N_j is the number of derived mutations present in j copies in the selected site data and S_j and E_j are the observed and expected number of mutations at frequency j in the synonymous site data. Expected numbers of sites come from the fit of a neutral demographic model. Conceptually similar corrections have been applied by Tataru et al. (2017); Glemin et al. (2015); Eyre-Walker et al. (2006).

C.1.2 Divergence correction

Likelihoods in DFE-alpha are calculated using the allele frequency vector (AFV). The AFV is a vector of counts for mutations at different frequencies, accounting for population size change and selection. In a sample of alleles drawn from a population, polymorphic sites may resemble fixed derived mutations due to sampling effects. For each element of the AFV, we calculate the binomial probability of observing 20 fixed derived alleles (the number of samples in the *M. m. castaneus* data). The expected proportion of spurious fixed derived sites is then subtracted from the observed data and re-distributed to the polymorphic bin using the AFV to apportion out the number of sites appropriately.

We implemented the divergence correction in an iterative fashion as follows. After fitting demographic models or selection models to the uSFS using DFE-alpha, we remove the number of sites from the fixed derived class of sites and redistribute the proportion removed to polymorphism bins using the AFS estimated from the model fitted as described above. The resulting uSFS is then fitted using DFE-alpha as before. This procedure is repeated until the likelihood difference between successive iterations is less than 1.0. In all cases of applying this correction, likelihoods converged within 5 iterations.

C.1.3 Ancestral effective population size, Ne-anc Model B

Between-species divergence and within-species polymorphism may not reflect the same suite of processes. If, for example, in the time since a focal species split from the outgroup used to estimate the uSFS the DFE for advantageous mutations changed, the between-species divergence at selected sites will reflect a combination of the current DFE and the previous one. Polymorphisms present in the population at the time of sampling, however, may only reflect the most recent DFE. This is because divergence is accumulated over all time since the split of the focal species and the outgroup while polymorphism may only reflect recent processes. If polymorphism and divergence have become decoupled including fixed derived sites when estimating selection parameters may lead to bias. Indeed, Tataru et al. (2017) showed that accurate estimates of selection parameters can be obtained solely from the polymorphism data in the uSFS.

We modified the likelihood function of Schneider et al. (2011) by adding an additional parameter, that we call the ancestral effective population size (N_{e-anc}). N_{e-anc} is fitted solely to the fixed derived class of sites. Fitting N_{e-anc} has the effect of absorbing the contribution of fixed derived sites, so that selection parameters are estimated from polymorphism and invariant ancestral sites only. We define N_{e-anc} size as the population size that, given the current estimates of the selection coefficients, satisfies the number of fixed derived mutations in the sample. We use Kimuras formula for the fixation probability (Q) of a selected allele in a population of size N:

$$Q = \frac{(1 - e^{-s_a})}{1 - e^{-2N_e s_a}} \tag{C.2}$$

where s is the selection coefficient for homozygotes, and N_e is the effective population size. Models that do not incorporate N_{e-anc} are nested within those that do, so likelihood ratio tests can be used for comparisons. Model B in the main text fits the N_{e-anc} parameter alongside the selection parameters. Because polymorphism and divergence could potentially become decoupled by multiple processes, the value of N_{e-anc} is difficult to interpret.

C.2 Supplementary Tables and Figures

Table C.1: Comparison of the fit of demographic models based on the analysis of 4-fold sites and CNE-flanks in *M. m. castaneus*.

	Epochs	ΔlnL	χ^2	# Estimated Parameters
4-fold	1	1,620	22,500	2
	2	159	2,930	4
	3	0.0	553	6
CNE-flank	1	19,100	53,500	2
	2	1,350	5,070	4
	3	0.0	975	6

Table C.2: Parameters of the best-fitting demographic model estimated from the analysis of 4-fold and CNE-flanking sites.

	4-fold	CNE-flank
N2/N1	0.40	0.07
t2/N1	0.44	0.17
N3/N1	0.40	1.00
t3/N1	1.10	0.63

Table C.3: Parameters of the 3-epoch demographic model at different sample sizes. Down sampled datasets were generated by randomly selecting alleles, with respect to frequency, from the full dataset of 10 individuals.

Parameter	Number of alleles sampled		
	$n = 10$	$n = 16$	$n = 20$
N2/N1	0.030	0.030	0.060
t2/N1	0.204	0.140	0.181
N3/N1	0.120	0.200	0.800
t3/N1	0.080	0.220	0.461

Table C.4: Likelihood differences between models of the deleterious DFE (dDFE) fitted with or without a single class of adaptive mutations.

Site Type	dDFE Model	ΔlnL	
		dDFE	**dDFE + Adaptive Mutations**
0-fold	1-Class	49,300	4.18
	2-Class	129	0.00
	3-Class	129	0.00
	Gamma	247	4.18
CNE	1-Class	51,000	245
	2-Class	1,660	3.41
	3-Class	1,480	0.00
	Gamma	2,310	19.3
UTR	1-Class	6,170	32.7
	2-Class	335	0.00
	3-Class	335	0.00
	Gamma	970	13.5

Table C.5: Parameter estimates for the scaled effect and frequency of advantageous mutations in three classes sites in *Mus musculus castaneus* when models incorporated either one class of advantageous mutations, or two.

# Adv. Mutation Classes	0-fold		UTR		CNE	
	1	**2**	**1**	**2**	**1**	**2**
	Model A: Full uSFS					
$N_e s_{a,1}$	7.27	7.27	5.32	5.32	9.17	9.17
$p_{a,1}$	0.003	0.003	0.0133	0.0133	0.0098	0.0098
$N_e s_{a,2}$	-	0.000	-	0.000	-	0.000
$p_{a,2}$	-	0.000	-	0.000	-	0.000
ΔlnL	-	0.000	-	0.005	-	0.000
	Model B: Ignoring sites fixed for the derived allele					
$N_e s_{a,1}$	8.30	8.30	6.96	6.96	8.60	8.60
$p_{a,1}$	0.010	0.010	0.0294	0.0294	0.008	0.008
$N_e s_{a,2}$	-	0.0925	-	33.6	-	0.240
$p_{a,2}$	-	0.000	-	0.000	-	0.000
ΔlnL	-	0.000	-	0.000	-	0.002

Table C.6: Parameters of the selection and demographic models when estimated from simulated data. The upper portion of the table shows the selection parameters. The selection model simulated contained two classes of deleterious mutational effects and a single class of advantageous mutations. The lower portion of the table contains the demographic parameters inferred for *M. m. castaneus* compared with those inferred from simulation data.

		Model A		**Model B**	
		Simulated	Estimated	Simulated	Estimated
	$N_e s_0$	-0.045	-0.70	-0.171	-0.40
	p_0	0.191	0.145	0.184	0.181
	$N_e s_1$	-104	-92.3	-100	-77.3
	p_1	0.806	0.784	0.806	0.799
	$N_e s_a$	7.27	0.950	8.30	4.91
	p_a	0.00300	0.0710	0.0100	0.0200
	M. m. castaneus				
$N2/N1$	0.40	-	0.20	-	0.12
$N3/N1$	1.4	-	1.0	-	0.9
$t2/N2$	0.31	-	0.46	-	1.2
$t3/N3$	0.79	-	1.4	-	1.3

Table C.7: Parameters of strongly selected mutations assumed in simulations.

$2N_e s_a$	p_a	$2N_e s_a p_a$
400	0.000218	0.0872
400	0.000109	0.0436
200	0.000436	0.0872
200	0.000218	0.0436
100	0.000436	0.0436

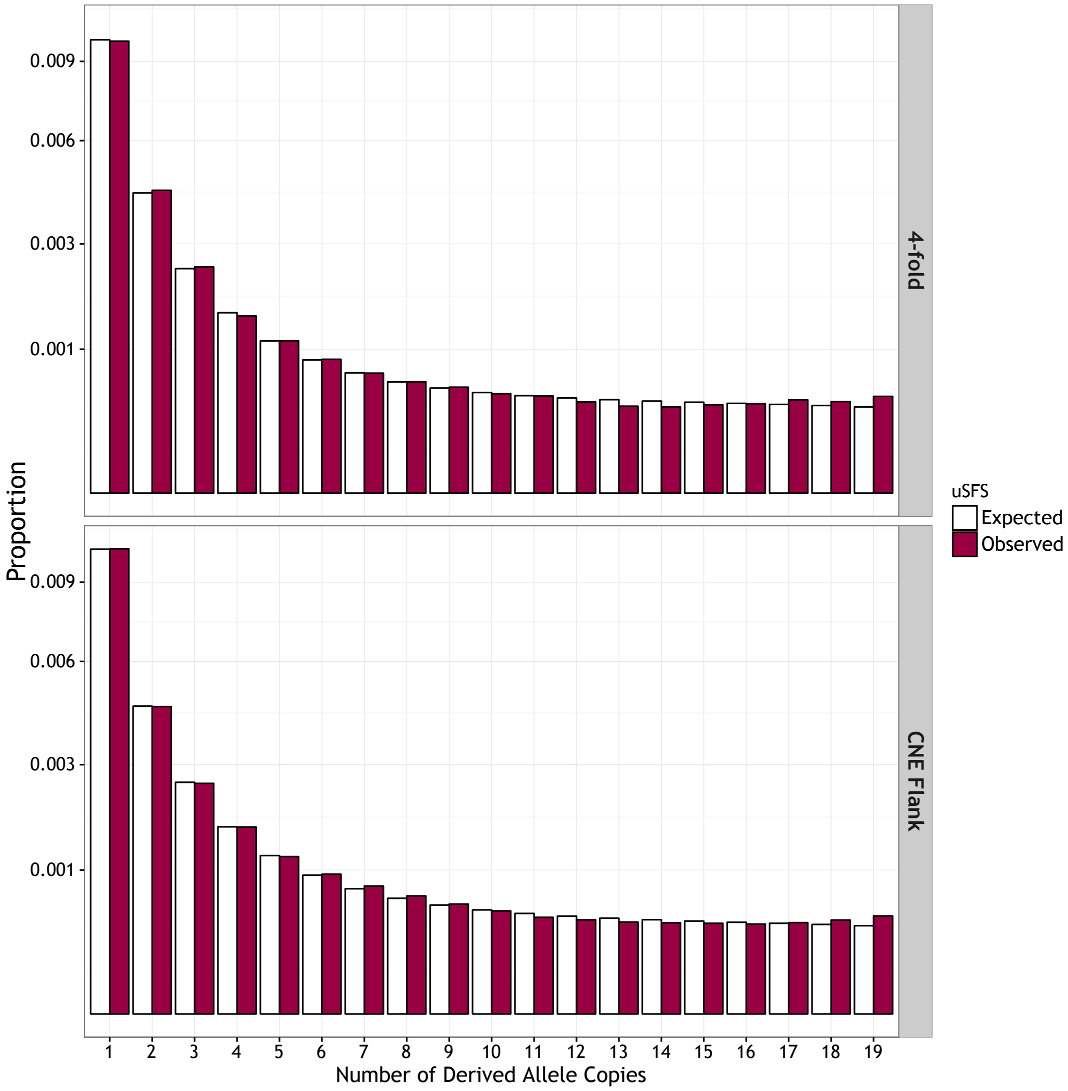


Figure C.1: A comparison of the uSFS expected and observed under the best-fitting demographic models for two classes of putatively neutral sites, 4-fold degenerate synonymous sites and CNE-flanking sequences.

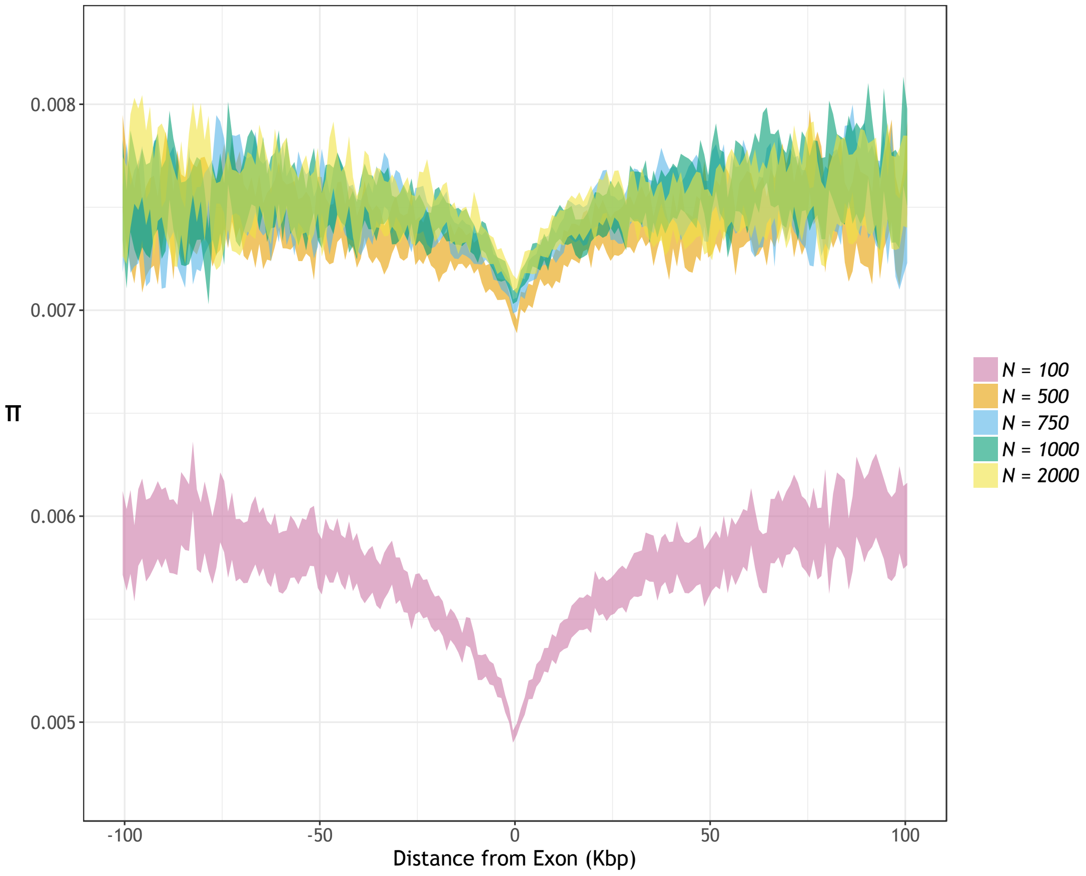


Figure C.2: The effect of population size on patterns of diversity obtained from simulated populations.

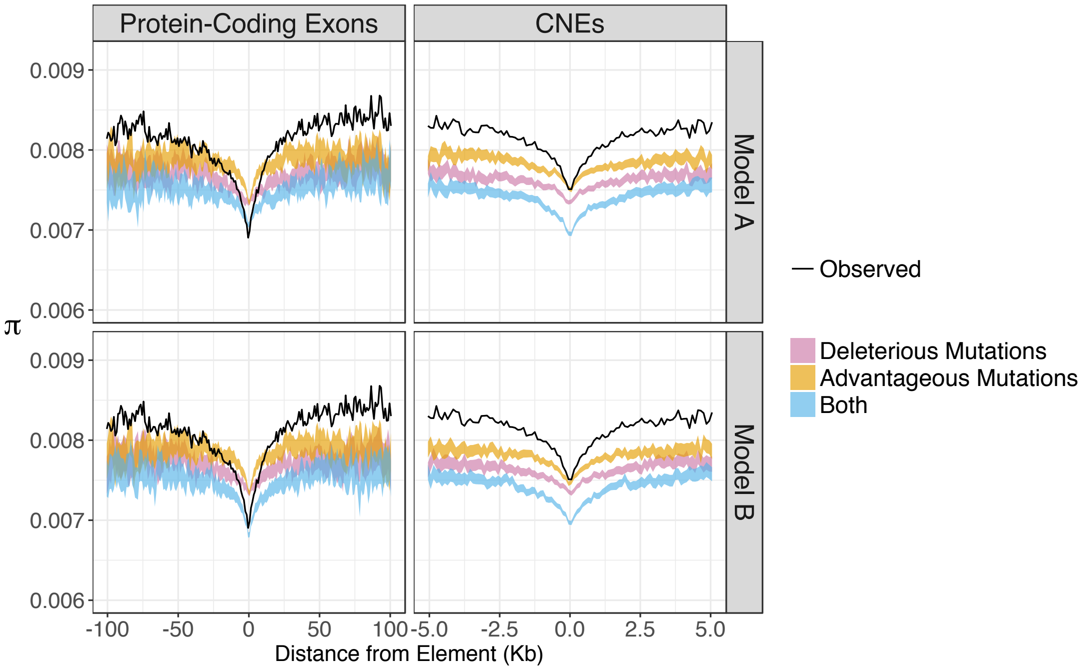


Figure C.3: Estimates of unscaled π around protein-coding exons and CNEs in *M. m. castaneus* (black line) compared to the values observed in simulated populations (coloured ribbons) when distances are measured using physical distances.

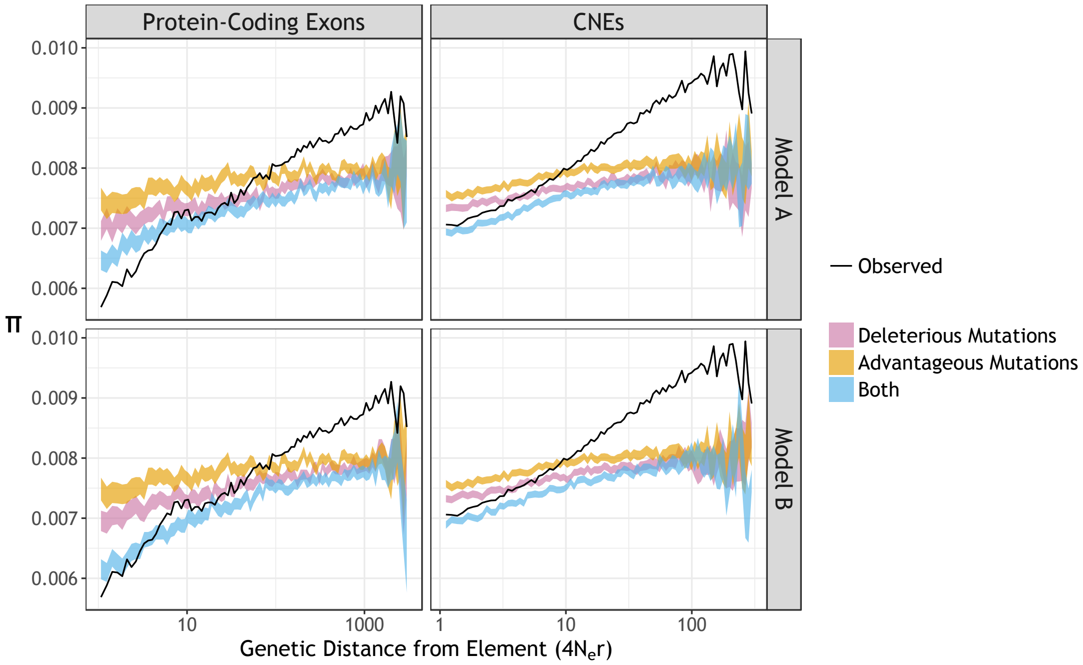


Figure C.4: Estimates of unscaled π around conserved non-coding elements in *M. m. castaneus* (black line) compared to the values observed in simulated populations (coloured ribbons) when distances are measured using population-scale genetic distances.

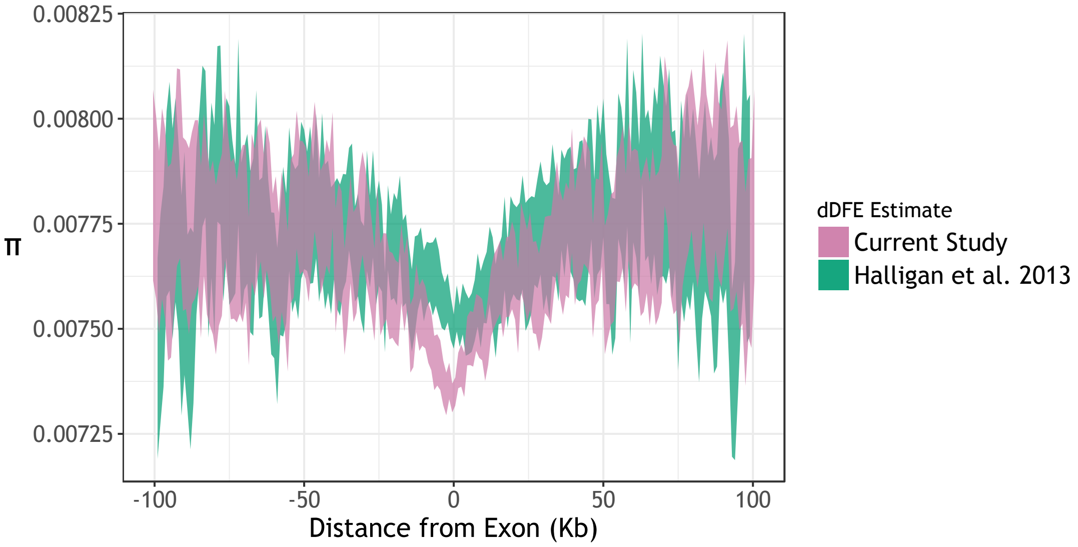


Figure C.5: Comparison of nucleotide diversity (π) around protein-coding exons in simulated populations under either the discrete-class dDFE estimated in the current study or the gamma dDFE estimated by Halligan et al. (2013)

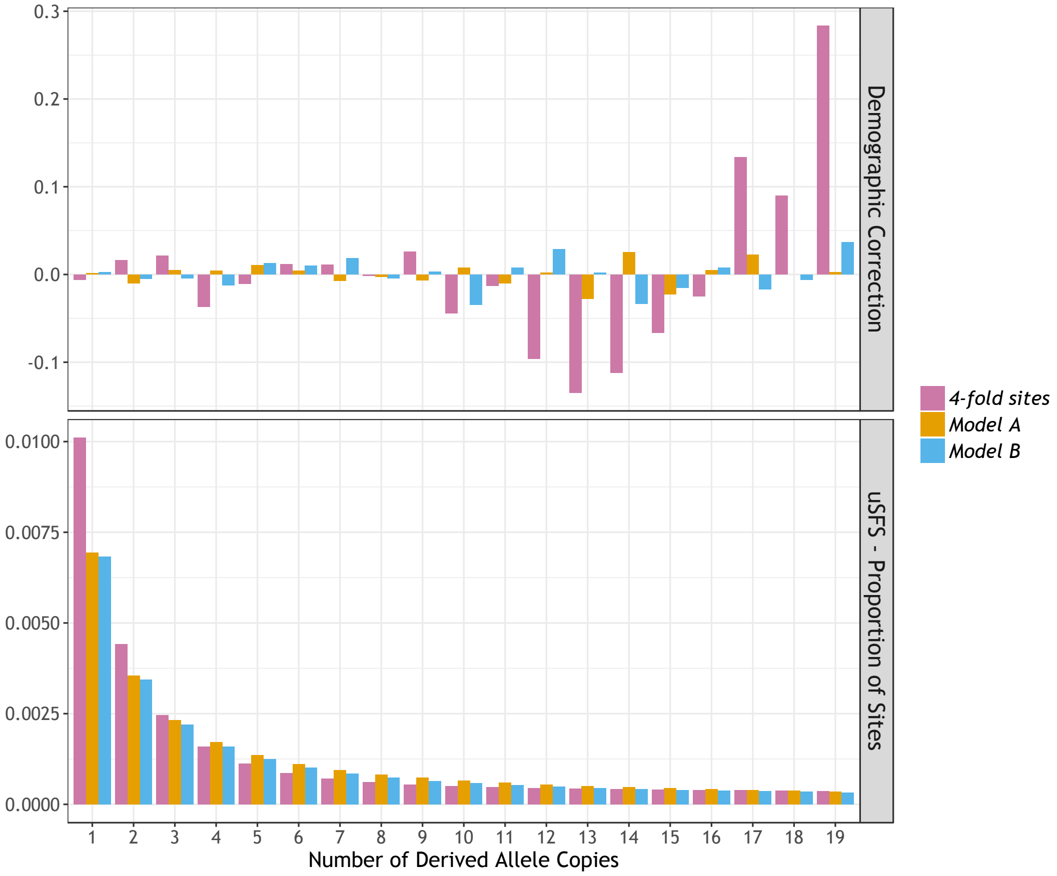


Figure C.6: The demographic correction and uSFS for 4-fold sites in *M. m. castaneus* and synonymous sites simulated under Model A or Model B selection parameters. The upper panel shows the proportional deviation between the observed uSFS and that expected under the best-fitting demographic model. The lower panel shows the uSFS.

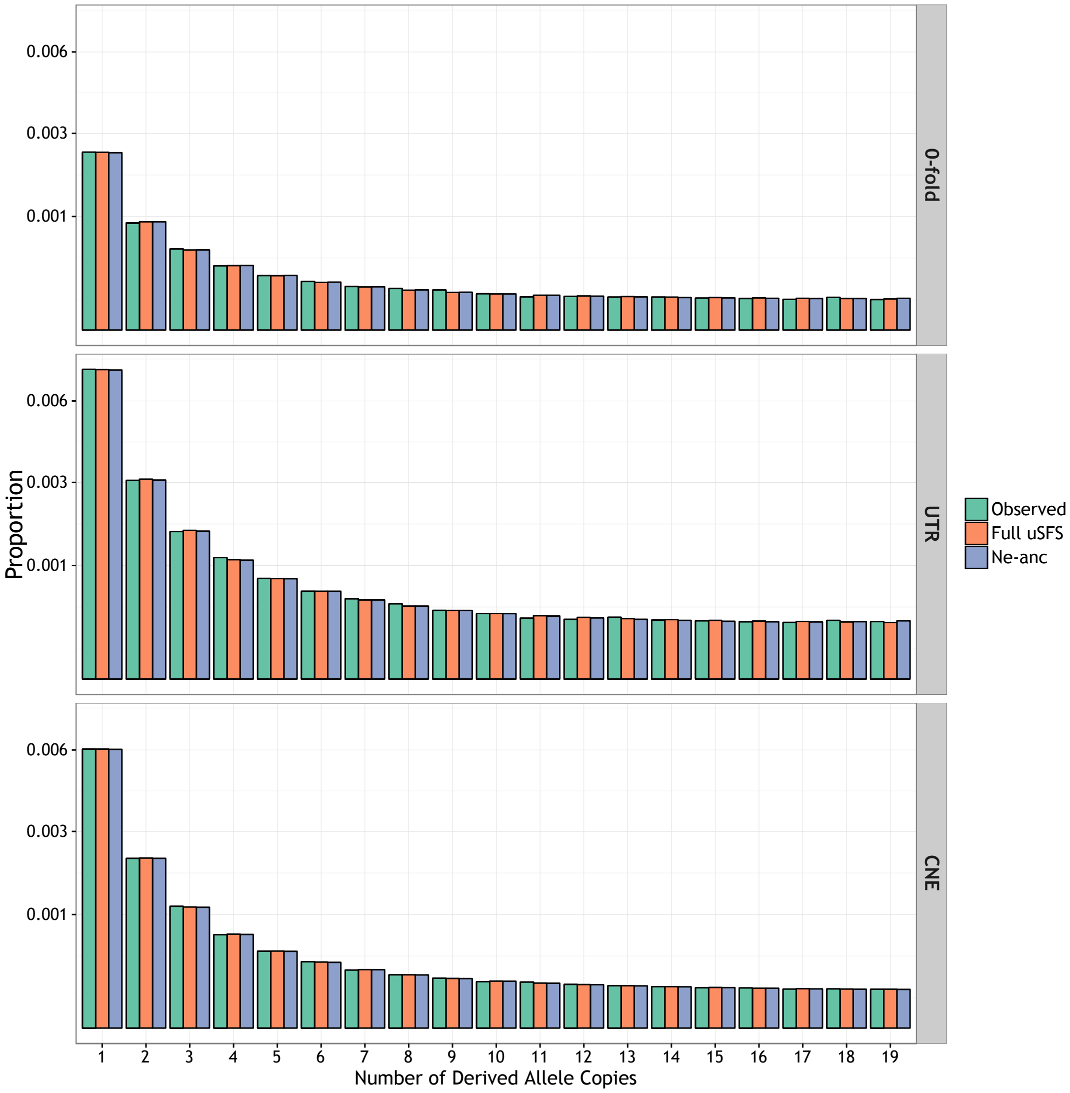


Figure C.7: A comparison of the observed uSFS with that expected under the best-fitting selection models for three classes of functional sites.

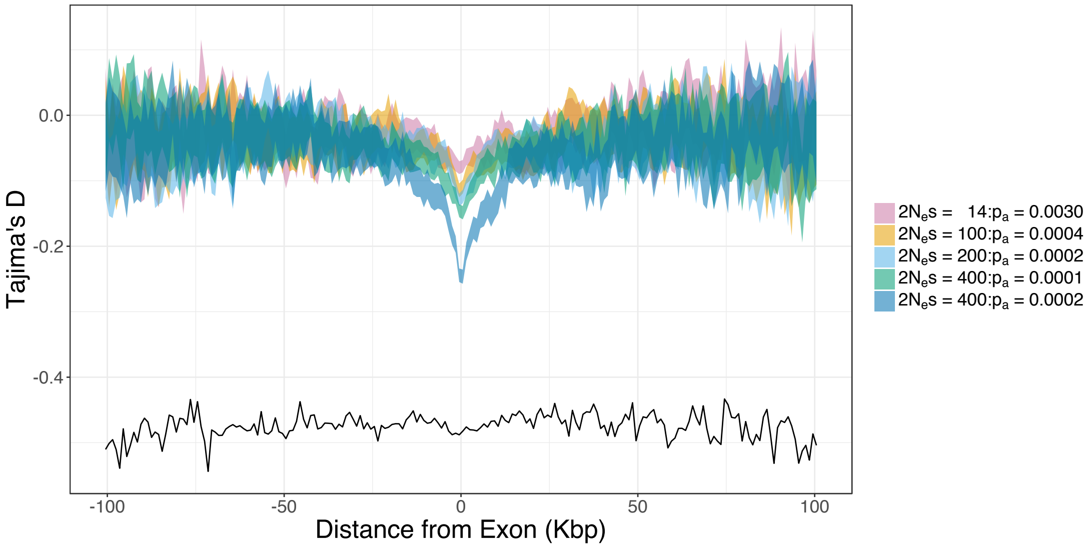


Figure C.8: The pattern of Tajimas D around protein-coding exons for simulations assuming strongly selected advantageous mutations.

Appendix D

Estimating sweep parameters